Weakly Supervised Labeling Strategies for Classifying User-generated Content

by

Matti Wiegmann

Dissertation to obtain the academic degree of
Dr. rer. nat.

Faculty of Media
Bauhaus-Universität Weimar
Germany

Advisor:	Prof. Dr. Benno Stein
Reviewer:	Prof. Dr. Paolo Rosso
Date of oral exam:	17. July 2025

Abstract

Weakly Supervised Labeling Strategies for Classifying User-generated Content

This dissertation presents a principled approach to weak supervision for creating large labeled datasets. Weakly supervised strategies derive labels for an unlabeled dataset from some distant knowledge, enabling large-scale dataset creation even for tasks where human annotations are infeasible. Instead of developing individual strategies depending on the task, available data, and required knowledge, we first determine the design parameters of weak supervision strategies and, based on these, develop strategies for specific tasks. We demonstrate and evaluate our principled approach to weakly supervised annotation through three case studies in the domain of user-generated content, where we create large labeled datasets and use them to answer research questions of societal interest.

The first case study is an analysis of why some debaters are more persuasive than others. For this study, we created a dataset of 3,801 Reddit users and their history of debate posts. A weak supervision strategy determines debater persuasiveness over time which provides a new perspective on the intersection of persuasiveness, experience, and argument characteristics.

The second case study is an investigation into author profiling technology in difficult situations: with little available text per author or with few authors per label. For this study, we created a dataset of 71,706 Twitter users and their complete timeline of tweets. A weak supervision strategy determines up to 239 personal attributes for each user, allowing model development and offering test cases for profiling understudied attributes.

The third case study tackles a novel task: assigning trigger warnings to fiction documents. For this study, we created a dataset of about 1 million fan fiction documents from 'Archive of our Own'. A weak supervision strategy labels each document with the appropriate trigger warnings from a unified 36-label trigger warning taxonomy that we created for this study.

Abstract (in German)

Weakly Supervised Labeling Strategies for Classifying User-generated Content

In dieser Dissertation wird ein grundsatzorientiertes Verfahren vorgestellt, um grosser Datensätze mit schwach überwachten Annotationsstrategien zu erstellen. Solche Strategien ermöglichen die Erstellung grosser Datensätze zu Problemen, bei denen manuelle Annotation ungeeignet ist. Anstatt individuelle Strategien in Abhängigkeit des Problems, der verfügbaren Daten und des erforderlichen Wissens zu entwickeln, bestimmen wir die Designparameter schwach überwachter Annotationsstrategien. Wir demonstrieren und evaluieren unser grundsatzorientiertes Vorgehen zur schwach überwachten Annotation anhand von drei Fallstudien im Bereich nutzergenerierter Inhalte, bei denen wir grosse Datensätze erstellen und Nutzen, um gesellschaftlich relevante Forschungsfragen eruieren.

In der ersten Fallstudie untersuchen wir, warum einige Debattanten überzeugender argumentieren als andere. Für diese Studie haben wir einen Datensatz von 3.801 Reddit-Nutzern und all ihren Debattenbeiträge erstellt. Eine schwach überwachte Annotationsstrategie bestimmt die Überzeugungskraft der Debattanten über Zeit. Damit bietet der Datensatz eine neue Perspektive auf den Zusammenhang Zusammenhang zwischen Überzeugungskraft, Erfahrung und Argumenteigenschaften. In der zweiten Fallstudie untersuchen wir Technologien zur Eigenschaftsanalyse von Autoren ('Profiling') unter schwierigen Bedingungen: mit wenig verfügbarem Text pro Autor oder mit wenigen Autoren pro Attribut. Für diese Studie haben wir einen Datensatz von 71.706 Twitter-Nutzern und all ihren Tweets erstellt. Eine schwach überwachte Annotationsstrategie identifiziert bis zu 239 persönliche Eigenschaften für jeden Nutzer und ermöglicht so die Entwicklung von Modellen und Testfällen für die Analyse selten untersuchter Eigenschaften. In der dritten Fallstudie untersuchen wir ein neues Problem: die Zuordnung von Inhaltswarnungen ('Trigger Warnings') zu Fan-Fiction-Dokumenten. Für diese Studie haben wir einen Datensatz von etwa 1 Million Dokumenten aus 'Archive of our Own' erstellt. Eine schwach überwachte Annotationsstrategie bestimmt für jedes Dokument die Inhaltswarnungen basierend auf einer für diese Forschung entwickelten 36-Label-Taxonomie.

Ehrenwörtliche Erklärung

Ich erkläre hiermit ehrenwörtlich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Teile der Arbeit, die bereits Gegenstand von Prüfungsarbeiten waren, sind ebenfalls unmissverständlich gekennzeichnet.

Weitere Personen waren an der inhaltlich-materiellen Erstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich hierfür nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten (Promotionsberater oder anderer Personen) in Anspruch genommen. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt. Ich versichere, dass ich nach bestem Wissen die reine Wahrheit gesagt und nichts verschwiegen habe.

Weimar, 17. Januar 2025

Matti Wiegmann

Contents

1 Introduction

This thesis studies weak supervision and its application to the creation of large labeled datasets for processing user-generated content on social media platforms. The research is organized around three case studies about problems relevant to society. Each study contributes a conceptual analysis, generic insights, and a technical solution to a problem in weak supervision.

Social media platforms[1] allow us to explore novel research questions about our society by supplying data that is otherwise unavailable. These platforms are central to today's information-sharing infrastructure. They are places to post user-generated content, which can be opinions and arguments on a topic of debate, news of events, advertisements for products or services, creative works in writing, drawing, photography, or film, or slices of life such as vacation photos or video logs. They are also places to interact with other users and their content, which may include (dis)liking, sharing, commenting, tagging, categorizing, curating, or aggregating this content. User-generated content and its metadata represent a model of society in data that is orders of magnitude larger and comparatively more accessible when compared to other means of quantitative research in computational social science, natural language processing, and related fields [69, 96, 198].

[1] It is difficult to define what counts as a "social media" platform [2] since the term's understanding changes to include the contemporarily popular sites. We assume a minimal definition and refer to all platforms that disseminate primarily user-generated content as a means of communication. This includes social networks (Facebook), forums (Reddit), blogs (Substack), microblogs (Twitter), recommendation-driven platforms (TikTok, YouTube), but excludes collaborative or non-communication sites (Wikipedia).

In practice, the available data is a source for answering relevant questions about society that otherwise lack an angle of attack because the required data is hard to obtain. For example, consider the following questions:

- *Can an author's writing reveal his personal characteristics?*
  This is a hard question to study because most forms of writing, such as letters or short messages, are not public, and the characteristics of an author, such as demographics, are not usually known, as is the case with texts found on the web. Social media platforms enable this kind of study because their users' texts are public, often along with a profile that may reveal relevant characteristics.

- *What framing makes an argument more persuasive?*
  This is a hard question to study because of the large number of arguments on each topic and the potential framings for each. Constructing different framings of arguments and testing their persuasiveness in a debate across different audiences is arduous at best. As social media users often post arguments in different frames, collecting persuasiveness ratings based on interaction data becomes feasible.

- *Why do people use hate speech and how can it be mitigated?*
  This is a hard question to study because people who produce and receive hate speech are difficult to recruit for a study, so their behavior can only be observed, not induced. Because hate speech is public and common on platforms, it is possible to collect large samples for study.

Computer science contributes to answering such questions by modeling the respective objectives as classification tasks, either because an effective classification involves answering the question (the classifier detects the hate speech, the demographic attribute, or the frame), or because classifying the data is the only way to determine the property of interest, the class, for enough data points to allow further research. The latter case is particularly common with user-generated content, where raw data is ubiquitous but class information is often scarce. Consequently, to classify the data, an initial labeled dataset is required to train probabilistic models based on the properties of the labeled dataset and to evaluate any (pre)trained models, such as a generative large language models, and any non-probabilistic methods, such as hand-crafted heuristics and regular expressions.

In both cases, training and testing, a complete and reliable relation between all data points and their associated labels is required. The completeness of the labels, i.e. no data point is without a label, is in most cases a

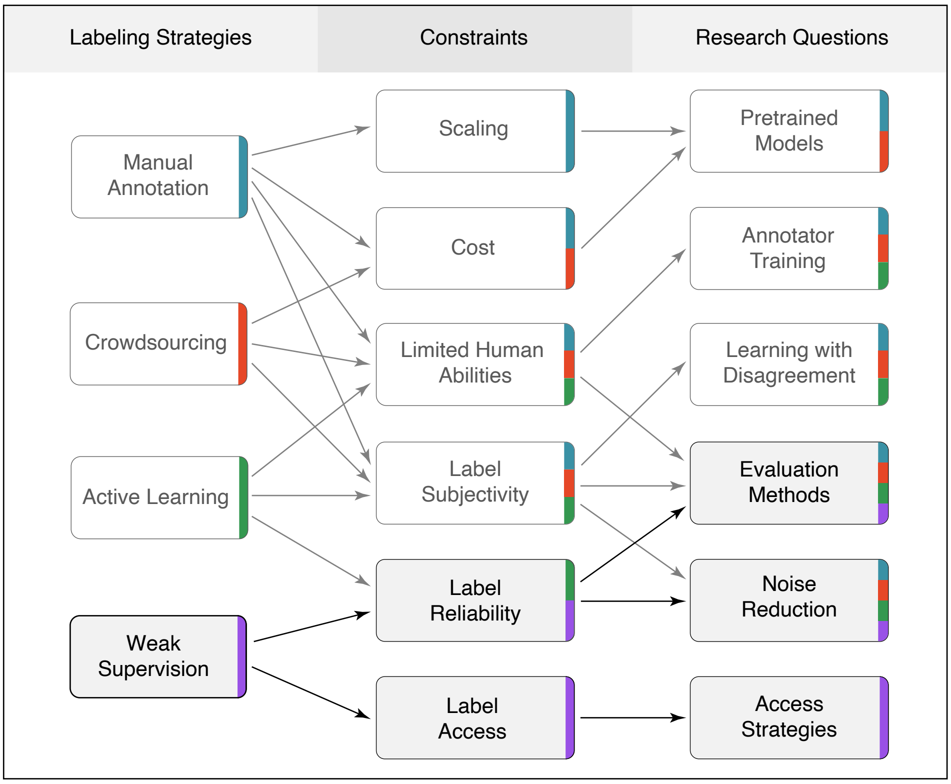


Figure 1.1: Several strategies exists to label data, the four most important are manual annotation, crowdsourcing, active learning, and weak supervision. Each strategy has its constraints that makes it more or less suited for certain tasks or data types. Research questions usually evolve around overcoming these constraints. This work is mostly concerned with weak supervision.

prerequisite for the classifier and can be guaranteed by preprocessing. The reliability of the labels, i.e. all associated labels are correct, complete, and unambiguous, is assumed if the labels result from a traditional annotation process by human experts or from structured and controlled annotation campaigns, at least in empirical computer science.

However, reliably labeled data is often not available in the quantities required for training a classifier (see Figure 1.1). We distinguish two cases:

1. Collecting data through a traditional annotation process is not feasible when human annotators cannot reliably assign a class in the problem domain, because (1) the labeling decision is subjective, such as for content moderation or relevance judgments, (2) domain expertise is required, for example in legal or biomedical domains, or (3) the

classes are complex and difficult for humans to discern from the data, such as for authorship attribution or user profile prediction.

2. Collecting data through a traditional annotation process is not feasible on a sufficiently large scale for data-intensive classification methods, such as training deep neural networks, because of the number of examples, the number of labels, or the effort required per label. Alternatives for scaling up the traditional annotation process are crowdsourcing, for example using Amazon Mechanical Turk, and active learning. However, both techniques still rely on subjective and error-prone manual annotation and time-consuming and expensive setup [133]. In addition, active learning faces issues with concept drift and crowdsourcing is being disrupted by large language models, which are used by an estimated 33-46% of all crowd workers to complete their tasks [224].

Under these circumstances, it is still possible to collect a large amount of labeled data using weak supervision. Weak supervision is an umbrella term for all techniques within supervised machine learning that determine the label of a data point using some form of "distant knowledge" that is relevant for the problem at hand. Distant knowledge can be metadata such as geotags, knowledge bases such as Wikidata, rule sets, models trained on data from a different task or domain, or other sources besides a human annotated gold standard or the data themselves.

An example from the field of natural language processing is weak supervision for relation extraction [139] to determine the relation (*is married to*, *is acquired by*, ...) between two entities in any given sentence. Here, as distant knowledge, triplets of two entities and their relation are extracted from a semantic knowledge base such as Wikidata, and this relation is then assigned to each sentence containing these two entities. These relations are then used to train a supervised classifier.

An example from the field of computer vision is weakly supervised image segmentation [49] to partition an image into pixel-level segments without pixel-level labels. Here, in a first stage, a pre-trained image-level classifier is used to determine the objects in the image and, in a second stage, the classifier's weighted, pixel-level features are used to determine the "weak label" of each pixel. Which pixels belong to the same segment is then determined heuristically: If neighboring pixels have the same label, they belong to the same segments. Finally, an image segmentation model is trained with these pixel-level labels using supervised learning.

Weak supervision is applicable to many content classification problems because distant knowledge is ubiquitous on social media platforms. Distant knowledge can be found in the rich metadata, such as author demographics from profile fields, geotags, and hashtags, or it can be inferred from user behavior, such as using likes as an indicator of whether an argument is persuasive. Because user-generated content is available in huge quantities, large datasets can be created even if only a fraction of the data can be labeled with weak supervision. Chapter 2 contains an extensive collection of related work using weak supervision to label user-generated content.

Developing a weak supervision strategy for data labeling requires a source of distant knowledge and making that knowledge accessible as labels. But, even if there is a lot of distant knowledge available for each data point, it is difficult to tap the distant knowledge needed for a particular problem due to sparsity and distance:

- **Sparsity:** The more specific the required knowledge is, the more sparse is the available data, because not every data point can be related to it. For example, non-specific information, such as users' gender, can be inferred from gender-typical names, read out from public profiles where the users' disclose it, or parsed from posts via self-referential phrases such as *"As a man, I* [...]*"*. In contrast, specific information, such as Myers-Briggs personality scores, are much rarer to find. The sparsity is exacerbated by sampling and de-biasing.
- **Distance:** While certain distant knowledge is closely related to the data, such as a date a post was created or the place of residence in the profile of a post's author, other knowledge is more distant and can require complex, multi-step heuristics to make a connection to the data. A large distance to the knowledge is harder to access and tapping into it may introduce more noise than less distant knowledge.

Finally, weak supervision generally trades label scale for label reliability, i.e., increasing sparcity and distance in a weak supervision strategy also increases the amount of label noise. Assessing label noise and controlling the reliability-to-scale ratio matters for weakly labeled datasets, albeit, as our survey in Chapter 2 shows, neither is common in practice. Large datasets created via weak supervision will contain label noise since the distant knowledge, or the way to access it, can be ambiguous or imprecise and the created labels cannot be individually verified at scale. This trade-off has different implications for training and for test datasets.

Classification models trained on noisy data generalize poorly. However, while the effectiveness of traditional machine learning based on feature engineering quickly decreases in the presence of label noise, current neural network models are robust to some percentage of label noise [194, 260]. The robustness depends on the amount of labeled data available, so increasing the size of the training data at the expense of some label noise often improves effectiveness. As Radford et al. [177] note in the context of the speech recognition model Whisper: *"moving beyond gold standard crowdsourced datasets such as ImageNet to much larger but weakly supervised datasets significantly improves the robustness and generalization of models."* Nevertheless, it is to be expected that there is a limit or equilibrium point to the relationship between dataset size, noise ratio, and model performance. If the labels are too unreliable, model effectiveness suffers, similar to the pivot from large corpora to high quality data for training large language models [117].

Test data, unlike training data, is susceptible to label noise and benefits less from scaling. Regarding the susceptibility, the label noise contained in the test data degrades the model scores and provides a false measure of the model's true effectiveness. In addition, the noise may also obscure model differences or even change the order of models when ranking them by effectiveness. Regarding scale, as Perlitz et al. [165] point out, making test datasets larger than necessary for validity is firstly more computationally expensive, especially if the evaluation is run often and with expensive models such as LLMs, and secondly it can emphasize imbalances or biases in the dataset.

1.1 Main Contributions

The contributions of this thesis are divided into (1) conceptual contributions in the form of the first systematic review (Chapter 2) on the data, knowledge, and strategies used to create and evaluate novel datasets with weak supervision and (2) practical contributions in the form of three case studies where the conceptual insights are applied to the creation of novel datasets to address specific research questions that could not previously be studied due to lack of data.

The first study (Chapter 3) examines the problem of using weak labels for computational analysis, specifically the persuasiveness of debaters on Reddit's ChangeMyView subreddit. The second study (Chapter 4) examines the problem of linking an external knowledge base to users on a social media platform, specifically by linking Wikidata to Twitter[2] influencers for demographic profiling from Twitter timelines. The third case study (Chapter 5) examines the problem of combining many different distant sources, in particular by inferring harmful content warnings for fan fiction documents from user-assigned free-form tags, and how to limit the increase of noisy labels. The case studies are based on the following five research questions:

Rq 1. **Why are some debaters more persuasive than others?** Research on persuasion is primarily concerned with arguments or debate contributions, not by whom they are delivered. However, some debaters are much more persuasive than others, even when they use similar arguments or frames. It is not well understood how the debater influences the outcome of a debate. In particular, there is a gap in research on the extent to which a debater's style, experience, and argumentative strategies affect whether an argument is persuasive.

Rq 2. **Can author profiling technology be effectively transferred between populations?** Author profiling is the task of predicting an author's attributes, such as demographics or personality, from a given text. The established technology for author profiling is supervised classification, which requires multiple texts for each attribute as training data. User-generated data has been used to collect large datasets for some attributes, such as age and gender, because many users proactively disclose them in their posts or biographies. However, using this strategy with less commonly discussed attributes, such as religious beliefs or education, introduces a bias in that the dataset contains only a certain population, such as highly religious or highly educated individuals. Transferring profiling models from a narrow or biased population to a general population without loss of effectiveness would allow profiling models for many rare attributes.

Rq 3. **Are the posts of a group of fans indicative of the demographic attributes of an influencer?** Many users on social media platforms like Twitter, while active users, post very little text themselves. This lack of text severely limits the effectiveness of author profiling technology, which becomes more accurate the more text a of a user is available. One possible mitigation strategy is to use network homophily, the phenomenon that connections on a platform are more likely to occur between users with shared

[2] We refer to the platform 𝕏 by it's former name Twitter, as much of this work place before the name was changed.

attributes [135]. Homophily results in some attributes being more prevalent within closely connected communities in a social graph, and since author profiling technology exploits statistical similarities in the writing of authors with shared attributes, it can be hypothesized that a user can be profiled using the text of connected users.

Rq 4. **Can trigger warnings be effectively assigned to documents via text classification?** A trigger warning is used to warn people about potentially disturbing content. Such warnings are common for traditional media, examples are the ESRB[3] or the Motion Picture Association rating systems,[4] but they are rarely integrated as a feature into any platform. However, the members of some online communities assign warnings voluntarily by prepending labels to their posts. Detecting triggering content via classification would be a desirable support mechanism for vulnerable individuals or simply for users who want to make more informed decisions about the content they engage with.

Rq 5. **How large is the influence of label noise in the dataset on the evaluation of trigger detection models?** One observation from the classification experiments in (Section 5.3) is that the label noise in the test dataset is likely to bias the evaluation results. The noise stems from the weak supervision heuristics as evaluated in Section 5.2, but sensitive authors often declare mild or implicit mentions of a harmful topic in their work, which leads to documents having a trigger warning without the text to support it.

1.2 Thesis Overview

This thesis presents conceptual contributions on how to create novel, labeled datasets with weak supervision and their practical application to societally relevant research questions. We first systematize weak supervision for labeling user-generated content and, then, present three case studies, each centered around a novel dataset created using weak supervision. Each dataset is then used to answer specific research questions. Table 1.1 shows the publications corresponding to each research question and case study and the chapters in which they are used and partially reprinted.

[3] ESRB rating for video games: `https://www.esrb.org/`

[4] MPA rating for films: `https://www.motionpictures.org/film-ratings/`

Chapter 2 presents our conceptual work on weak supervision for labeling user-generated data. Different types of supervised learning have emerged organically to enable supervised learning under label scarcity, since it is a common problem. We present a unifying view of the three types of supervision used, semi-, self-, and weak supervision, provide comparative definitions and etymological notes, and contrast the scenarios in which these types are effectively applied (Section 2.1). The perspective that emerges is that while weak supervision is well established for data labeling, it is not well understood conceptually, and there is a lack of overview of the design parameters. To fill this gap, we conducted a systematic review of 35 high-quality publications that use weak supervision to create datasets of user-generated content (Section 2.2). The review identifies the main design parameters: the tasks, platforms, data types, the seven types of distant knowledge, the different strategies for linking data and knowledge, and the five common ways of evaluating the linkage. The review shows that the effective and efficient evaluation of weakly labeled datasets is a substantial open research problem.

Chapter 3 presents our first case study on the persuasiveness of debaters on Reddit. While most research on persuasion focuses on arguments, we examine the role of the debater in argumentation. The primary research question is why some debaters are more persuasive than others, and how a debater's style, experience, and argumentative strategies affect the outcome of a debate (Section 3.1).

For this study, we use weak supervision to create a dataset of 3,801 users from Reddit's debate forum (`reddit.com/r/changemyview`). The dataset contains all debate contributions of the users, whether the contribution was persuasive or not, and the persuasiveness of the debater over their active period (Section 3.2).

We find that persuasiveness improves over time for the average debater, that the distribution of 'frames' in debaters' arguments can play an important role in persuasiveness, and that argumentative characteristics based on the presence of certain types of arguments in debaters' text do not seem sufficient to indicate persuasiveness (Section 3.3).

Chapter 4 presents our second case study on influencer profiling on Twitter. Author profiling aims to correlate writing style with an author's personal attributes, such as demographics or personality types, with applications in marketing, forensic linguistics, psycholinguistics, and the social sciences. Influencers are a good population with which to develop profiling technology because they provide many writing samples and many personal attributes are public knowledge (Section 4.1).

For this study, we use weak supervision to create a dataset of 71,706 influencers on Twitter for author profiling. The dataset includes each influencer's full Twitter timeline and up to 239 attributes from Wikidata. The influencer's Twitter account is linked to the corresponding Wikidata page by searching Wikidata for multiple, generated variants of the user's name and handle, and heuristically discarding mismatches based on Wikidata properties.This method achieves a high precision (0.994) with a reasonable recall (0.723), as evaluated on the Twitter handles registered in Wikidata for some authors (Section 4.2).

This dataset is used to investigate whether author profiling technology can be transferred across populations without losing effectiveness. First, we organized a shared task at the 2019 PAN workshop to collaboratively develop state-of-the-art profiling technology for this dataset (Section 4.3). Second, we conducted a series of transfer learning experiments between our influencer dataset and four established profiling datasets with authors from the general population. We find that even though the best models on each dataset are generally the ones trained on it, the loss in effectiveness during transfer is small at about 0.05 F_1 (Section 4.4).

The dataset is also used to investigate whether influencers on Twitter can be profiled using only their fans' posts. For this question, we extended the influencer dataset to include the complete timelines of 10 followers for 2,320 influencers. This extended dataset will again be used in a shared task at the PAN workshop to evaluate whether a classification model can effectively predict an influencer's attributes from the texts written by a group of fans (Section 4.5).

Chapter 5 presents our third case study about assigning trigger warnings to fan fiction stories on the "Archive of our Own". A trigger warning is prepended to documents to warn people about potentially disturbing content. These labels are often requested by online communities, especially by vulnerable groups, but as a new phenomenon they are rarely integrated into mainstream social media platforms (Section 5.1).

For this study, we use weak supervision to create a dataset of 1 million fan fiction works from Archive of Our Own that are labeled with appropriate trigger warnings. The warnings originate from an original 36 warning taxonomy which, since no set of warnings existed for fiction or any other text content, we compiled in a principled manner based on several authoritative sources. Each document's trigger warnings are determined by combining distant knowledge from multiple sources: the free-form content descriptors added by authors, the relationships between content descriptors added by site moderators, and various heuristics and rules for translating the content descriptors into trigger warnings. An evaluation of the effectiveness of the labeling strategy using spot checks shows a near ideal F_1 of 0.95. An additional evaluation using verbatim warnings, i.e., tags containing the terms 'trigger warning', shows a recall of 0.86 across all tags. The remaining percentage points are largely due to different interpretations of what kind of content warrants a warning. In other words, many verbatim warnings signal topics that are not considered 'triggering' by the sources on which our taxonomy is based (Section 5.2).

This dataset enables a series of three experiments into whether trigger warnings can be assigned to documents via classification with sufficient quality. The first set of experiments is limited to warnings for *Violence* with a deep analysis of the important features. We find that (1) classification is highly effective with an F_1 of 0.89, (2) long-document classification techniques are essential because the triggering content can be anywhere in the document, and transformer-based models with truncation often miss it and, (2) it is common for certain topics or characters to be very strongly associated with violence, which may be a notable confounder. The second set of experiments tests multi-label classifiers while varying the characteristics of the labels and dataset, such as granularity, openness, and sampling criteria. The third set of experiments develops state-of-the-art multi-label classifiers for trigger detection in a shared task setting. We find that hierarchical and long-text classifiers are superior, that all models are less effective on rare labels than on common ones and more effective on popular works, suggesting that authors' diligence in tagging their works varies and that this introduces label noise (Section 5.3).

This dataset also enables us to investigate the influence of label noise on the evaluation of the classification models, since (1) prior evidence suggests that the dataset contains label noise from various sources and that it may degrade the model evaluation, and (2) information on label reliability is available, positively when authors explicitly state trigger warnings and negatively via artificially injected label noise. We propose a novel approach called "prompt-based rank pruning" to reduce the amount of noise in the test data by using large language models to determine the amount of textual support for a warning in each document and removing documents with little support. We find that prompt-based rank pruning is effective, as it removes most works with artificially added noisy labels, increases the overall model test score, and reveals differences between models that would otherwise be hidden behind label noise (Section 5.4).

1.3 Publication Record

Ch.	Venue	Type	Pages	Year	Publisher	Ref.
	Case 1: Analyzing the Persuasiveness of Debaters on Reddit					
3.1– 3.3	COLING ***Matti Wiegmann****, Khalid Al-Khatib, Vishal Khanna, and Benno Stein.* *Analyzing Persuasion Strategies of Debaters on Social Media.*	conference	6897–6905	2022	ACL	[235]
	Case 2: Profiling Influencers on Twitter					
4.1– 4.4	ACL ***Matti Wiegmann****, Benno Stein, and Martin Potthast.* *Celebrity Profiling.*	conference	2611–2618	2019	ACL	[239]
4.3	CLEF ***Matti Wiegmann****, Benno Stein, and Martin Potthast.* *Overview of the Celebrity Profiling Task at PAN 2019.*	workshop	–	2019	CEUR-WS	[240]
4.5	CLEF ***Matti Wiegmann****, Benno Stein, and Martin Potthast.* *Overview of the Celebrity Profiling Task at PAN 2020.*	workshop	–	2020	CEUR-WS	[241]
	Case 3: Trigger Warning Assignment					
5.1– 5.3	ACL ***Matti Wiegmann****, Magdalena Wolska, Christopher Schröder, Ole Borchardt, Benno Stein and Martin Potthast.* *Trigger Warning Assignment as a Multi-Label Document Classification Problem.* Nominated for an Outstanding Paper Award.	conference	12113–12134	2023	ACL	[245]
5.3	EMNLP *Magdalena Wolska,* ***Matti Wiegmann****, Christopher Schröder, Ole Borchardt, Benno Stein and Martin Potthast.* *Trigger Warnings: Bootstrapping a Violence Detector for Fan Fiction.*	conference	–	2023	ACL	[247]
5.3	CLEF ***Matti Wiegmann****, Magdalena Wolska, Benno Stein, and Martin Potthast.* *Overview of the Trigger Detection Task at PAN 2023.*	workshop	2523–2536	2023	CEUR-WS	[244]
5.4	CLEF ***Matti Wiegmann****, Benno Stein, and Martin Potthast.* *De-Noising Document Classification Benchmarks via Prompt-based Rank Pruning: A Case Study.*	conference	172–178	2024	Springer	[242]

Table 1.1: Peer-reviewed publications by the author used in this dissertation.

Venue	Type	Pages	Year	Publisher	Ref.
COLING	conference	1498-1507	2018	ACL	[170]
*Martin Potthast, Tim Gollub, Kristof Komlossy, Sebastian Schuster, **Matti Wiegmann**, Erika Patricia Garces Fernandez, Matthias Hagen, and Benno Stein. Crowdsourcing a Large Corpus of Clickbait on Twitter.*					
LNCS	chapter	123–160	2019	Springer	[171]
*Martin Potthast, Tim Gollub, **Matti Wiegmann**, and Benno Stein. TIRA Integrated Research Architecture. In: Information Retrieval Evaluation in a Changing World.*					
ISCRAM	conference	814-824	2019	ISCRAM	[98]
*Jens Kersten, Anna Kruspe, **Matti Wiegmann**, and Friederike Klan. Robust Filtering of Crisis-related Tweets.*					
ISCRAM	conference	872–880	2020	ISCRAM	[236]
***Matti Wiegmann**, Jens Kersten, Friederike Klan, Martin Potthast, and Benno Stein. Analysis of Detection Models for Disaster-Related Tweets.*					
NHESS	journal	1431-1444	2021	COPERNICUS	[237]
***Matti Wiegmann**, Jens Kersten, Hansi Senaratne, Martin Potthast, Friederike Klan, and Benno Stein. Opportunities and Risks of Disaster Data from Social Media: A Systematic Review of Incident Information.*					
IN2WRITING	workshop	39–45	2022	ACL	[243]
***Matti Wiegmann**, Michael Völske, Martin Potthast, and Benno Stein. Language Models as Context-sensitive Word Search Engines.*					
OSSYM	symposium	–	2022	OSSYM	[25]
*Janek Bevendorff, **Matti Wiegmann**, Martin Potthast, and Benno Stein. The Impact of Online Affiliate Marketing on Web Search.*					
ECIR	conference	236–241	2023	Springer	[72]
*Maik Fröbe, **Matti Wiegmann**, Nikolay Kolyada, Bastian Grahm, Theresa Elstner, Frank Loebe, Matthias Hagen, Benno Stein, and Martin Potthast. Continuous Integration for Reproducible Shared Tasks with TIRA.io.*					
EAAI	conference	15807–15815	2023	ACM	[62]
*Theresa Elstner, Frank Loebe, Yamen Ajjour, Christopher Akiki, Alexander Bondarenko, Maik Fröbe, Lukas Gienapp, Nikolay Kolyada, Janis Mohr, Stephan Sandfuchs, **Matti Wiegmann**, Jörg Frochte, Nicola Ferro, Sven Hofmann, Benno Stein, Matthias Hagen, and Martin Potthast. Shared Tasks as Tutorials: A Methodical Approach*					
DH	conference	357–359	2023	ADHO	[151]
*Andreas Niekler, Magdalena Wolska, Marvin Thiel, **Matti Wiegmann**, Benno Stein, and Manuel Burghard. Marco Polo's Travels Revisited: From Motion Event Detection to Optimal Path Computation in 3D Maps*					
ECIR	conference	-	2024	Springer	[26]
*Janek Bevendorff, **Matti Wiegmann**, Martin Potthast, and Benno Stein. Is Google Getting Worse? A Longitudinal Investigation of SEO Spam in Search Engines*					
WOWS	workshop	-	2024	CEUR-WS	[238]
***Matti Wiegmann**, Jan Heinrich Reimer, Maximilian Ernst, Martin Potthast, Matthias Hagen. and Benno Stein. A Mastodon Corpus to Evaluate Federated Microblog Search*					
CHIIR	conference	-	2024	Springer	[27]
*Janek Bevendorff, **Matti Wiegmann**, Martin Potthast, and Benno Stein. Product Spam on YouTube: A Case Study*					

Table 1.2: Peer-reviewed publications of the author not used in this thesis, in chronological order.

<table>
<tr><th>Venue</th><th>Type</th><th>Pages</th><th>Year</th><th>Publisher</th><th>Ref.</th></tr>
<tr><td>CLEF</td><td>workshop</td><td>402–416</td><td>2019</td><td>Springer</td><td>[48]</td></tr>
<tr><td colspan="6">*Walter Daelemans, Mike Kestemont, Enrique Manjavacas, Martin Potthast, Francisco Rangel, Paolo Rosso, Günther Specht, Efstathios Stamatatos, Benno Stein, Michael Tschuggnall,* ***Matti Wiegmann****, and Eva Zangerle. Overview of PAN 2019: Bots and Gender Profiling, Celebrity Profiling, Cross-domain Authorship Attribution and Style Change Detection.*</td></tr>
<tr><td>CLEF</td><td>workshop</td><td>508–516</td><td>2020</td><td>Springer</td><td>[23]</td></tr>
<tr><td colspan="6">*Janek Bevendorff, Bilal Ghanem, Anastasia Giachanou, Mike Kestemont, Enrique Manjavacas, Ilia Markov, Maximilian Mayerl, Martin Potthast, Francisco Rangel, Paolo Rosso, Günther Specht, Efstathios Stamatatos, Benno Stein,* ***Matti Wiegmann****, and Eva Zangerle. Overview of PAN 2020: Authorship Verification, Celebrity Profiling, Profiling Fake News Spreaders on Twitter, and Style Change Detection.*</td></tr>
<tr><td>ECML-PKDD</td><td>workshop</td><td>–</td><td>2020</td><td>CEUR-WS</td><td>[105]</td></tr>
<tr><td colspan="6">*Konstantin Kobs, Martin Potthast,* ***Matti Wiegmann****, Albin Zehe, Benno Stein, and Andreas Hotho. Towards Predicting the Subscription Status of Twitch.tv Users – ECML-PKDD ChAT Discovery Challenge 2020.*</td></tr>
<tr><td>CLEF</td><td>workshop</td><td>–</td><td>2020</td><td>CEUR-WS</td><td>[99]</td></tr>
<tr><td colspan="6">*Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff,* ***Matti Wiegmann****, Efstathios Stamatatos, Benno Stein, and Martin Potthast. Overview of the Cross-Domain Authorship Verification Task at PAN 2020.*</td></tr>
<tr><td>CLEF</td><td>workshop</td><td>567–573</td><td>2021</td><td>Springer</td><td>[21]</td></tr>
<tr><td colspan="6">*Janek Bevendorff, Berta Chulvi, Gretel Liz De La Peña Sarracén, Mike Kestemont, Enrique Manjavacas, Ilia Markov, Maximilian Mayerl, Martin Potthast, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, Benno Stein,* ***Matti Wiegmann****, Magdalena Wolska, and Eva Zangerle. Overview of PAN 2021: Authorship Verification, Profiling Hate Speech Spreaders on Twitter, and Style Change Detection.*</td></tr>
<tr><td>CLEF</td><td>workshop</td><td>1743–1759</td><td>2021</td><td>CEUR-WS</td><td>[100]</td></tr>
<tr><td colspan="6">*Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff,* ***Matti Wiegmann****, Efstathios Stamatatos, Benno Stein, and Martin Potthast. Overview of the Cross-Domain Authorship Verification Task at PAN 2021.*</td></tr>
<tr><td>CLEF</td><td>workshop</td><td>382–394</td><td>2022</td><td>Springer</td><td>[22]</td></tr>
<tr><td colspan="6">*Janek Bevendorff, Berta Chulvi, Elisabetta Fersini, Annina Heini, Mike Kestemont, Krzysztof Kredens, Maximilian Mayerl, Reyner Ortega-Bueno, Piotr Pezik, Martin Potthast, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, Benno Stein,* ***Matti Wiegmann****, Magdalena Wolska, and Eva Zangerle. Overview of PAN 2022: Authorship Verification, Profiling Irony and Stereotype Spreaders, and Style Change Detection.*</td></tr>
<tr><td>CLEF</td><td>workshop</td><td>459–481</td><td>2023</td><td>Springer</td><td>[19]</td></tr>
<tr><td colspan="6">*Janek Bevendorff, Mara Chinea-Ríos, Marc Franco-Salvador, Annina Heini, Erik Körner, Krzysztof Kredens, Maximilian Mayerl, Piotr Pęzik, Martin Potthast, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, Benno Stein,* ***Matti Wiegmann****, Magdalena Wolska, and Eva Zangerle. Overview of PAN 2023: Authorship Verification, Multi-Author Writing Style Analysis, Profiling Cryptocurrency Influencers, and Trigger Detection.*</td></tr>
<tr><td>CLEF</td><td>workshop</td><td>-</td><td>2024</td><td>CEUR-WS</td><td>[20]</td></tr>
<tr><td colspan="6">*Janek Bevendorff,* ***Matti Wiegmann****, Jussi Karlgren, Luise Dürlich, Evangelia Gogoulou, Aarne Talman, Efstathios Stamatatos, Martin Potthast, and Benno Stein. verview of the “Voight-Kampff” Generative AI Authorship Verification Task at PAN and ELOQUENT 2024*</td></tr>
<tr><td>CLEF</td><td>workshop</td><td>-</td><td>2024</td><td>Springer</td><td>[11]</td></tr>
<tr><td colspan="6">*Abinew Ali Ayele, Nikolay Babakov, Janek Bevendorff, Xavier Bonet Casals, Berta Chulvi, Daryna Dementieva, Ashaf Elnagar, Dayne Freitag, Maik Fröbe, Damir Korenčić, Maximilian Mayerl, Daniil Moskovskiy, Animesh Mukherjee, Alexander Panchenko, Martin Potthast, Francisco Rangel, Naquee Rizwan, Paolo Rosso, Florian Schneider, Alisa Smirnova, Efstathios Stamatatos, Elisei Stakovskii, Benno Stein, Mariona Taulé, Dmitry Ustalov, Xintong Wang,* ***Matti Wiegmann****, Seid Muhie Yimam, and Eva Zangerle. Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification*</td></tr>
</table>

Table 1.3: Workshops co-organized by the author, in chronological order.

2
Weak Supervision for Labeling User-generated Content

The classification problems discussed in Chapter 1 are primarily approached as learning problems. The literature [82, 90] distinguishes between supervised learning, which requires a set of (input, output) tuples to learn from, and unsupervised learning, where only the inputs are known and groups with similar properties are discovered. Classification problems, where all eligible classes are known a priori, are typically modelled for supervised learning.

There are three different variants of supervised learning for classification when labeled data is scare: semi-supervised, self-supervised, and weakly supervised learning.[1] Since all of these types are suitable solutions to the classification problems considered, in Section 2.1 we offer a comparison of the different types of supervision, how they are related, how they are used in related work, and what sets weakly supervised learning apart regarding the classification of user-generate content.

Even though weakly supervised learning is often used to label user-generated data, there is little systematic understanding of when it works well, in particular (1) the situations in which weakly supervised learning can and has been used to label data, (2) what kind of external knowledge is appropriate, (3) what trade-offs are to be expected and what mitigation measures, such as evaluation or noise reduction, need to be taken. Without

[1] The literature uses *weakly supervised learning* and *weak supervision* synonymously.

this systematic understanding, it is difficult to identify the design space of weakly supervised learning, i.e., what are the options for creating a dataset and what are the research gaps. Therefore, in Section 2.2, we present a systematic review of related work that uses weakly supervised learning to create resources from user-generated content.

2.1 A Unifying View on the Types of Supervision

We consider five types of supervision: supervised, semi-, self-, and weakly supervised learning, and unsupervised learning:

1. **Supervised learning** takes a set of (input, output) tuples and, from those, learns to determine the corresponding output for a given input.[2] Supervised learning requires a corresponding output for each input in the dataset and it is assumed that this output is reliable, such as from a traditional annotation process.

2. **Semi-supervised learning** is a subtype of supervised learning that assumes a labeled set of (input, output) tuples and uses a second, unlabeled set of inputs to improve the learning, especially when the labeled set is small.

3. **Self-supervised learning** is a subtype of supervised learning that uses an unlabeled set of inputs and determines the output by exploiting a part of the input to use as the output, especially for autoregressive training.

4. **Weakly supervised learning** is a subtype of supervised learning that uses an unlabeled set of inputs and derives the output from some distant knowledge, such as databases or heuristics.

5. **Unsupervised learning** takes a set of inputs and discovers groups with similar properties between while the output is generally not known.

All five types of supervision are learning tasks with the goal of training a model function $y : \mathbf{X} \to C$ that determines, given an input $\mathbf{x} \in \mathbf{X}$ the corresponding output $c \in C$. Here, the input $\mathbf{x}$ is a feature vector of fixed length with values of numeric or categorical type, and C is a finite set with few

[2] A note on terminology: *inputs* and *features* are used synonymously in this work and *label* is used instead of *output* when referring to classification problems.

elements, for example $\{0, 1\}$ in binary classification problems. The model function is fitted and tested on a multiset D containing a number n of examples, which are generally (input, output) tuples. This multiset D differs between the types of supervision.

Hastie et al. [82] note that *supervised learning* problems consider "the presence of the outcome variable to guide the learning process", and more specifically, that (1) for each output there is a measured input, (2) that the input has some influence on the output, and (3) "the goal is to use the inputs to predict the values of the outputs" [82]. This means that supervised methods optimize the fit $y : X \rightarrow C$ of the model function $y(\cdot)$ on a multiset D of inputs $\mathbf{x}$ with known, corresponding labels c:

$$\text{Supervised Data:} \quad D = \{(\mathbf{x}_1, c_1), \ldots, (\mathbf{x}_n, c_n)\} \subseteq X \times C \tag{2.1}$$

In contrast, Hastie et al. [82] note that *unsupervised learning* problems consider "only the features and have no measurements of the outcome", which here refers to the class, and that "the task is rather to describe how the data are organized or clustered". That is, unsupervised methods operate on a set of inputs and discover groups with similar properties between those inputs:

$$\text{Unsupervised Data:} \quad D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subseteq X \tag{2.2}$$

This means that the output Y is generally not know for unsupervised problems, which is not of further interest for this work.

Inputs The input is a representation, i.e., a feature vector $\mathbf{x} \in \mathbf{X}$ of the data object $o \in O$ from the set of all data objects to be classified. In the case of textual user content, which is the primary type of data used in this work, these features can be manually engineered or learned via representation learning. Engineered features often capture abstract properties of the text content, such as weighted counts of selected words or n-grams, the linguistic structure, such as Part-of-Speech tag or phrase structure n-gram frequency, or task-specific features that are tailored to a specific task, such as, for spam classification, the credibility of the source or, for clickbait classification, if a post starts with a number [170]. Learned features, such as word2vec vectors [137] or the output of pre-trained transformer models like BERT [56], represent the text content in a latent space, where a small distance between two feature vectors typically indicates similarity. Engineered features are commonly used with traditional linear, tree, kernel, or

Bayesian-based learning methods, while learned features are the standard in deep learning.

Label Functions Supervised learning in general, as shown in Equation 2.1, assumes that the output $c \in C$ for each example is both known and reliable, as determined by an ideal labeling function $\gamma(o), \gamma(o) \in C, o \in O$, so that

$$\text{Ideal Supervision:} \quad D = \{(\mathbf{x}_1, \gamma(o_1)), \dots, (\mathbf{x}_n, \gamma(o_n))\} \tag{2.3}$$

An example of an ideal labeling function is the traditional annotation process, where human annotators are presented with the data objects o, say an email, and determine the appropriate label, spam or not spam, based on their understanding of the problem and their perception of the objects. Supervised learning with an ideal labeling function is the preferred setting to optimize the performance of the model function $y(\cdot)$.

As discussed in Chapter 1, the requirement of completeness and reliability of labels often does not hold in practice, especially for user data from social media platforms. Consequently, semi-, self-, and weakly supervised learning either relax these criteria by adapting non-ideal labeling functions or by modifying the model function to learn from examples with unknown labels.

2.1.1 Semi-supervised Learning

Semi-supervised learning, as described by Chapelle et al. [40], refers to all methods that use both labeled and unlabeled data to fit a model function $y(\cdot)$:

$$\text{Semi-Supervision:} \quad D = \{(\mathbf{x}_1, c_1), \dots, (\mathbf{x}_k, c_k), \mathbf{x}_{k+1}, \dots, \mathbf{x}_n\}. \tag{2.4}$$

This means that, unlike unsupervised learning, the label set C is already defined and some labeled data is available, e.g. to fit an initial model function.

In natural language processing, according to Søgaard [208], semi-supervised methods can be either wrapper methods around supervised algorithms, methods that use unsupervised learning to augment the supervised dataset, and nearest neighbor methods. Wrapper methods are the most commonly used and include self-training and its variants such as co-training or expectation maximization (EM). Here, a supervised classifier is trained on the labeled data, then used to classify the unlabeled data, and finally trained on both. Biemann [29] introduces self-training as *bootstrapping*, which is the more popular term in natural language processing:

> "Bootstrapping starts with a few training examples, trains a classifier, and uses thought-to-be positive examples as yielded by this classifier for retraining. As the set of training examples grows, the classifier improves, provided that not too many negative examples are misclassified as positive, which could lead to deterioration of performance." [29]

Jurafsky and Martin [93] mention several typical applications of semi-supervised learning under the term *bootstrapping*: for lexicon construction by labeling a set of seed words and propagating the labels according to a similarity measure such as cosine in an embedding space, for slot filling in dialog systems to discover new utterances [212], and for relation extraction.

Semi-supervised and weakly supervised learning (introduced below) overlap in some cases when external information besides the labeled seed data is used for expansion, as in relation extraction. Here, semi-supervised learning starts with a seed set of relation triplets with two entities and their relation. These seeds are used to discover anchor text, for example using Hearst patterns [83], in the sentences of a corpus, which is the text between or around two entities of a source triplet. The anchor text is assumed to indicate the relation of the source triplet, and for each sentence with this anchor text and two unknown entities, a new triplet is created indicating that the unknown entities fulfill the relation indicated by the anchor text. So the iterative expansion of the set of labeled data is a semi-supervised strategy, and the use of external patterns to facilitate the expansion is a weakly supervised technique.

Semi-supervised learning comes with the trade-off of introducing label noise, or what Jurafsky and Martin [93] call semantic drift, that needs correction. Whenever the model function erroneously introduces a new rule or label, this error changes the model function and makes future misclassification more likely.

2.1.2 Self-supervised Learning

Self-supervised learning refers to all methods that use labeling functions $\beta(o), \beta(o) \in C, o \in O$ that derive the label from the data object o by modifying or splitting of parts.

$$\text{Self-Supervision:} \quad D = \{(\mathbf{x}_1, \beta(o_1)), \ldots, (\mathbf{x}_n, \beta(o_n))\}. \tag{2.5}$$

This strategy allows the creation of very large training datasets, and the most prominent examples are pre-trained foundation models: word embedding models such as word2vec [137] predict the center word in a window given the context words (or vice versa), autoregressive language models predict the last word in a sequence given the preceding sequence, and image generation models such as stable diffusion [195] modify an image, for example by adding noise, and predict the original given the modified version. Jurafsky and Martin also mention sentence coherence as an application, where a sentence is scrambled, for example by permuting the words, and the original sentence becomes the target.

Self-supervised learning and weakly supervised learning overlap in some cases where the label is extracted from the data object, but the decision of when to extract is based on external knowledge, for example when a post contains a `#Sarcasm` hashtag and is subsequently labeled as such [53]. Compared to semi- and weakly supervised learning, erroneous labels are less of an issue for self-supervised learning since the inferred output is (part of) the original data object. However, self-supervised learning is also limited in its application to the problems of interest in this work, besides being widely used for pre-training large models.

2.1.3 Weakly Supervised Learning

Weakly supervised learning refers to all methods that use labeling functions $\delta(o), \delta(o) \in C, o \in O$ that derive the label for a data object using an external or distant source of information:

$$\text{Weak Supervision:} \quad D = \{(\mathbf{x}_1, \delta(o_1)), \ldots, (\mathbf{x}_n, \delta(o_n))\}. \tag{2.6}$$

Eisenstein [59] describes this concept under the term *distant supervision*, where "noisy labels are generated from an external resource" [59]. The concept and its applications are equivalent, but distant supervision as a term is more popular in natural language processing. Jurafsky and Martin [93] traces the origin to Mintz et al. [139], who first used the term distant supervision when using relation triples from Freebase (now Wikidata) to annotate a corpus for relation extraction. They in turn note that similar ideas had appeared in earlier systems in bioinformatics by Morgan et al. [145] as "weakly labeled data" and by Snow et al. [207] when using WordNet to extract hypernym (is-a) relations between entities. Although distant supervision is the more common term in natural language processing, and has been adopted by the social media processing community [61], weakly supervised learning is more common in the machine learning community.

In more recent work, Ratner et al. [188] attempt to unify and generalize different weakly supervised learning methods in a framework called *programmatic weak supervision*, which allows the combination of various weak supervision labeling functions to label the same dataset. Zhang et al. [256] present a follow-up survey on programmatic weak supervision and specifically categorize the types of common labeling functions: (1) rules and heuristics, (2) existing knowledge (databases, classifiers, or other tools), (3) and noisy human sources such as crowd sourcing. The latter, noisy human sources, is excluded from weakly supervised learning by almost all other definitions, so we exclude it as well. Zhou [259] also present a survey on weakly supervised learning, where they distinguish three subtypes: (1) *incomplete supervision*, where a subset of the training data is labeled while the other data remains unlabeled and which is equivalent to semi-supervised learning, (2) *inexact supervision*, where only coarse-grained labels are given and fine-grained labels are inferred, as in the image segmentation example in Chapter 1, and (3) *inaccurate supervision*, where the given labels are noisy and not considered to be ground truth.

Weak supervision is widely used in natural language processing, especially to create resources in various areas: For example, in linguistic structure parsing, Li et al. [118] uses weakly supervised learning by leveraging Wiktionary for part-of-speech annotations, in information extraction, Mintz et al. [139] using Freebase to annotate relations, in word sense disambiguation, Navigli and Ponzetto [149] using WordNet and Wikipedia to construct BabelNet, Wang et al. [226] use heuristics to transfer document-level sentiment labels to the sentence level. Lin et al. [119] use rules such as "the statement contains a number" to compile common sense reasoning questions for BERT pre-training, and Hedderich et al. [84] investigate weakly supervised learning for various applications in low-resource languages.

Compared to other types, weakly supervised learning makes it possible to collect a large amount of labeled data without an initial labeled set, with a wide range of applications, and without the semantic drift that occurs with semi-supervision. The disadvantages of other types of supervision are that the size and quality of the resulting dataset depends on both the distant resource and the supervision method, both of which can be sources of label noise or bias.

2.1.4 Conclusion

This section outlines the different types of supervision and that they all operate on a dataset consisting of tuples of data objects and labels. In weakly supervised learning, these labels are determined by a label function $\delta(o)$ that uses some form of distant knowledge. The label function $\delta(o)$ distinguishes weakly supervised learning from other forms of supervised learning: supervised learning in general only assumes that the labels come from a reliable function $\gamma(o)$, such as a traditional annotation process, and self-supervised learning determines the labels via a label function $\beta(o)$ that splits the data object into data and label parts. Semi-supervised learning is an exception in that it uses knowledge from labeled and unlabeled data to extend the dataset while the specific label function is not relevant.

2.2 Weak Supervision in User Content Labeling

Weak supervision has been used extensively to derive labels for user content. Related works study a broad range of (1) tasks, such as sentiment classification [52], personality profiling [32], and topic modeling [142], (2) genres, such as microblog posts from Twitter, classified individually or as timelines, or forum threads from Reddit, (3) sources of distant knowledge, such as knowledge bases, Wikipedia, metadata from other platforms, or lists that map content features to labels, and (4) methodologies for connecting the distant knowledge to the user-generated content, such as simple lookups, multi-step heuristics, or graph propagation.

Despite its widespread use, there is no systematic analysis and comparison of weakly supervised learning in social media and user-generated content analysis, beyond the general overview in Section 2.1.3. So it is difficult to grasp the principles of weak supervision strategies. For example, what parts of the data, such as hashtags, or metadata, such as links in user profiles, can be exploited with little error? What external knowledge can be linked up with the user content? The lack of such a systematic analysis can lead to errors in the application of weak supervision, resulting in smaller, noisier, and poorly evaluated datasets as elaborated in Chapter 1.

As part of our work, we conduct a systematic review of weak supervision methods used to derive labels for user content. The review is based on 303 related and highly relevant papers from 26 leading venues in four areas of empirical computer sciences that frequently use user content, retrieved from Semantic Scholar. We manually filter the documents to identify the publications that use weak supervision to construct a dataset using user-generated content, and review the remaining 35 publications in detail. In particular, we ask the following questions:

1. Which tasks in natural language processing and computational social science effectively use weak supervision?
2. What platforms and data are used to create the datasets and how large are the resulting dataset?
3. What sources of distant knowledge are used, and how are they linked to the data?
4. How much label noise do the resulting datasets contain, and what methods are used to evaluate or mitigate label noise?

2.2.1 Review Method

We collect a set of publications that use weak supervision to derive labels for user-generated content, contain high quality work, and cover the important tasks, platforms, data types, sources of distant knowledge, and evaluation strategies. However, we also limit the number of publications to make an in-depth analysis feasible.

The review set of papers is constructed in a two-step process: (1) an initial retrieval, which collects 303 papers using high-recall queries with many hits and (2) a manual filtering, where the 35 most relevant papers are selected using three coarse inclusion criteria. The use of a multi-stage review method is necessary because both terms "weak supervision" and "social media" may not always be mentioned explicitly. For example, authors may not mention the term 'weak supervision' out of unawareness, by using an alternative phrase, or by referring to another type of supervision as discussed in Section 2.1. Consequently, a single-stage retrieval where any doc-

Venue		**Relevant Publications**
Natural Language Processing		
`ACL`	Association for Computational Linguistics	47
`EMNLP`	Empirical Methods in Natural Language Processing	37
`NAACL`	North American Chapter of the ACL	17
`COLING`	Computational Linguistics	11
`EACL`	European Chapter of the ACL	6
`LREC`	Language Resources and Evaluation	6
`CONLL`	Conf. on Computational Natural Language Learning	3
`IJCNLP`	Int. Joint Conf. on Natural Language Processing	1
Information Retrieval and Data Mining		
`WSDM`	Web Search and Data Mining	16
`SIGIR`	Conf. on Research and Development in Inf. Retrieval	13
`CIKM`	Conf. on Information and Knowledge Management	12
`ECIR`	European Conf. on Information Retrieval	2
`CLEF`	Conf. and Labs of the Evaluation Forum	0
`PKDD`	Data Mining and Knowledge Discovery	0
`VLDB`	Very Large Data Bases Conf.	0
Machine Learning and AI		
`AAAI`	Conf. on Artificial Intelligence	30
`NeurIPS`	Neural Information Processing Systems	22
`ICLR`	International Conf. on Learning Representations	21
`ICML`	International Conf. on Machine Learning	14
`ECML`	European Conf. on Machine Learning	0
`PMLR`	Proceedings of Machine Learning Research	0
`TPAMI`	Trans. on Pattern Analysis and Machine Intelligence	0
Web and Social Media		
`WWW`	The Web Conference	19
`ICWSM`	International Conference on Web and Social Media	15
`ASONAM`	Advances in Social Networks Analysis and Mining	11
`SNAM`	Social Network Analysis and Mining	0

Table 2.1: Venues for high quality, peer reviewed research that are included in this systematic review. The venues are manually curated based on their relevance for the four fields of interest for the review. According to these venues, citation count, and publication year, 303 publications are considered relevant.

ument matching a keyword query is considered relevant would find very few publications, and we instead use a high-recall first stage followed by a high-precision stage.

Initial Retrieval In the initial retrieval step, we collect 303 high quality publications from the leading natural language processing and computational social science venues. The retrieval is based on 12 queries sent to SemanticScholar, the results of which are then filtered by time and impact.

Eight of these 12 queries search for publications at the intersection of weak supervision and user-generated content; the remaining four search at the intersection of datasets and user-generated content. All 12 queries are *AND* queries, where all terms must be present in a document to yield a match. The eight supervision queries contain exactly one platform phrase from `{social media, twitter, reddit, facebook}` and one supervision phrase from `{weak supervision, distant supervision}`. The four dataset queries again contain exactly one platform phrase and `dataset`. Separating the three areas into two sets of intersections is necessary because of the aforementioned problems of term ambiguity: there are only few papers that mention a term from all three areas, even if they belong to the the intersection set.

We retrieve the results of each query using the SemanticScholar Academic Graph API's relevance search endpoint,[3] which uses term weighting based on the title, abstract, and body content of the publication [103]. SemanticScholar is preferable to using the publishers' sites because computer science publishing is very fragmented and most publishers and venues use a different publishing system which often does not support a systematic search. SemanticScholar's ranked search is preferable to other aggregators because it matches terms in the title, abstract, and body of a publication, unlike the Google Scholar search which takes references into account and produces many false positives because of it and unlike the SemanticScholar API's bulk search endpoint,[4] which only considers title and abstract.

The additional filtering criteria are a curated list of venues, citation count, and publication year. All 26 eligible venues are shown in Table 2.1 and include a manual selection of venues for high-level peer-reviewed research concerning weak supervision to label user-generated content in the four larger fields of interest to this thesis: *Natural Language Processing*, *Information Retrieval and Data Mining*, *Machine Learning and AI*, and *Web and Social Media*. Only publications with at least 100 citations are considered, as an indicator of relevance and quality of the work. Publications released before 2008 were discarded since only few social media platforms existed before that year. In the end, the supervision queries retrieved 193 publications and the dataset queries 110 publications before deduplication for annotation.

[3] `https://api.semanticscholar.org/api-docs/graph#tag/Paper-Data/operation/get_graph_paper_relevance_search`

[4] `https://api.semanticscholar.org/api-docs/graph#tag/Paper-Data/operation/get_graph_paper_bulk_search`

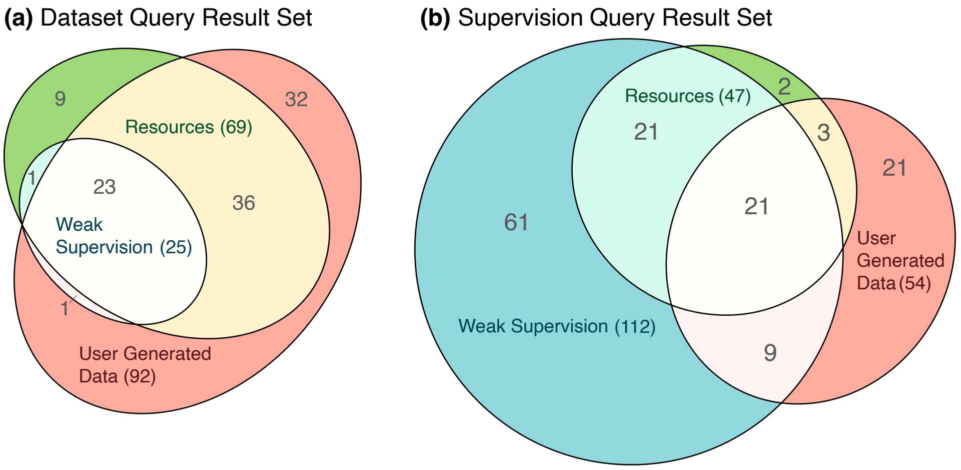


Figure 2.1: Publications returned by the dataset queries (**a**) and the supervision queries (**b**). The Euler diagram shows how many publications satisfy any or all of the three preconditions: the publication presents a newly created *Resource*, labels *User-generated Data*, and uses *Weak Supervision* for the labeling, as described in Section 2.2.1

Manual Filtering In the manual filtering step we check three preconditions for the research to be included in this review: (1) the research present the creation or modification of a resource or dataset, (2) the data objects are user-generated content from a social media platform, excluding sites where users can only read or comment, like news sites with a comment section, or where the content is collaborative, such as Wikipedia, and (3) the research uses a weak supervision method to label the data object. Figure 2.1 shows the result of the precondition checks. As expected, many of the retrieved works fulfil only one or two preconditions. However, about 10-20% of the retrieved publications, or 35 after deduplication, fall into the relevant intersection of the three fields. In the following, we systematically evaluate these 35 relevant publications in order to find answers to the initial research questions posed in Section 2.2.

The review methodology selects a set of highly cited publications from prestigious venues as this is both feasible for a systematic review and lends credibility to the review. However, the review methodology does not allow for any claims of completeness due to the aforementioned limitations of the computer science literature, where the collection of eligible publications across all areas is complex and extensive in scope. This means that although the review results reflect high quality work, it must be expected that they are not comprehensive and that conclusions for the larger field should be drawn with caution.

2.2.2 Results

The results of the systematic review consist of the extraction and systematic analysis of the following seven aspects of a weak supervision labeling process for each of the 35 relevant publications:

1. The *task* being studied and the scientific community interested in that task. We also annotate the task for all 303 publications from the initially retrieved set for comparison.
2. The *platform* whose content is used as data objects for the dataset, such as Twitter or Facebook.
3. The *data* object that is extracted, such as the posted text or image.
4. The *number* of data objects in the labeled dataset.
5. The type of *distant knowledge* used. We distinguish between seven categories of distant knowledge found in the relevant publications: curated list, database, web data, metadata, distant metadata, computed metadata, and classifiers.
6. The individual *weak supervision strategy* that is used to link the distant knowledge with the data objects. We analyze these strategies qualitatively, since they are highly specific to the individual case.
7. The *evaluation strategy* that is applied (or not) to evaluate the quality and usefulness of the created dataset. We distinguish five evaluation strategies found in the relevant publications: spot checks, weak labels, annotated data, models, and no evaluation.

The results (see Tables 2.2, 2.3, 2.4, and 2.5) are discussed in the following.

Tasks

Weak supervision is used in a variety of tasks and by a variety of communities. However, the most frequently studied areas that use user-generated content are text classification and computational social science. Figure 2.2 shows the 43 tasks identified in the publications found in the initial retrieval step. Most of the tasks in the sample stem from natural language processing, with the most common communities being text classification, information extraction, and applications. Tables 2.3, 2.4, and 2.5 show the tasks addressed by the publications selected as relevant for this review, which mostly address text classification and computational social science. The differences are explained by many publications not constructing a dataset, but

Source	Platform	Data Type	Knowledge	Count	Evaluation
[53](2010)	Twitter	Text Post	Curated List	75K	Spot Checks
[16](2010)	Flickr	Image Post	Database	870K	Model
[52](2010)	Twitter	Text Post	Curated List	1.5K	Annotated Data
[61](2010)	Twitter	User	Metadata	9.5K	Model
[76](2010)	LiveJournal	Text Post	Classifier	20M	Annotated Data
[121](2011)	Twitter	Token	Web Data	63K	Model
[130](2011)	Twitter, Friendfeed	User	Metadata	156K	–
[191](2011)	Twitter	Text Post	Database	?	Annotated Data
[228](2011)	Twitter	Token	Curated List	29K	Spot Checks
[175](2012)	Twitter	Text Post	Curated List	1M+	Annotated Data
[193](2012)	Twitter	User	Metadata	450K	Model
[199](2012)	Twitter	User	Metadata	1.2M	Model
[220](2012)	Twitter	Token	Comp. Metadata	1K+	Model
[250](2012)	Twitter	Text Post	Curated List	?	Model
[80](2013)	Twitter	Text Post	Distant Metadata	35K	–
[94](2013)	Twitter, Foursquare	User	Metadata	2.6M 2.7M	Model
[112](2013)	Twitter	Text Post	Curated List	24K	Model
[131](2013)	Twitter	User	Curated List	25K	Annotated Data
[116](2014)	Twitter	Token	Distant Metadata	300K	Model
[215](2014)	Twitter	Text Post	Curated List	?	Annotated Data
[47](2015)	Twitter	User Group	Database	1.5K	Model
[8](2016)	Twitter	Text Post	Database	200K	Model
[30](2016)	Twitter	User	Database	2.8M	Model
[146](2016)	Twitter	User	Curated List Metadata	95K 6K	Model
[192](2016)	YouTube	Video Post	Distant Metadata	14K	Model
[55](2017)	Twitter	Text Post	Curated List	200M	Model
[147](2017)	Twitter	Dialog	Curated List	250K	Annotated Data
[255](2017)	Twitter	User	Curated List	45K	Model
[227](2017)	Twitter Facebook	User	Distant Metadata	7K 8K	Model
[12](2018)	Twitter	User	Curated List	29K	Weak Labels
[85](2018)	Twitter	Text Post	Curated List	400K	Annotated Data
[57](2018)	Twitter Reddit	Text Post	Classifier	2M 7.3M	Model
[248](2018)	Twitter	Text Post	Database	3.6K	Model
[231](2019)	Twitter	User	Curated List	50M	Annotated Data
[252](2020)	Twitter	User	Database	3K	Model
Ch. 4(2019)	Twitter	User	Database	71K	Weak Labels
Ch. 3(2022)	Reddit	User	Metadata	3.8K	Model
Ch. 5(2023)	AO3	Text Post	Curated List Metadata	1M	Spot Checks Weak Labels

Table 2.2: Works considered in the survey. Shown are the Platform of the data that is labeled, the data type, the kind of distant knowledge used, the number of data objects labeled, and the method of evaluation. A question mark indicates an unclear statement in the publication.

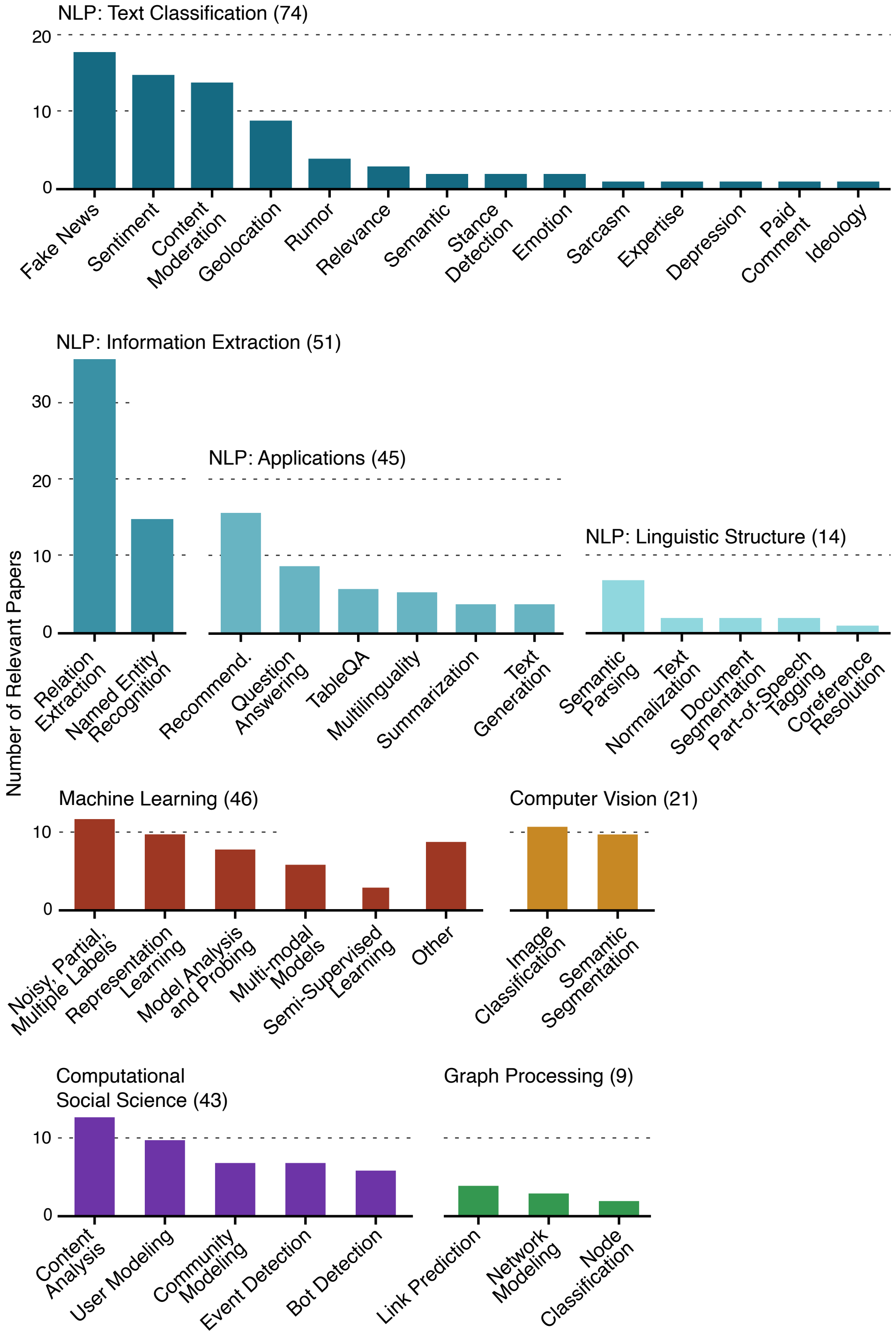


Figure 2.2: Overview of the task tackled by the publications collected by the initial retrieval step, grouped by the communities that are interested in those tasks.

using weak supervision as part of an algorithm. This is popular in computer vision, as in the image segmentation example in Chapter 1, in information extraction, where *distant supervision* has become a reference term for a family of relation extraction algorithms, and in machine learning, where weak supervision is integrated in the model training loop.

The three case studies conducted in this thesis cover two novel tasks and one established task. Both trigger detection and persuasiveness analysis have not previously been studied with weakly labeled data. They are examples of cases where weak supervision allows the study of a task for which the data would otherwise be missing. Author profiling is often studied using weakly labeled data, but for a smaller set of author attributes such as place of residence and ethnicity. However, our weak supervision strategies allow the study of many rare and understudied attributes.

Platforms

The dominant platform across all publications in the reviews is the microblogging service Twitter, which is used in 32 publications (91%). Seven other platforms are used, some in addition to Twitter: the image-sharing platform Flickr, the blogging service LiveJournal, the (now defunct) feed aggregator Friendfeed, the location-discovery service Foursquare, the video platform YouTube, the social network Facebook, and the online forum Reddit. Several factors may explain this imbalance: First, the review only considers venues that mainly deal with textual data, so platforms with a focus on image (Instagram) or video data (YouTube) are less represented. Second, the review considers publications with many citations, and thus may be biased toward popular platforms, with Twitter being among the most popular during the review period. Finally, Twitter provided a free, public API that allowed search access to the tweet archive and access to the last 3,200 tweets of each known user. Collecting data from other popular text-based platforms, such as Facebook and Reddit, is more complicated, if not impossible. At the time of this writing, Facebook prohibits access to its data, and Twitter and Reddit's APIs have become much more restrictive in terms of access, pricing, and redistribution policies.

The three case studies use data from three different platforms. The first study on persuasiveness analysis uses Reddit because of its unique metadata on the persuasiveness of comments. The second study on author profiling uses Twitter for the same reason: there are many authors, their text is public, and the API makes it easy to access this data. However, Twitter is not a prerequisite for the linking strategy; it works with any platform that tags influencers in some way. The third study on trigger warnings uses Archive of Our Own, a platform where users share fan fiction stories, which is not used in any other study. The linking strategy is also tailored to this platform and is not applicable elsewhere.

Data and Dataset Size

The most common data objects are single text posts, used by 14 publications, and user-level collections of posts, used by 13 publications (about 40%). Text posts are mostly Twitter microblogs of up to 280 characters or Reddit posts of any length. An example is geolocation prediction, where each text post is tagged with a location, such as a city, and the task is to predict the location given the text [193]. User-level collections always refer to Twitter user "timelines", the set of the most recent up to 3,200 tweets of a user, where a label is usually assigned to all posts collectively. An example of this is user profiling, where each user is labeled with an attribute, and the task is to predict the attribute given all posts by that user [231]. The less common data types are tokens, for example labeling noisy word forms with a normalized version [121] or labeling named entities [191], images, for example by assigning them the ID of the event they were taken at [16], videos, for example by labeling them with the number of likes or shares to indicate popularity [192], or multi-turn dialogs [147].

The smallest dataset contains about 1,000 labeled data objects [220]. However, most of the datasets are much larger, often exceeding 50,000 labeled data objects. The largest dataset presented in the reviewed publications contains 200 million text posts in multiple languages with sentiment labels inferred from emoticons [55].

The three case studies are very different in size and fall within the range of the other studies. It should be noted that our author profiling dataset, with about 71,000 data points, is one of the largest ever created for this task, and by far the largest in terms of the number of attributes known for multiple authors. The trigger warning dataset, with about 1 million tagged documents, is one of the largest in the review, which is particularly relevant given its complex, multi-stage linking strategy.

Distant Knowledge

Two aspects of distant knowledge play a key role in creating a dataset via weak supervision: the source of the knowledge and the linking strategy that connects the knowledge to the data. Regarding the source of the knowledge, certain general cases can be observed in the relevant publications. On the other hand, the linking strategy in the reviewed publications depends more on the individual case, and systematic differences are limited to the distance between the data object and the knowledge, and thus how involved the linking strategy must be to bridge the distance.

Source of the Knowledge The knowledge sources in the relevant publications can be classified into seven categories: curated list, database, web data, metadata, distant metadata, computed metadata, and classifiers. As noted before, these classes are determined based on the set of relevant publications, and there may be other sources of knowledge that do not fit this taxonomy.

Curated Lists is the most common type used by 15 of the 35 publications. This type includes small, often manually curated lists of knowledge, often created specifically for use with the supervision strategy. For example, Davidov et al. [53] allowed annotators to curate a list of all hashtags and emoticons found in a Twitter post and select those that indicate an emotion, which then became the list of labels automatically assigned to posts containing them. As another example, Mostafazadeh et al. [147] uses a set of phrases (*"I'm 35 as of today."*) that reveal the age or gender of the author and assigns the appropriate label based on the phrase match.

Databases are the second most common type, used by nine publications. This type includes large repositories of data that exist independently of the supervision strategy, and often contain much more data than is used. For example, Blodgett et al. [30] uses U.S. census data to assign demographic data to user cohorts, and Yang et al. [252] uses `botwiki.org` to determine whether a Twitter user is a known bot. Our second case study also uses knowledge from a database, Wikidata.

Web Data is a special case of databases, where an unstructured collection of documents is accessed instead of a structured collection. We have only observed this in one case, Liu et al. [121], where a search engine is used to find text that uses certain tokens.

Metadata is used by six publications. This source refers to all data about the actual data that is directly accessible on the social media platform. This source is often closely related to the data, and the corresponding linking strategy is trivial. For example, Eisenstein et al. [61] determines that the location of the first geo-tagged posts in a Twitter user's timeline is the user's location. There are also more creative uses of metadata. For example, Morstatter et al. [146] created "honeypot" accounts that post stereotypical but largely nonsensical phrases and use the interaction metadata (follows, likes, ...) to identify suspected bots. Our first case study about persuasiveness analysis also uses Metadata to infer the labels: the Δ that the creator of a debate on Reddit hands out to mark a persuasive comment.

Distant Metadata, used by four publications, is a variant of the metadata class where the metadata originates from a different platform than the data object and requires a linking strategy to connect those accounts. For example, Li et al. [116] links Twitter accounts to Facebook and Google+ by identifying links to the other platform in a user's profile and then assigning metadata from the Facebook and Google+ profiles to the Twitter timeline.

Computed Metadata is a rare variant of the metadata class where the metadata is first computed in some way. For example, Tsur and Rappoport [220] first compiles a collection of hashtags from tweets, computes their counts, and uses those counts as labels for popularity.

Classifier is a rare class adjacent to semi-supervised learning, where a classifier is trained on similar but unrelated data and then used to label the data objects. For example, Gilbert and Karahalios [76] trains a classifier on an existing dataset of LiveJournal posts manually labeled with emotion classes, and applies it to label a newly collected set of posts.

Notably, our third case study on trigger warnings is the only one that considers knowledge from multiple sources to infer a label: the tags that authors assign to their works, and the curated list of nodes in the tag graph that link to a particular trigger warning. At most, other work uses multiple sources to create multiple, non-overlapping datasets, such as Morstatter et al. [146].

Linking Strategy The linking strategy, i.e. how the data objects are connected to the distant knowledge, often depend strongly on the specific problem, i.e. the structure of the data and the knowledge used and the distance between them. By distance, we mean how the direct data and knowledge are connected, or how many assumptions the linking strategy has to make. We distinguish three distances: none, close, and far.

	Task	Description
[53]	Sarcasm Detection	Hashtag `#sarcasm` as sarcasm class.
[76]	Emotion Detection	LiveJournal users assign tags with negative emotion words
[52]	Sentiment Analysis	Emoticons or high-polarity hashtags as sentiment class.
[193]	Location Prediction	Geo-location metadata of a post as the place of writing.
[175]	Emotion Detection	Emoticons as emotion class.
[112]	Racism Detection	Users are identified as racists by parsing statements in their biographies.
[131]	Ideology Analysis	Assign a party if the user posted "I voted for X today".
[215]	Sentiment Analysis	Emoticons as sentiment class.
[55]	Sentiment Analysis	Emoticons as (multi-lingual) sentiment class.
[147]	Dialog Generation	Use first reply to a post with an image as dialog response.
[255]	Depression Detect.	Search for depression-related terms in user biography.
[231]	User Profiling	Identify gender or age via revealing phrases.

Table 2.3: Linking strategies of publications where there is *no distance* between the object and the source of the distant knowledge.

Table 2.3 lists the linking strategies of publications with *no distance* between the objects and the distant knowledge. Typical for these publications is that the data object is a text, and part of this text is directly used to "look up" the distant knowledge. For example, Davidov et al. [53] finds tweets that contain the hashtag *#sarcasm* and flags them as sarcastic for classification. The hint text is typically removed so as not to bias the classifiers. Another typical case of "no distance" is when direct metadata is used. For example, when using the geo-tag of a tweet as the location it was sent from [193] or when identifying depression-indicating terms is a user's profile description [255].

Table 2.4 lists the linking strategies of publications with a *close distance* between the objects and the distant knowledge. Typical for these publications is that there is an additional computational step or heuristic involved in linking the data and the knowledge source. For example, Tsur and Rappoport [220] first heuristically determine the popularity of a hashtag via aggregation, and then associate each hashtag with the text posts that use it as data. Another example is user-level location prediction, where a user's location is computed based on multiple geo-tagged posts, either by the first post [61], the centroid over the user's timeline [94], or the centroid over connections on the social graph [199]. Our third case study on persuasiveness analysis

	Task	Description
[61]	Location Prediction	Geo-location of a post is the author's place of residence. (anxiety, worry) as post metadata.
[228]	Sentiment Analysis	Assign sentiment polarity based on polarity of hashtags (inferred via co-occurrence graph).
[199]	Location Prediction	User's location of residence is inferred from the location of close social connections.
[220]	Trend Prediction	Popularity of a hashtag inferred from its frequency.
[250]	Profanity Detection	A post is profane if it's author uses a lot of profanity in general.
[94]	Location Prediction	Twitter: Centroid of post's geo-locations is user residence.
[146]	Bot Detection	(1) An account is a bot if it was deleted after a short while. (2) A honeypot account creates (poor) posts with key terms. Each acquired follower is a bot.
[12]	Ideology Analysis	Infer political ideology based on the news sources shared in the interaction graph.
[85]	Misinformation	A post is trustworthy when send from a trusted outlet.
Ch. 3	Persuasiveness	Persuasive posts are marked by users with a Δ.

Table 2.4: Linking strategies of publications where there is a *close distance* between the object and the source of the distant knowledge.

uses a linking strategy with close distance: the metadata assigned to some comments in a debaters history.

Table 2.5 lists the linking strategies of publications with a *far distance* between the objects and the distant knowledge. Typical of these publications is that the distant knowledge often comes from a different platform or external database without an implicit link to the data, and that multiple computational or heuristic steps are required. For example, Becker et al. [16] labels images from Flickr with the event at which the image was taken. The events are chosen from a list of concerts on Last.fm, and a connection is made when the pictures are tagged with an ID from Last.fm. Another example is the distantly supervised named entity tagging of tweets via Freebase, done by Ritter et al. [191] in the tradition of Mintz et al. [139]. Two of our case studies use a linking strategy with far distance. The second case study on author profiling links entities from Wikidata to Twitter accounts using heuristic rules. The third case study on trigger warnings, uses graph propagation to distribute warning labels from 2,000 manually annotated nodes to about 2 million other tags along various node relations independently created by community volunteers.

	Task	Description
[16]	Event Detection	Users assign event IDs from the `last.fm` events catalogue
[130]	Graph Analysis	Connect Twitter and Friendfeed social graph via account links. as tags to Flickr photos.
[121]	Normalization	Identify clean forms of noisy words in web search snippets using the context as query.
[191]	NER	Extract entities from Freebase and annotate all occurrences in a text corpus as named entity (distant supervision for NER).
[80]	Enrichment	Search posts with an URL to specific news articles and link them.
[116]	Profile Entity Extraction	Find users with connected Google+, Twitter, and Facebook. Annotate education, profession, and familial relations entities via Freebase distant supervision.
[47]	Demographic Analysis	Get a website's user demographic distribution from Quantcast. Link Twitter account. Predict demographics from followers.
[8]	Topic Classification	Identify influencers in a topic via `wefollow.com`. Each post of that user belongs to that topic.
[30]	Ethnicity Analysis	Assign all post of a user a ethnicity distribution from census data based on geo-location. A users' ethnicity is the average over all posts.
[192]	Trend Prediction	Find posts with the url to a video. Number of posts over time indicates popularity.
[227]	Recommend.	Find links between users' Twitter/Facebook and `trip.com` accounts. Use social graph as recommendation signal.
[248]	Misinformation	For posts containing a verbatim news article title: (1) A post is trustworthy when send from a trusted outlet; (2) A post is trustworthy when the news article was cleared by an external fact checking website.
[252]	Bot Detection	User is a bot when it is registered in `botwiki.org`. Verified users are negative (non-bot) examples.
Ch. 4	Author Profiling	Heuristically link Twitter accounts to Wikidata entities.
Ch. 5	Trigger Detection	Map document tags to labels from a taxonomy by propagating them along graph of tag relations.

Table 2.5: Linking strategies of publications where there is a *far distance* between the object and the source of the distant knowledge.

Evaluation and Noise Assessment

The evaluation schemes in the relevant publications can be categorized into five classes: spot checks, weak labels, annotated data, models, and no evaluation. Of the five categories of evaluation schemes used in the relevant publications, only spot checks and weak labels are capable of directly assessing the level of label noise. The remaining evaluation schemes only assess the quality of labeling in relative terms: by showing that a trained model achieves an acceptable level of effectiveness, or by comparing the effectiveness of models using weak labels to models using traditional labels. Of all the publications in the review, only three do a direct assessment of the level of noise in the dataset. This is typically not due to misconduct, but because noise assessment faces similar problems to traditional annotation in that it is limited by cost and the ability of assessors to correctly determine labels, as we discussed in Chapter 1. Table 2.2 shows which scoring scheme is used by the relevant publications considered in the review.

In evaluation via *Spot Checks*, used by three publications, human assessors manually check a sample of weakly labeled data points for correctness. For example, Davidov et al. [53] assign hashtags or emoticons from a curated list as sentiment labels and, present human assessors with a text post and a selection of potentially matching emotion labels, including the weakly labeled one, and ask the assessor to determine the correct label. They measure the accuracy of weak supervision to be 77% for hashtag-based labels and 84% for emoticon-based labels.

In evaluation via *Weak Labels*, used by one publications, a second, unrelated source of distant knowledge is used to evaluate the correctness of the first weak supervision strategy. For example, Badawy et al. [12] infers the political leanings of users on Twitter by aggregating the political leanings of news sources shared by the user, where the political leanings of news sources are derived from distant knowledge, media analysis sites, and lists of known trolls[5] accounts. The second source of distant knowledge is the personal website indicated by the weakly labeled users in their profile: if this website is that of a political party or a (partisan) news source, the user's political leaning is assumed to match that source, and if this assumed leaning differs from the previously computed one, the computed label is counted as noisy. Overall, the authors estimate an accuracy of 90% for their weak supervision strategy. It is apparent that this evaluation strategy depends on the existence of a second, unrelated source of distant knowledge.

In evaluation via *Annotated Data*, used by nine publications, the weakly labeled dataset is compared to a traditionally annotated dataset, either via direct comparison, or by training a model with one and evaluating it on the other. For example, Davidov et al. [52] evaluate a multi-class classi-

[5] Trolls are platform users who deliberately post offensive or provocative messages with the intention of disrupting online discourse [31].

fier trained on their hashtag-based emotion recognition dataset on test data from the same source against human-annotated test data. They find that the classifier performs worse on the weakly labeled dataset (0.54 F_1) compared to the human annotated one (0.83 F_1) because *"the hashgold standard is noisy (containing non-sarcastic tweets) and is biased towards the hardest (inseparable) forms of sarcasm where even humans get it wrong without an explicit indication"* [52].

In evaluation via *Models*, used by 21 publications, the weakly labeled dataset is typically used to both train and test a model. The validity of the weak supervision strategy is assumed if the model reaches acceptable scores when also tested on weakly supervised data but, compared to *Annotated Data*, the model is not evaluated on traditionally annotated data. This evaluation strategy is used when the distant knowledge is close, the linking strategy makes few assumptions, and the labeled data are assumed to be largely correct. Evaluation via *Models* is the most common but also the weakest form of evaluation, but it is applicable in almost every scenario without additional cost. It should be noted that most publications do not explicitly use models to evaluate the dataset, but in most cases, the subject of the study is the modeling of the phenomenon, and an effective model implies that the label noise in the dataset does not invalidate the results, hence, the dataset is effective, too. For example, Roller et al. [193] determine the location where a text post was written based on its geo-tag, and uses multi-class classification and regression models to reason about the results. Similarly, Kwok and Wang [112] evaluate their method for classifying racism, and Yazdavar et al. [255] for classifying signs of depression.

In rare cases, *no evaluation* is done by authors when no other means of evaluation is applicable, the data does not support classification, and the weak supervision is convincingly reliable. For example, Magnani and Rossi [130] links the social graphs of Twitter with the feed aggregator Friendfeed to study the properties of the combined graph, where the link is based on Friendfeed users revealing their Twitter accounts in the corresponding metadata field.

The three case studies also use these evaluation strategies, although we put more emphasis on evaluation than most of the reviewed work. In the first case study on persuasiveness prediction, the dataset is evaluated as most other works via models. The second case study on author profiling is evaluated, in addition to modeling experiments, via weak labels: By checking the Twitter profiles added as metadata in some Wikidata entities, the recall (0.723) and precision (0.994) of the linking strategy can be estimated. The third case study on trigger warnings utilizes, first, spot checks to estimate the F_1 score (0.96) on a sample of 1,000 labeled tags via manual annotation, and second, weak labels to estimate the recall (0.86) across all tags by using freeform tags that contain 'trigger warning' verbatim.

2.2.3 Conclusion

This section presents a systematic review of 35 relevant, high quality publications that construct a dataset of user-generated content using weak supervision. The review identifies the following parameters of weak supervision within the domain: The most common tasks are classification tasks such as misinformation or location prediction, but also language processing tasks such as text normalization and information extraction, and computational social science tasks such as trend prediction or bot detection. Most publications study Twitter data due to its popularity and ease of access, and the most common data objects are text posts and users. The review identifies seven types of distant knowledge: curated lists, databases, web data, metadata, distant metadata, computed metadata, and classifiers. The review also outlines that supervision strategies, or how the data and knowledge are linked, are very task, data, and knowledge specific, but that they can be loosely grouped according to the distance between the data and knowledge, from immediate to cross-platform. Finally, the review identifies five evaluation strategies, spot checks, weak labels, annotated data, and models.

An effective and efficient evaluation of weak supervision strategies is also the biggest open research problem. Most publications only evaluate by training a model and evaluating the model effectiveness because other ways of evaluation are not possible or too expensive. However, model effectiveness is a poor way to evaluate the validity of a dataset if, conversely, the dataset is intended to evaluate the model effectiveness. The label noise in the dataset can cause the model to learn to detect the noise, where the model would appear to be effective but is not, and thus the dataset would appear to be valid but is not. Consequently, direct evaluation of a weakly labeled dataset is a prerequisite for determining the validity of that dataset.

The three case studies presented in this thesis build on the theoretical foundations established in this review and push the boundaries in several parameters. The three datasets we created with weak supervision enable new tasks, such as trigger detection, or extend existing tasks, such as author profiling, by providing larger datasets with a broader set of labels. The new datasets also apply weak supervision to new domains and platforms, such as fan fiction documents on Archive of Our Own, and present new linking strategies that even combine multiple distant knowledge sources, which was not done by any of the reviewed works. In learning from the review of evaluation strategies, we use weak labels, spot checks, or both.

Limitations The major limitation of the study design is the low recall of publications. By restricting to the major venues, some important publications may be missed. An example of this is the pioneering works by Go et al. [78], which was independently published but was highly influential with over 4,000 citations. Similarly, the study misses all works with less than 100 citations, although many of these works create novel and interesting datasets using weak supervision. The citation-based filtering also restricts the platforms and data types used by the reviewed publications, as studies on mainstream social media platforms attract more academic attention and thus affords more citations. Finally, the review misses all works that do not use the terminology expected by the review. We assume that works studying weak supervision also use the appropriate terms, however, many works create a dataset using heuristics and distant knowledge without being aware of the theoretical underpinnings.

3

Case 1: Analyzing the Persuasiveness of Debaters on Reddit

The first case study is an investigation into the persuasiveness of debaters on Reddit's debate forum ChangeMyView (`reddit.com/r/changemyview`). The study uses a dataset constructed via weak supervision to label if a comment is persuasive or not. The research question addressed in this chapter is:

RQ 1. Why are some debaters more persuasive than others?

For this case study, we use weak supervision to create a dataset of debaters, all of their contributions to various debates in chronological order, and the persuasiveness of the contributions. The distant knowledge in this case are the delta (Δ), a metadata flag assigned by the debate opponent to the most persuasive comment in a debate. By aggregating the delta comments, we can assess each debater's persuasiveness and how it changes over time. Weak supervision is essential to enable this study for the two reasons discussed in Chapter 1: labeling persuasiveness is subjective and requires domain knowledge, at least in a debate setting, and the analysis is very data-intensive, requiring many comments annotated for many debaters each.

Persuasiveness describes why some arguments are more convincing to an audience than others, and why some debaters perform well in a debate while others do not, even if they use the same arguments. Studies on persuasion in online discussions focus on *comments* and ignore the role of the *debaters*. By analyzing the debaters' persuasion strategies over multiple discussions, we seek to uncover the behavioral characteristics (e.g., engagement), language style (e.g., used frames), and argumentative techniques that distinguishes debaters with different levels of effectiveness (e.g., good vs. poor). To do so, we categorize ChangeMyView debaters based on their effectiveness in persuasion and examine key differences in their behaviors and skills (i.e., engagement and experience), as well as their argument's style at the semantic, syntactical, lexical, and pragmatic levels. Finally, we propose the task of identifying effective debaters and present a machine learning-based solution using existing argumentative features along with newly utilized ones such as syntactic complexity, semantic similarity, and argument framing; the latter is shown to play a role in the debater's persuasiveness.

Section 3.1 introduces the core concepts of ChangeMyView as required for our study, and the primary studies on modeling persuasion on ChangeMyView and related platforms. Section 3.2 describes the Reddit dataset and the weak supervision strategy used to construct it. Then, in Section 3.3, we analyze the dataset in terms of how style and debate strategy vary between debaters, change with experience, and if this indicates persuasiveness.

All resources developed for this case study, the dataset and analytical code, can be found at `https://doi.org/10.5281/zenodo.7034173`.

3.1 Persuasiveness in Online Debate Forums

Persuasion, a primary goal of argumentation, is the ability to convince people to take a certain action or form a certain belief [154]. Persuasion has always influenced the dynamics of communication and social interaction, either positively, by raising awareness of critical issues such as climate change, or negatively, by spreading propaganda and fake news.

Due to its growing role in shaping beliefs, social media has attracted considerable interest as a means to gain a deeper understanding of persuasion [229]. In particular, the ChangeMyView subreddit has been used in various studies that model text persuasiveness using a variety of linguistic, argumentative, and behavioral features (e.g., [86], [126], and [81]).

However, most scholarly work on online persuasion has focused on studying *persuasive comments* in individual discussions, without considering the importance of analyzing *persuasive debaters*. [128]. As a result, debaters' strategies and their effectiveness have not been adequately studied. Understanding effective debating strategies and debaters' persuasiveness can be highly beneficial for media analysis, rhetorical review, and learning debating skills. Moreover, it can advance the development of various applications, where effective strategies can be recommended in writing assistants and dialog management systems, or encoded in the backbone of text generation tools.

ChangeMyView

ChangeMyView is an open platform for users to engage in civilized discussions using sound arguments. ChangeMyView discussions are actively moderated to maintain the quality of argumentation. All comments and original posts must abide by the community rules.[1] These rules provide a predictable structure for ChangeMyView discussions, making them easy to manage.

A CMV discussion begins with a user, called the *original poster*, posting a marked request, called *original post*, to the ChangeMyView subreddit. The subreddit prohibits non-debative posts. The original post states the original poster's stance on a controversial topic, relevant justifications and explanations of that stance, and an (implicit) request to "change my view". All other users of ChangeMyView, called *debaters*, can challenge the original poster's stance and post opposing argumentative top-level comments. All debaters can respond to other comments to counter, cross-question, or defend their arguments, creating multi-layered and complex threads of conversation.

ChangeMyView offers two mechanisms to indicate comment persuasiveness: The delta (Δ) and the comment score. The delta mechanism allows the original poster to mark up to one comment as persuasive. The "awarded" deltas are aggregated, and the number of Δ per debater is publicly displayed. Reddit's comment score is the sum of upvotes and downvotes per comment. The highest scoring comments are displayed first. The comment score on ChangeMyView serves as an alt-metric indicating persuasiveness as perceived by the community.

[1] ChangeMyView rules are posted on their wiki: `https://www.reddit.com/r/changemyview/wiki/rules`

Related Work

The major work on the analysis of argument persuasiveness on social media (cf. [214], [257], [166], and [86].) tries to determine how persuasive a comment is by solving the task: given two comments with a shared original poster, identify the persuasive one. In contrast, this thesis provides a higher-level analysis. We try to determine how persuasive a debater is by studying the debaters across multiple discussions, striving to disclose their persuasion strategies.

Employing argumentative features to predict comment persuasiveness is a well-established strategy; Egawa et al. [58] annotated ChangeMyView discussions with elementary argumentative units (EUs) in a token-level five-class scheme: testimony, fact, value, policy, and rhetorical statement. The authors propose a Bi-LSTM-based sequence classifier for EU labeling. They conclude that EUs indicate persuasiveness if used effectively, 'fact' is the most persuasive, that the proportional distribution of types distinguishes ChangeMyView comments from original posts, and that persuasiveness is not indicated by the mere presence or absence of certain EUs.

Similarly, Hidey et al. [87] annotated ChangeMyView discussions regarding arguments' claims as interpretation, evaluation, agreement, disagreement, or premises as ethos, logos, and pathos. The authors show that the relative positional distribution of argumentative components in a ChangeMyView comment is a signal for its persuasiveness. Additionally, Li et al. [115] demonstrated the effectiveness of arguments' structural features in persuasiveness prediction. Multiple features were developed based on the usage of the proposition types reference, testimony, fact, value, and policy in the debaters' texts. The feature analysis showed that the presence of 'value' and 'testimony' bi-grams is more prevalent in persuasive argumentative texts, indicating that justifying claims with personal experiences is an effective persuasion strategy.

Several papers have examined various characteristics and behaviors of debaters. Addressing debaters' behavior, Tan et al. [214] examined the role of debaters' interaction dynamics with the original poster in persuasion and found that debaters who respond early in the discussion tend to be more successful, that engaging with the original poster improves a debater's chances of success up to a threshold, and that higher debater participation in a discussion improves the chances of persuasion.

Focusing on the characteristics of debaters, Al-Khatib et al. [3] modeled debaters' beliefs, personality traits, and interests based on their past activity on Reddit and used them to tackle the task of predicting persuasiveness. The study found that the similarity between the characteristics of the original poster and the debaters is influential for effective persuasion. In comparison, this thesis groups debaters based on their persuasiveness so that we can explore the different strategies used by good vs. poor debaters.

Analyzing the discussion structure, Guo et al. [81] hypothesized that persuading the original poster in a ChangeMyView discussion happens gradually throughout a multi-turn conversation rather than immediately. A prediction task was performed to model the cumulative effect of a sequence of comments in a ChangeMyView discussion and detect the position where the persuasion of the original posters occurs. Besides, a user study to evaluate the persuasiveness of debaters' arguments' was conducted, concluding that the perception of persuasiveness differs across individuals and that it is influenced by one's idiosyncrasies i.e. the same argument could be persuasive for one person but not persuasive for another. Likewise, Wei et al. [234] considered the relevance ranking of ChangeMyView comments by their score in a discussion. They found the comment's score to be influenced by its temporal entry order as well as the past credibility of its corresponding debater. The credibility is measured by the number of prior deltas received by a debater. Several feature classes were used for the relevance ranking task, including linguistic features derived from the comment's text, interaction-based features obtained by modeling the ChangeMyView discussion as a tree, and argumentative features such as the proportion of argumentative text and argument relevance and originality.

The only work targeting the debater-level is by Luu et al. [128]; They investigate how debaters' skill improves over time as they learn how to interact with other debaters. They present a strong estimator of the development of a debater's persuasive skill over time using several linguistic features, such as length of comments, co-occurrence of hedges, and fighting words.

Our work is distinct in several respects: First, we analyze ChangeMyView, as opposed to `Debate.org`, which is more strict and conventional regarding debate structure. Second, we analyze the relationship between the debaters' engagement, experience, and writing style across linguistic dimensions, accounting for the argumentative nature of debate texts. Finally, we address different levels of debater persuasiveness and scrutinize the differences in their argumentation strategies.

3.2 A Dataset for Debater Analysis

To conduct our study of debater persuasion strategies, we created a dataset of $3,801$ ChangeMyView debaters, equally sampled for good, average, and poor debater persuasiveness. Here, we detail our data collection method, quantification of debater persuasiveness, and sampling method to balance the dataset by debater persuasiveness.

Quantifying Debater Persuasiveness

We define the persuasiveness of a debater d with a given sequence of comments $c_1, \ldots, c_{i,\Delta 1}, \ldots, c_{j,\Delta k}, \ldots, c_n$ in ChangeMyView as ratio of delta comments c_Δ to all comments:

$$\text{Persuasiveness}(d) = \frac{k}{n}.$$

The persuasiveness is, hence, the number of debater's delta comments normalized by her total comment count. As Table 3.1 shows, this normalization is necessary because the delta-comment count correlates strongly with the total comment count.

Based on the persuasiveness score, we categorize debaters into three groups as follows:

1. **Good debaters** with a persuasiveness of 5% or above.
2. **Average debaters** with a persuasiveness between 0% and 5%.
3. **Poor debaters** with a persuasiveness of 0%; These debaters did not receive any delta during their active period on ChangeMyView.

The separation of debaters with a non-zero persuasiveness is based on the observation that obtaining any c_Δ is already challenging. Hence, the highly persuasive tail should be studied as a separate population. The 5%-threshold used in categorization separates the non-poor debaters into two groups of approximately equal size.

Collecting Debater Comments

We obtained an initial set of ChangeMyView debates from the Webis ChangeMyView corpus [3], which comprises all ChangeMyView debates

	Persuasiveness	**Δ Count**	**Score**
Comments	0.02	0.72	0.03
Active Period	-0.03	0.13	0.15

Table 3.1: Pearson ρ between three success measures, persuasiveness, Δ comment count, and median reddit comment score and two absolute experience measures, active period duration and the number of comments.

from June 2005 to September 2017. We extracted all top-level comments from this corpus and grouped them by debater. We discarded all inactive debaters with less than 10 comments and obtained an unbalanced dataset of $13,254$ ChangeMyView debaters along with their top-level comments on various debates. We only considered top-level comments as they serve as debate starters, while lower-level comments are either rebuttals or non-argumentative content like corrections, clarifications, or thanks.

Sampling Data

In the intermediate dataset, 80% of the debaters are of poor persuasiveness and have never been awarded a Δ. Since we aim for a controlled analysis, we resampled the dataset in such a way that the distribution of ChangeMyView debaters is balanced by persuasiveness. Overall, we end up with $3,801$ entries, evenly distributed across the three debater categories.

Our resampling strategy first added a "good" debater to the dataset by random and then selected one "average" and one "poor" debater with the same number of comments, or the closest number to that. If multiple candidate debaters existed, we minimized the absolute difference in mean comment length. Since both comment count and length are indicative of persuasiveness, the resampling minimizes this bias in the dataset.

3.3 An Analysis of Persuasion Strategies

The analysis of persuasion strategies has three parts: The first part concerns the role of engagement, i.e. whether active and frequent debaters are more persuasive. The second part concerns the role of style in persuasion, i.e. which lexical, syntactic, semantic, and pragmatic features identify good and experienced debaters. The third part concerns the prediction of persuasiveness, i.e., whether the previous findings and the studied phenomena can be used to predict a debater's persuasiveness.

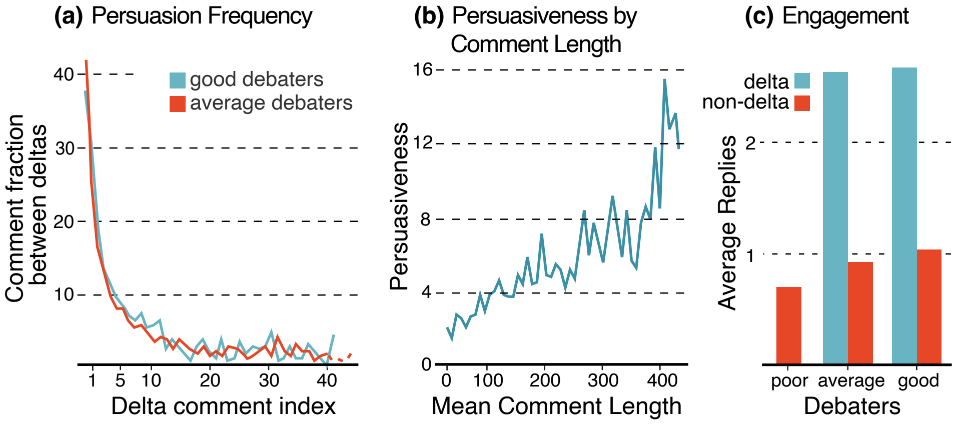


Figure 3.1: (a) Evolution of the frequency of persuasive comments. (b) Persuasiveness by debaters' average comment length. (a) Engagement of debaters by persuasiveness.

3.3.1 Debater Engagement and Experience Analysis

The first analysis concerns the relationship between the debaters' persuasiveness and their engagement with and experience on the ChangeMyView subreddit. We presume that engagement on ChangeMyView may correspond to rebuttals in live debates. Our findings suggest that a high engagement is indicative of persuasive debaters. We further inspect the relationship between experience and persuasiveness in both absolute measures such as comment count and active period and relative measures such as changes in style and persuasiveness with experience gain. Our findings suggest that debaters become more persuasive with increased experience, especially average debaters. However, the experience effect is not reflected in absolute experience measures, and hence it is hard to operationalize for classification.

3.3.2 Engagement

Figure 3.1(c) shows that persuasive comments and persuasive debaters are more engaging. We measure debater engagement by the average number of replies to persuasive and non-persuasive comments. Persuasive debaters get about 10% more replies to their total comments compared to average debaters and about 30% more replies compared to poor debaters. Persuasive comments get about 250% as many replies as non-persuasive comments.

Absolute Experience

Table 3.1 shows that the absolute measures of experience are insufficient. We can observe that neither the active period—the time between the first and latest comment—nor the comment count correlates with persuasiveness or Reddit score. We disregard the correlation between the total comment count and the number of persuasive comments as evidence of debater experience without observing a correlation with persuasiveness.

Relative Experience

We model the relative experience of a debater on ChangeMyView as seen from the comment: A debater is inexperienced for her first comment and very experienced for her last; that is to say, the experience of the debater d of a comment c_t in a sequence $c_1, \ldots, c_n$ is $\text{Experience}(c_t) = \frac{t}{n}$. We analyze the impact of experience gain of good and average debaters on persuasiveness, persuasion frequency, comment length, as length is the most indicative feature in comment classification, and average comment score, which represents the ChangeMyView community's opinion on persuasiveness.

Persuasiveness Figure 3.2(b) shows that the overall persuasiveness of good debaters is largely unaffected by experience while the persuasiveness of average debaters almost doubles.

Persuasion Frequency Figure 3.1(a) shows that the persuasion frequency increases sharply up to the 5th persuasive comment for both good and average debaters and increases slightly up to the 15th persuasive comment. This indicates that debaters learn to replicate persuasive strategies and become more persuasive with experience. We measure persuasion frequency as the number of non-delta comments that occur between two consecutive delta comments, as a fraction of the total comments made. A decreasing delta-to-non-delta rate indicates more frequent persuasions.

Comment Length Figure 3.2(a) confirms the established assumption that length is highly indicative of persuasiveness. There is no indication that relative experience has any substantial impact on the length of delta or non-delta comments.

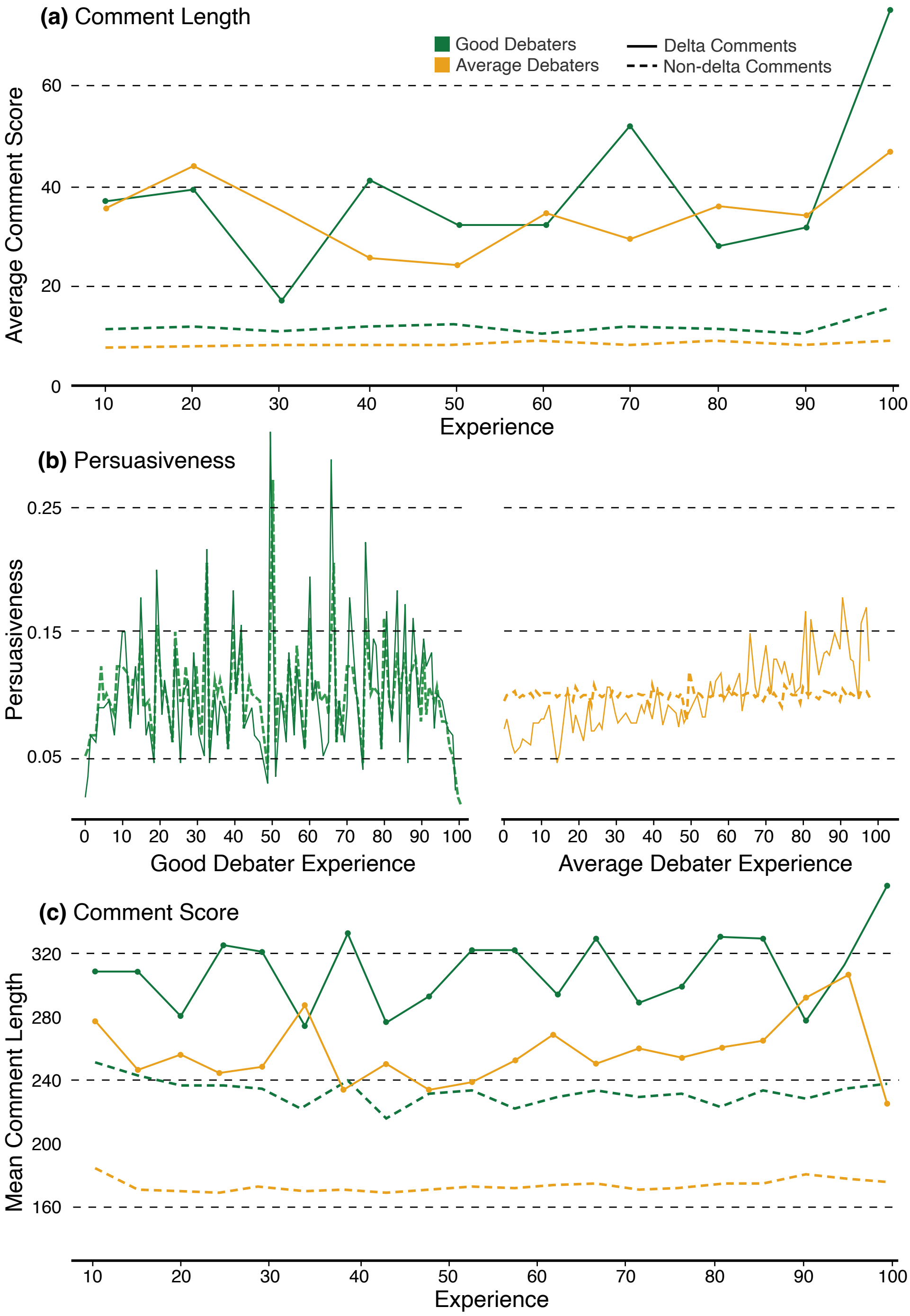


Figure 3.2: Changes of various debater-level features with increasing relative experience. The color indicates debater persuasiveness.

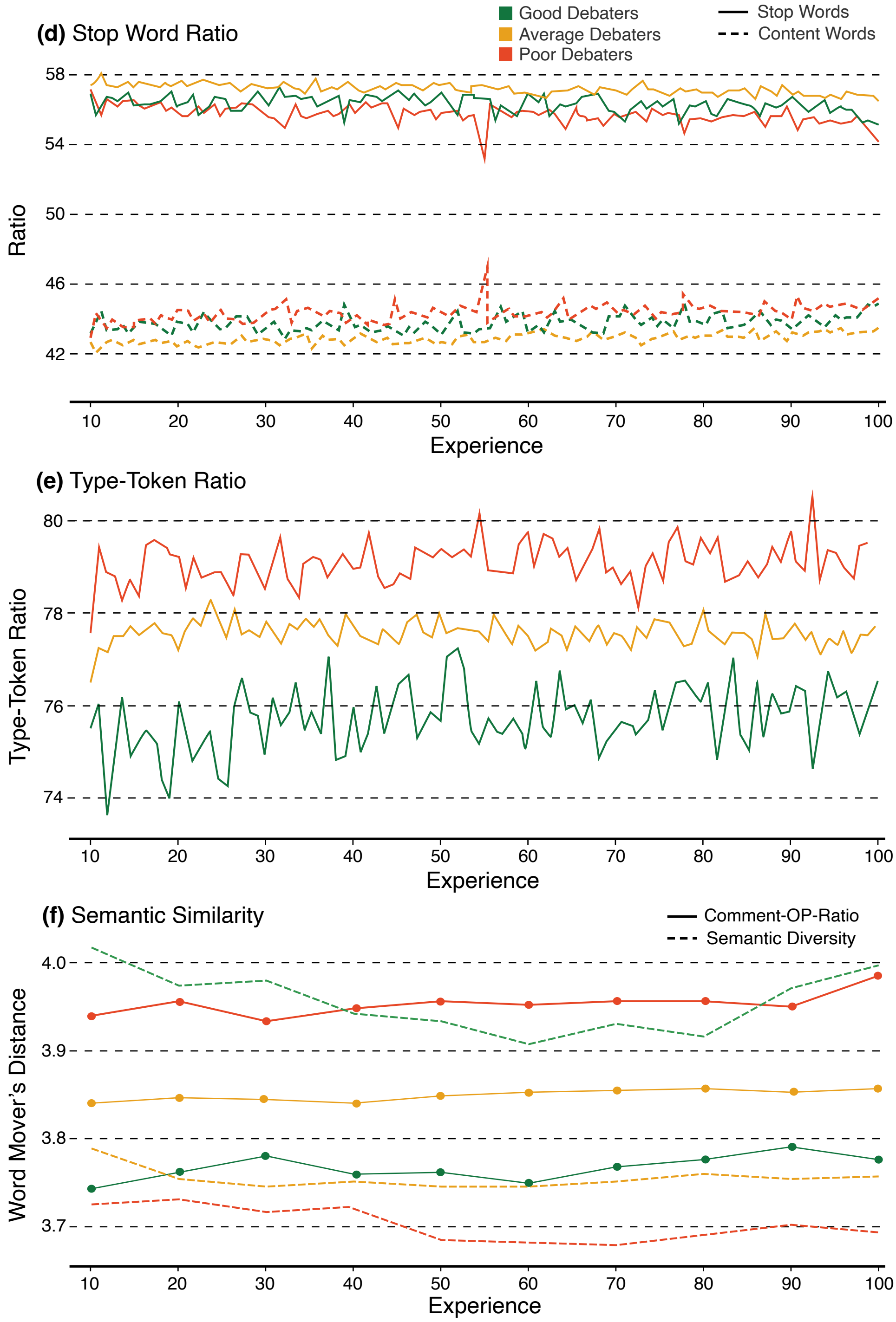


Figure 3.3: Changes of various debater-level features with increasing relative experience (continued). The color indicates debater persuasiveness.

Average Comment Score Figure 3.2(c) shows that the mean-comment score, the alt-metric for community persuasiveness, increases with experience but not consistently. On average, however, debaters score higher on persuasive comments with increasing experience. The effect. however, is negligible on non-delta comments.

3.3.3 Debater Style Analysis

Stylistic features are frequently used to determine the characteristics of authors. Since stylistic features are indicative of persuasive comments, we consider stylistic features to also be indicative of persuasive debaters. In particular, we study the relationship between a debater's persuasiveness and the lexical, syntactic, semantic, and pragmatic dimensions. We found notable differences in persuasiveness in each dimension. The most substantial feature is again comment length. Additionally, we found that better debaters tend to have lower lexical diversity and syntactic complexity, but a higher semantic diversity. We also found correlations between certain word class patterns and certain patterns of elementary argumentative units, particularly rhetorical statements. Lastly, found that persuasive debaters use political and cultural identity frames more often.

Lexical Dimensions

Within the lexical dimension of style, we analyze the relation between debater persuasiveness and the (1) comment length and the (2) lexical diversity, in particular the stop-word and type-token ratio.

Comment Length Figure 3.1(a) shows that debaters with a higher mean comment length are also, consistently and without apparent bound, more persuasive on average. Figure 3.2(a) shows, independently of the debater's experience, that persuasive comments are longer than non-persuasive comments and that good debaters write longer (~20%) comments. These findings are consistent with previous evidence (cf. Section 3.1) and suggest that the comment length is highly indicative of the persuasiveness of comments and debaters alike.

Lexical Diversity Figure 3.3(d) shows that the differences in the stop-word ratio are consistently small (<1%) and have no direction since good

WC n-gram	ρ	WC n-gram	ρ
`IN JJ`	0.11	`PRP VBP`	-0.13
`NN IN JJ`	0.10	`PRP`	-0.12
`JJ NN IN`	0.09	`WRB VBP`	-0.11
`VBG DT JJ`	0.08	`NN WRB`	-0.11

Table 3.2: Top Pearson ρ between a word class n-gram and persuasiveness.

debaters are between poor and average ones. However, Figure 3.3(e) shows that the type-token ratio has a higher effect size of 2% among the debater groups and has a direction. This suggests that good debaters write comments with lower lexical diversity.

Syntactic Dimensions

Within the syntactic dimension of style, we analyze the relationship between persuasiveness and syntactic complexity and the word class n-gram distribution.

Syntactic Complexity The complexity of a debater's text was measured based on the dependency parse trees of all sentences in her top-level comments. We measure the Pearson correlation between debater persuasiveness and three common syntactic complexity measures:[2] Outdegree centrality ($\rho = -0.17$), Closeness centrality ($\rho = -0.16$), and the number of dependents per word ($\rho = 0.17$). Since a high centrality indicates complex syntax, and persuasiveness is negatively correlated with centrality, our results suggest that good debaters use less complex syntax. However, all correlations are weak ($\rho <= 0.25$).

Word class n-grams Table 3.2 shows the word class 1–3-grams with the strongest correlation with persuasiveness. Here, better debaters use adjectives more and `PRP VBP` (e.g. you did ...) as well as `WRB VBP` (e.g. how did ...) less frequently. Although the correlation is weak and word-class n-grams are difficult to interpret, these results may indicate an impact of certain syntactical structures on debater persuasiveness as for comment persuasiveness (cf. Tan et al. [214]). We determined the word class n-grams

[2] We measured the complexity using `https://github.com/tsproisl/textcomplexity`

using NLTK and the Penn tagset since all ChangeMyView comments are English. We only inspected the 1,000 most frequent n-grams.

Semantic Dimension

Within the semantic dimension of style, we measure the relation between debater persuasiveness and the (1) semantic similarity between a debater's comment and the original post and the (2) semantic diversity within the comments of a debater. We use Word Movers Distance[3] (WMD, Kusner et al., 2015) to measure the semantic similarity.

Similarity between Comment and Original Post Figure 3.3(f) shows that the WMD is lower the more persuasive a debater is. Hence persuasive debater's comments are semantically more similar to the original post.

Semantic Diversity Figure 3.3(f) shows the semantic diversity for debaters with different persuasiveness, whereas the semantic diversity is higher for better debaters.

Semantic diversity indicates if a debater prefers semantic depth (few different concepts discussed) or breadth (many different concepts discussed) within each comment. For lack of a better, lexeme-agnostic intra-document semantic similarity measure, we use a sentence-based heuristic:

$$\mathrm{SemDiv}(c_k) = \frac{2}{n^2 - n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathrm{WMD}(s_i, s_j).$$

Here, the semantic diversity of a debater is the average diversity of the comments $c_k = s_1, \ldots, s_n$, and the diversity of the comments is the average WMD between each pair of sentences (s_i, s_j). We assume WMD captures the semantic diversity between two sentences in this way.

Pragmatic Dimension

Within the pragmatic dimension of style, we measure the relation between debater persuasiveness and (1) the distribution of argumentative units: elementary units, claims, and premises, (2) framing strategies.

[3] We use Gensim with fastText embeddings

Unit n-gram	ρ	**Unit n-gram**	ρ
rhetoric	-0.194	policy	-0.110
value	-0.126	rhetoric rhetoric	-0.101
rhetoric value	-0.114	rhetoric rhetoric none	-0.063

Table 3.3: Argumentative units with largest absolute Pearson ρ with ChangeMyView debaters' persuasiveness. All other combinations correlated with $\rho \leq 0.05$

Argumentative Units Table 3.3 shows the argumentative unit n-grams which correlate the strongest with debater persuasiveness, while all other unit n-grams do not correlate with $\rho \leq 0.05$. All correlating units are elementary units, with rhetorical statements being the most persuasive. No claim or premise types correlate in a meaningful way with persuasiveness.

We measure the Pearson correlation between persuasiveness and the relative frequency of elementary unit 1–3-grams, where each sentence of a debater's comment is assigned one unit. We use the five elementary units testimony, fact, value, policy, and rhetorical statement proposed by Egawa et al. [58] for ChangeMyView comments. We determine the elementary unit of a sentence with a BERT-based classifier trained on Egawa et al. [58]'s annotated dataset of ChangeMyView comments and original posts; The classifier reaches a 6-class (including *None*) micro-accuracy of 0.75 on the standard split. Since the dataset annotates units on a token level, we assign each sentence the unit assigned to its tokens, discarding sentences with multiple units annotated.

We also measure the Pearson correlation between persuasiveness and the relative frequency of 1–3-grams of claim and premise types, where each sentence of a debater's comment is assigned one type. We use the 2-stage classification scheme proposed by Hidey et al. [87] for ChangeMyView comments. Each sentence is first classified with a BERT model as claim, premise, or neither. Claims are then classified as interpretation, evaluation/rational, evaluation/emotional, or agreements. Premises are classified into eight classes, one for each combination of ethos, logos, and pathos using three binary classifiers. We trained each of the five needed classifiers on Hidey et al. [87]'s datasets of ChangeMyView discussions.

Frames Figure 3.4 shows how often debaters with different persuasiveness use certain frames in their comments. Most frames are used equally often independently of persuasiveness, except for the *political* and *cultural identity* frames, which are used notably more often by better debaters.

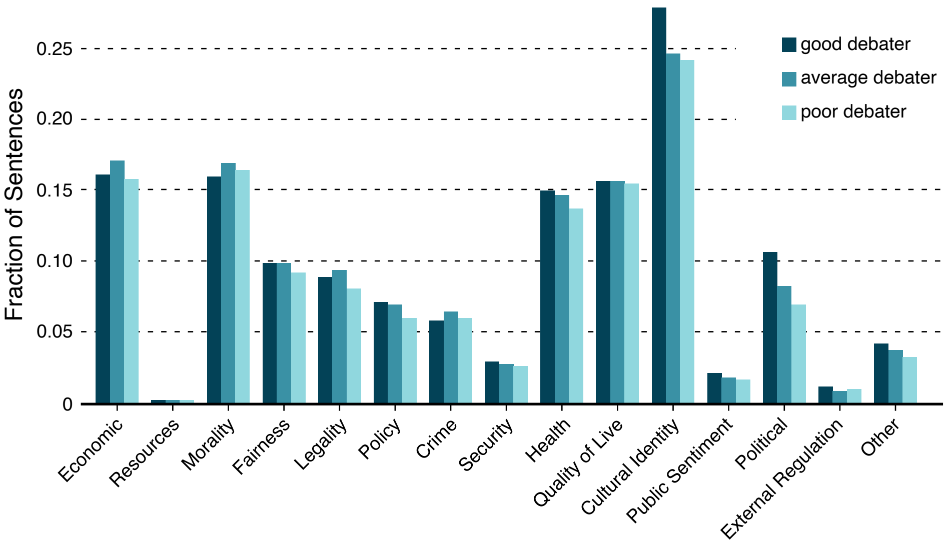


Figure 3.4: Distribution over the 15 sentence-level frames for good, average, and poor debaters.

We determined frames by classifying each sentence of each comment of a debater with one of the 15 frames used in Card et al. [36]'s Media Frames corpus of manually annotated news articles. We trained a BERT classifier to classify the sentences, which reaches a micro accuracy of 0.68 in 5-fold random cross-validation.

3.3.4 Predicting Persuasiveness

In addition to the analytical scrutiny of debater persuasiveness, we conduct an experimental validation of our findings by classifying debaters by persuasiveness. We define the general task of debater-level persuasiveness prediction as: Given a debater d with comments $c_1, \ldots, c_n$, classify this debater as persuasive (good) or non-persuasive (average or poor). To conclusively supplement our analysis, we individually inspect the classification performance of the introduced features (cf. Section 3.3.3).

We encoded the syntactic, semantic, and pragmatic features of our analysis for each of the 3,801 debaters in our ChangeMyView debaters' corpus. Each encoding was chosen to obfuscate comment length as far as reasonable. We encoded the word class and all argumentative unit n-grams tf-idf vectors of the aggregated comments. We encoded the numerical features, outdegree centrality, closeness centrality, and the number of dependents for text complexity and comment-op distance and within-comment distance for

Features	**Good vs**	
	Average	Poor
Baseline Features		
Bag of Words	0.60	0.68
Stylometry	0.62	0.67
Vocabulary Interplay	0.58	0.67
Syntactic Features		
Word class n-grams	0.57	0.51
Text Complexity	0.65	0.61
Semantic Features		
Word Mover's Distance	0.59	0.63

Features	**Good vs**	
	Average	Poor
Pragmatic Features		
Elementary Units	0.51	0.59
Claim or Premise	0.47	0.55
Claim Type	0.48	0.58
Premise Type	0.48	0.58
Claim and Premise Types	0.48	0.58
Frames	**0.70**	**0.72**

Table 3.4: Macro F1 score of the two classification settings: Good vs. Average debaters and Good vs. Poor debaters.

WMD, by averaging comment-level counts per debater. We encoded each of the 15 frames with the absolute and relative number of comments that utilize a frame.

As baselines, we selected feature sets previously used for *comment persuasiveness* prediction: Bag-of-Words, vocabulary interplay after [214], which covers original poster and commenters' vocabularies' absolute and relative overlap and Jaccard similarity, and common stylometrics, which cover counts of words, selected word classes, links, word lists, symbols, type-token ratio, and readability scores. The baseline feature sets were implemented trivially following the related work.

We consider two binary classification settings for our experimental validation: (1) good vs. average and (2) good vs. poor. We maintained a balanced distribution of the classes (1,267 each) and used a logistic regression classifier with default parameters on a random 80-20 train-test split. The effectiveness of the classifiers is reported as macro F_1 (see Table 3.4).

The classification results reveal several findings: First, most features distinguish good from poor debaters better than good from average ones. Syntactic features are the only exception to this trend, which can not be explained by our analysis. Second, Bag-of-words is a strong feature for the two settings as it outperforms most of the other features. Besides, the weak effectiveness of the argumentative features is similar to the observations of Egawa et al. [58]; the mere distribution of argumentative units in the text is insufficient to identify its persuasiveness. Third, the distribution of the frames in the debaters' comments results in the best scores across the two experimental settings. The most discriminating frames are 'Quality of Life', 'Morality', and 'Health and Safety', all with negative weights towards the 'good debater' class.

3.3.5 Conclusion

The persuasion skills of debaters vary, depending on different cultural and social aspects, among others. Understanding how people argue and what makes some debaters more successful than others are interesting research questions that have been neglected in the literature. This chapter has contributed in this regard by modeling debater effectiveness in ChangeMyView and analyzing their behavior and argumentative stylistic choices, demonstrating several interesting insights that can be utilized for improving the persuasion skills of new debaters and assessing the development of advanced text generation and writing assistant tools.

In particular, our analysis of debater strategies shows that(1) the persuasiveness improves over time for average debaters, (2) the distribution of 'frames' in the debaters' arguments can play a major role in persuasion, and (3) argumentative features based on the presence of certain types of arguments in the debaters' text do not seem sufficient to indicate the effectiveness of persuasion.

Limitations Although this study substantially contributed to understanding the role of the debater in the persuasiveness of arguments, we think that there is room for further analysis. First, we quantified the persuasiveness of ChangeMyView debaters based on awarded Δs only. Although it appears to be a standard method in previous work, we believe that a more comprehensive quantification, possibly using human judgments and a more fine-grained scale, would account for a degree of subjectivity to consider the evaluating user's idiosyncrasies. Guo et al. [81] touches on this briefly, finding that despite general agreement about what is persuasive, there are differences in the assessment of persuasion based on the positions of the evaluating party.

Second, while argumentative features based on the distribution of argumentative units did not perform well in our prediction task, possible improvements can be achieved through modeling features that can capture the effective use of argumentative units. A possible direction is to identify the interdependencies between the different argumentative units in the text [115] as well as their relative arrangement [87].

Third, other, more in-depth features can disclose useful insights into debater persuasiveness. Conceivable are features that better model behavior such as experience and the dynamic of debater interaction or the velocity of experience gain.

Fourth, using more sophisticated models in the prediction task may lead to better results, although our logistic regression is ideal to compare class separability by feature. Guo et al. [81] proposed using conditional random fields (CRF) to model the cumulative effect of persuasion in ChangeMyView discussions, and Li et al. [115] used bi-LSTM and BERT to model their persuasiveness task.

4

Case 2: Profiling Influencers on Twitter

The second case study is an investigation of the application of author profiling technology to influencers on Twitter. The study uses a dataset constructed using weak supervision to label an influencer with a large number of personal attributes. The research questions addressed in this chapter are:

RQ 2. Can author profiling technology be effectively transferred between populations?

RQ 3. Are the posts of a group of fans indicative of the demographic attributes of an influencer?

For this case study, we use weak supervision to create a dataset of Twitter accounts of 71,706 influencers, all the tweets on their public timeline consisting of an average of 29,968 words, and up to 239 pieces of personal information for each influencer. The distant knowledge in this case is properties extracted from the Wikidata entries of the influencers. As a linking strategy, we use several heuristics to transform Twitter usernames into candidate names for Wikidata pages, collect the matching pairs, and then use another set of heuristics to filter out false positive links. The links are evaluated using a 'weak label' strategy: some Wikidata pages list the Twitter account of the influencer, which we use to measure the high precision of 0.994 and the reasonable recall of 0.723.

The first research question concerns the rare label problem in author profiling, where we propose to transfer models from a specific group of people to the genral population as an angle of attack. Author profiling aims at inferring personal attributes from an author's texts, such as demographics or personality types. Training a good classifier requires sufficient data for each attribute of interest, and collecting such data is only possible for ubiquitous but often uninteresting attributes, such as age and gender, or for rare attributes, but only for a small group of people.

Here, we propose to use a proxy, influencers, to build a classification model and apply it to the general population. Using influencers has several advantages: they are prolific social media users who provide many writing samples, many personal details are public knowledge, and they attempt to create a consistent public persona either themselves or with the help of agents. Our proposed solution is to develop and evaluate several state-of-the-art profiling models submitted to a shared task on datasets sampled from the corpus and to demonstrate the effectiveness of transferring models between our dataset and much smaller general population datasets.

The second research question concerns the related problem that many users on social media platforms are passive and provide few writing samples for a profiler. We propose to use the theory of social media homophily, which suggests that close connections also share many attributes, to profile authors through the texts of their friends and followers. Our proposed solution is to extend our dataset to a systematic sample of 2,380 influencers, including the timelines of ten of their fans, and to again create evaluative profiling models for this dataset in a second shared task.

Section 4.1 elaborates on how influencers fit into profiling research, introduces the state of the art and open problems in author profiling, and in Subsection 4.1.1 presents our systematic review of different methods for creating author profiling datasets. Section 4.2 then describes the influencer profiling dataset and the weak monitoring strategy used to construct it. Building on this dataset, Section 4.3 describes the development of an effective profiling technique, and Section 4.4 describes our experiments on transferring models between datasets. Finally, Section 4.5 describes the extension of our dataset with fan timelines and the results of the corresponding classification experiments.

All resources developed for this case study, the dataset and the analysis code, can be found at `https://github.com/webis-de/ACL-19`.

4.1 Influencers as Authors in Profiling Research

Author profiling is about predicting personal traits of individual authors based on their writing style. Frequently studied traits are demographics such as gender, age, native language or dialect, and even personality, with applications in marketing, forensic linguistics, psycholinguistics, and the social sciences.

Given the high expectations that are implied by these and similar applications, the creation of a valid automatic profiler for a given trait, let alone many, depends on the availability of carefully constructed corpora. Corpus construction for author profiling has always been difficult for lack of large-scale distant supervision sources that provide for genuine pieces of writing from many different authors alongside personal information. In part, the aforementioned selection of demographics that are frequently studied reflects the availability of corresponding ground truth. In this regard, one source of ground truth, available in large quantities, high diversity of traits, and near-perfect label reliability, has been overlooked: influencers.

On social media, influencers occupy an exalted position. Rallying up to millions of followers, they serve as role models to many and exert a direct influence on public opinion, sometimes for the better, e.g., by lending their voices to the disenfranchised, and sometimes for the worse. Unsurprisingly, the "rich and famous" are subjects to research in the social sciences and economics alike, especially with regard to their presence on social media.

Related Work

The study of author profiling techniques has a rich history, with the pioneering works done by Pennebaker et al. [163], Koppel et al. [108], Schler et al. [202], and Argamon et al. [9], focusing on age, gender, and personality from genres with longer, grammatical documents such as blogs and essays. Table 4.1 overviews most of the works done in author profiling over the past 20 years, reporting on text genre, author count, word count, and the demographics studied. The most commonly used genre in recent years is Twitter tweets, first used in 2011 to predict gender [32] and age [161]. Later work also used Facebook posts [67], Reddit [77], and Sina Weibo [230]. Recently added demographics include education [37], ethnicity [225], family status [230], income [174], occupation [172], location of origin [67], religion [179], and location of residence [37].

At the PAN workshop (`pan.webis.de`), author profiling has been studied since 2013, covering different demographics including age and gender [183, 184, 186], personality [180], language variety [185], genres including blogs, reviews, and social media posts [186], cross-domain prediction [187]and profiling author characteristics outside the domain of demographics, such as the authors inclination to spread fake news [181] of detecting if an author writes like a bot [182]. Profiling research related to aspects such as behavioral traits [110], medical conditions [44], and native language identification (NLI) have been excluded from our survey, since these have developed into subfields of their own right.

Methodologically, author profiling has been comparatively stable over the last decade: most approaches utilize supervised machine learning based on the authors' texts and varying stylometric and psycholinguistic features to encode non-lexical information. The additional features proved to be important to the degree that even advanced neural network architectures are only competitive if these features are explicitly encoded [75]. The biggest methodological improvements, experimentally shown for selected demographics, are the usage of message-level attention, recently proposed by Lynn et al. [129] and of network homophily by encoding information from the social graph. The pioneering work by Kosinski et al. [109] shows that the common likes of Facebook users suffice to predict demographics like gender, sexual orientation, ethnicity, and substance use behavior with up to 0.9 accuracy. Recent advances in graph encoding algorithms [79] motivated the use of node embeddings as supplemental features when predicting age and gender on Facebook [66], occupation and income [4], racism and sexism [140], and suicide ideation [141] on Twitter. Similar approaches have also been explored in related fields to, for example, profile the bias and factuality of news agencies [13]. An even more advanced approach to predict the occupation of authors was suggested by Pan et al. [155], who jointly encoded the adjacency matrix of the follower graph with the biographies of all authors in the network using graph convolutional neural networks.

4.1.1 Labeling Strategies in Author Profiling

We analyzed 29 publications on author profiling the authors of which explicitly describe their data acquisition and corpus construction strategies. The strategies have been reviewed, abstracted, and mapped into a taxonomy, which in turn enabled us to identify specific quality criteria. Table 4.1 overviews these publications and reports key figures, personal traits, and the underlying acquisition strategy.

Src	Genre	Lang.	Authors	Words	Personal Traits	Label Acquisition Strategy
[138]	Blogs	1	100	20,323	Gender	AIS
[150]	Blogs	1	1,997	27,303	Age	AIS+U
[196]	Blogs	1	24,500	(?)	Age	AIS
[202]	Blogs	1	37,478	7,885	Gender	AIS
[184]	Blogs	2	346,100	632	Age, Gender	AIS
[230]	Sina Weibo	1	742,323	(?)	Age, Education, Gender, Relationship	AIS
[32]	Tweets	12+	183,729	283*	Gender	AIU
[37]	Tweets	1	5,000	17,195*	Education, Residence	AIU
[77]	Comments	1	23,503	24,861	Personality (MBTI)	AIU
[168]	Tweets	1	1,500	12,880	Gender, Personality (MBTI)	AIU
[172]	Tweets	1	5,191	26,415*	Occupation (SOC)	AIU
[179]	Facebook	1	1,019	2,178	Age, Education, Gender, Personality (Big Five), Religion	AIU
[185]	Twitter	4	19,000	1,195	Dialect, Gender	AIU
[223]	Twitter	6	18,168	25,400	Gender, Personality (MBTI)	AIU
[173]	Tweets	1	13,651	23,717*	Politics	AIU
[64]	Emails	1	1,033	3,259	Age, Gender, Education, Native lang., Personality (Big Five), Residence	ARS
[65]	Emails	1	1,033	2,085	Age, Education, Gender, Personality (MBTI)	ARS
[67]	Facebook	4	479	2,156	Age, Birthplace, Gender, Education, Extroversion, Nat. lang., Occupation	ARS
[120]	Essays	1	500	145	Age, Education, Gender, Personality	ARS
[174]	Tweets	1	4,098	16,785*	Age, Education, Gender, Income, Race	ARS
[180]	Tweets	4	1,070	1,205	Age, Gender, Personality (Big Five)	ARS
[218]	Tweets	1	250	31,011*	Personality (Big Five)	ARS
[222]	Essays	1	749	976	Age, Birthplace, Gender, Personality (Big Five)	ARS
[173]	Tweets	1	3,938	15,587*	Age, Gender, Politics	ARS
[204]	Facebook	1	136,000	4,129	Age, Gender, Personality (NEO-PI-R)	ARS
[45]	Tweets	4	8,618	12,700*	Gender	ORS
[63]	Tweets	1	6,610	31,750*	Gender	ORS
[225]	Tweets	1	5,000	2,540	Age, Children, Education, Gender, Income, Intelligence, Optimism, Political alignment, Ethnicity, Religion, Relationship, Satisfaction	ORS
[95]	Essays	1	186	286	Age, Gender	OIS
[18]	Papers	1	4,500	(?)	Gender, Native language	OIS
We	Tweets	37	71,706	29,968	up to 239	OIS

Table 4.1: Survey of author profiling corpora. A star indicates an estimated word count, assuming 12.7 words per tweet; a question mark indicates unavailable information. Rows are grouped by acquisition strategy.

	Independent		**Requested**
	Structured	**Unstructured**	**Structured**
Author	(AIS) Profile forms	(AIU) Posts, Comments	(ARS) Questionaires
Others	(OIS) Wikidata	(OIU) News, Mentions	(ORS) Crowdsourcing

Table 4.2: Taxonomy of label acquisition strategies with typical instances.

Three criteria describe the quality of the surveyed resources: the representativeness of the targeted group of people, the comprehensiveness in terms of author, text, and label size, and the reliability of label attributions. Table 4.2 shows our taxonomy of label acquisition strategies for reliability and comprehensiveness evaluation: labels provided by the author or by others (A/O), labels provided independently or on request (I/R), and labels retrieved in structured or unstructured form (S/U). We disregard all combinations of requests for unstructured data (R-U) as inapplicable.

There is generally no strategy that clearly increases the resource quality:

1. Requested labels by experts, volunteer annotators, or crowdsourcing workers are prone to subjectivity or misunderstandings; Self-reported labels by authors are prone to deception and self-serving bias.

2. Requested labels are prone to self-selection bias and have a high per-author cost; Independently reported labels are prone to few and stale choices of attributes to profile.

3. Unstructured data is prone to imprecision, incompleteness, and misunderstandings; Structured labels are prone to restricted choices.

Note that profiling research related to aspects such as behavioral traits [110], medical conditions [44], or native language identification (NLI) have been excluded from our survey, since these have developed into subfields of their own right.

4.2 An Influencer Dataset for Author Profiling

This section introduces the Webis Celebrity Corpus 2019, detailing how we identified influencers at scale, compiled a large corpus of their writing, and linked it with Wikidata to obtain personal profiles. A corpus analysis and validation follows.

I	alphanum. characters of the display name	**IV**	first and last part of **I**, split at spaces
II	reference name split at capitalization	**V**	all but the last part of **I**
III	reference name split at display name	**VI**	all but the last two parts of **I**

Table 4.3: Rules to generate name candidates for Wikidata matching from Twitter reference and display names.

4.2.1 Who is a Influencer?

To operationalize the term "influencer", we say that a person has an influencer-like status, with reach in a large or narrow community, if he or she possesses a *verified* Twitter account, and at the same time, is deemed *notable* enough to be the subject of a Wikipedia article and a Wikidata item.

Importantly, Twitter used to verify "that an account of public interest is authentic" [221], awarding a blue checkmark badge: ✓.[1] Notability at Wikipedia pertains to people who are "worthy of notice," "remarkable," or "famous or popular" [246].

The intersection of these two sources is necessary and provides a good and scalable approximation of what makes an influencer. Verified accounts alone are insufficient because author profiling is only concerned with attributes of humans and Twitter also verified brands, institutions, organizations, and other non-person entities. On the other side, Wikipedia and Wikidata also considers historical persons to be notable, which are not influencers. So, to collect influencer profiles at scale, we join these sources of information.

4.2.2 Corpus Construction

We crawled all 297,878 verified Twitter accounts and linked them with Wikidata items. The verified accounts could easily be identified through the `@verified` Twitter account,[2] who followed all other verified accounts, hence we collected a list of all followed accounts via Twitter's API. We then then filtered out all influencers who declared a non-English profile language,[3] or were born before 1940, as well as all tweets that did not contain mainly text.

[1] It should be noted that the verification system described here was replaced in 2023. The current system, as of time of writing, rewards blue checkmarks to paid subscribers, gray checkmarks to government accounts, and yellow checkmarks to verified organizations.

[2] `https://web.archive.org/web/20180403094636/https://twitter.com/verified`

[3] Note that the dataset still contains some bilingual and non-English tweeting influencers.

The method for linking Twitter accounts to Wikidata entities works in two stages: candidate generation, which finds potential account-entity pairs, and filtering, which removes ineligible candidates. In the first stage, the account-entity pairs are found by generating six normalized variants of the unique, static Twitter *"@"-handles* and free-form *display names*, and then querying Wikidata for entities with the variant name as a name property. The six variants are generated using the following heuristics:

I Remove all non-alphanumeric characters from the *display name*. This removes most embellishments.

II Split the *handle* at each capitalized character. The handles show the intended capitalization and many users use `@FirstLast`-style handles.

III Split off the *display name* from the *handle*. Many users use their first or last name as their display name, but a concatenation as their handle.

IV Split the alphanumeric variant **I** at whitespace and use only the first and last parts. This removes middle names or nickname injections, which are often not part of the Wikidata name property.

V Split **I** on whitespace and use all but the last part.

VI Split **I** on whitespace and use all but the last two parts. **V** and **VI** remove (self-assigned) titles or other prefixes used as embellishments.

Linking accounts by name is a non-trivial task, since a Twitter account name and its corresponding Wikidata entity need not be an exact string match, and there may be false matches. Common reasons for mismatches include omission of middle or last names, use of nicknames, frequent name changes to reflect current events, or embellishments such as adding emoticons.

In the second stage stage, all pairs where the Wikidata page does not belong to a human influencer are removed, the results of which are shown in Table 4.5a. Only one matching pair is kept for each Twitter account and matches with a higher rank (descending from **I** to **VI**) are preferred. We filtered out matches non human and memorial accounts, as well as ambiguous and erroneous matches:

1. Each Wikidata entity has an `instance of` property, which can either be *human* or any other, in which case we consider it *not human*.
2. When a linked entity contained one of the eight death-related Wikidata properties and a date of death from before Twitter was launched in March 2006, we consider the account to be *memorial*.

Label	Occurrences		Most frequent value	
Sex	65,035	90.1%	Male	71.7%
Occupation	63,017	87.9%	Actor	15.3%
Date of birth	60,493	84.4%	-	-
Educated at	28,134	39.2%	Harvard	2.1%
Sport	18,688	26.1%	Football	30.8%
Languages spoken	12,094	16.9%	English	54.9%
Political party	6,703	9.4%	Republican	16.4%
Genre	6,699	9.3%	Pop Music	21.6%
Race	3,531	0.5%	African Am.	66.5%
Religion	2,960	0.4%	Islam	23.5%

Table 4.4: Selection of relevant personal traits studied in the related work, how often they have been assigned in our corpus and the most frequent value.

3. When different entities match for the same Twitter account, and these matches differ by language, we consider the account to be *ambiguous*.
4. All mismatches identified during our subsequent corpus validation were marked as *error* (see Section 4.2.4).

After excluding matches with private timelines, 71,706 valid account-entity matches remained.

4.2.3 Corpus Analysis

The corpus contains an average of 29,968 words per author and 1,523 different Wikidata properties, of which 239 were manually identified by us as personal traits relevant for profiling. Table 4.4 shows a selection of these properties, the most common value, and how many influencers they are annotated for. The remaining properties are 1,224 external references (i.e., links to other sites) and 60 miscellaneous properties (mostly internal references and multimedia data). Of the 239 properties, 45 are attributed to more than 1,000 users, and 5 are attributed to more than 55,000 users simultaneously. The extracted Wikidata traits are highly specific, often having over 100 different values per trait in our corpus, although most are Zipf distributed and can be easily aggregated or reduced to smaller dimensions (see Section 4.3). Note that labels such as ethnicity, religion, and native language are mostly present for members of minority groups, so the distribution of these traits in the corpus is highly skewed.

We collected a total of 156,411,899 tweets ($\approx$ 3 billion words), with an average of 2,181 tweets per influencer. The corpus contains a maximum

(a)

	Celebrity	**Error**	**Memorial**	**Not hum.**	**Ambig.**
all	71,706	124	2,666	60,232	896
I	91.8%	50.0%	70.4%	77.6%	82.6%
II	2.8%	3.2%	2.6%	6.2%	1.8%
III	0.1%	0.0%	0.0%	0.1%	0.0%
IV	1.8%	23.3%	5.6%	3.8%	5.3%
V	2.9%	21.8%	9.2%	10.6%	9.6%
VI	0.3%	1.6%	12.3%	1.9%	0.8%

(b)

Dataset	**Authors**	
	Training	Test
PAN15 [180]	152	142
PAN16 [186]	428	78
PAN17 [185]	3,600	2,400
PAN18 [187]	2,000	1,900
Celebrities	31,861	13,614

Table 4.5: (**a**) Evaluation of matching success as per generation rule. (**b**) Sizes of the datasets used for evaluation.

of 3,200 tweets per influencer, since this is the limit imposed by Twitter's Timeline API endpoints. However, since the total number of posts is visible for each user, we can estimate that the corpus contains 98.05% of all tweets sent by the included influencers.

Of all collected tweets, 29.3% are retweets and 20.9% are replies. Of the remaining 49.7% of tweets, an average of 989 (13,938 words) per influencer are longer than 20 characters and do not contain links, providing a conservative estimate of tweets suitable for style analysis. Although influencers tweeted in 50 different languages, 77% of all timelines consisted of tweets written exclusively in English, followed by 7% in Spanish and 4% in French, while 2,104 influencers tweeted at least bilingually.

4.2.4 Evaluation of the Linking Strategy

Wikidata implicitly provides a large ground truth for evaluating our Twitter-Wikidata matches: 89,451 human entities contain a Twitter username; 28,454 of these usernames intersect with the 297,878 verified Twitter accounts we crawled. Comparing these 28,454 true matches with those obtained by our matching heuristic, we distinguish three cases: (1) 20,579 are correctly linked, (2) 124 are incorrectly linked (0.6% error rate), and (3) 7,751 are not linked (27.7% miss rate). Thus, our heuristic achieves a very high precision of 0.994 at a reasonably high recall of 0.723.

Table 4.5a (bottom row group) breaks down the number of matches by type and name candidate. The most successful name candidate is **I**, yielding 92% of all matches, but only half the erroneous ones. Name candidates **II**, **III**, and **VI** contribute negligibly, while candidates **IV** and **V** provide only for 5% of the matches combined, but 45% of all errors. At an overall error rate of 0.6%, though, candidates **IV** and **V** produced 3,416 correct and only 56 incorrect matches, rendering them still viable.

Representation We may cautiously claim to have obtained a wide cross-section of people with a large reach on Twitter. However, influencers are excluded who do not use Twitter, whose account is not verified (which is exceedingly unlikely, the more famous they are), or who have no Wikipedia article about themselves. There are no reliable estimates of the true number of influencers worldwide, but it is safe to assume that our corpus has a bias towards Western culture, and particularly English-speaking influencers.

Comprehensiveness The corpus provides for comparably long samples of writing per author and a rich set of traits, albeit many traits are available only for a subset of profiles. Most influencers provide genuine writing samples of themselves at Twitter, but some employ public relations staff to manage their account. Though a problem for generic author profiling, this does not impede *influencer profiling*. influencers craft public personas as their own unique brands. If an influencer decides to employ staff to do so, approving their impersonations, these personas are no less genuine and normative than personally crafted personas.

The information about the traits of influencers obtained from Wikidata can be considered highly reliable. Dedicated volunteers collect all kinds of personal information about influencers, which are often referenced and under constant review by other Wikipedia and Wikidata editors. As per our taxonomy of label acquisition strategies in Table 4.2, we employ an OIS strategy: we obtain labels from third-party expert annotators (O), who are independent (I), supplying data in structured form (S).

4.2.5 Conclusion

This section introduces the "Webis Celebrity Corpus 2019", the first corpus of its kind comprising a total of 71,706 influencer profiles, 239 profiling-relevant labels, and 3 billion words. Its quality is due to Twitter's verification process, Wikidata's accuracy, and our low-error linking strategy between the two sites. Its generalizability qualities for gender prediction has been demonstrated using state-of-the-art approaches.

Potential future work includes improvements to the corpus by including verified accounts from other social media, and by inferring new labels for previously unlabeled influencers through link prediction.

4.3 Influencer Profiling

The usefulness of the dataset for building effective profiling technology can best be demonstrated by training state-of-the-art classification models. For this purpose, we organized a competitive shared task, the "Celebrity Profiling Task 2019", as part of the PAN evaluation lab.[4] and invited other research labs to develop classification models on the influencer dataset and submit their best systems for evaluation.

The task's goal was to evaluate technology for predicting four demographic attributes of an influencer from their history of tweets:

- *Gender* was to be predicted in three classes as male, female, or, for the first time, diverse.
- *Year of Birth* was to be predicted precisely as opposed to the more conventional way of predicting fixed age groups. However, this demographic was evaluated more leniently within a novel, variable-bucket evaluation scheme.
- *Renown* was to be predicted in three classes as low, medium, or high.
- *Occupation* or "Claim to Fame" was to be predicted in eight classes as athlete, entertainer, creative, politician, manager, scientist, professional, or clergy.

The evaluation data for this task was sampled from the "Webis Celebrity Profiling Corpus 2019" (see Section 4.2). The gender, year of birth, and occupation labels were obtained from Wikidata, and renown was derived from the number of followers. Participants were given a large training dataset of 33,836 influencers with up to 3,200 tweets each, and submissions were evaluated on a test dataset of 14,499 influencers using the TIRA evaluation service [72]. Performance was evaluated using a combination of the multi-class F_1 scores of each demographic (Section 4.3.1).

[4] https://pan.webis.de/clef19/pan19-web/celebrity-profiling.html

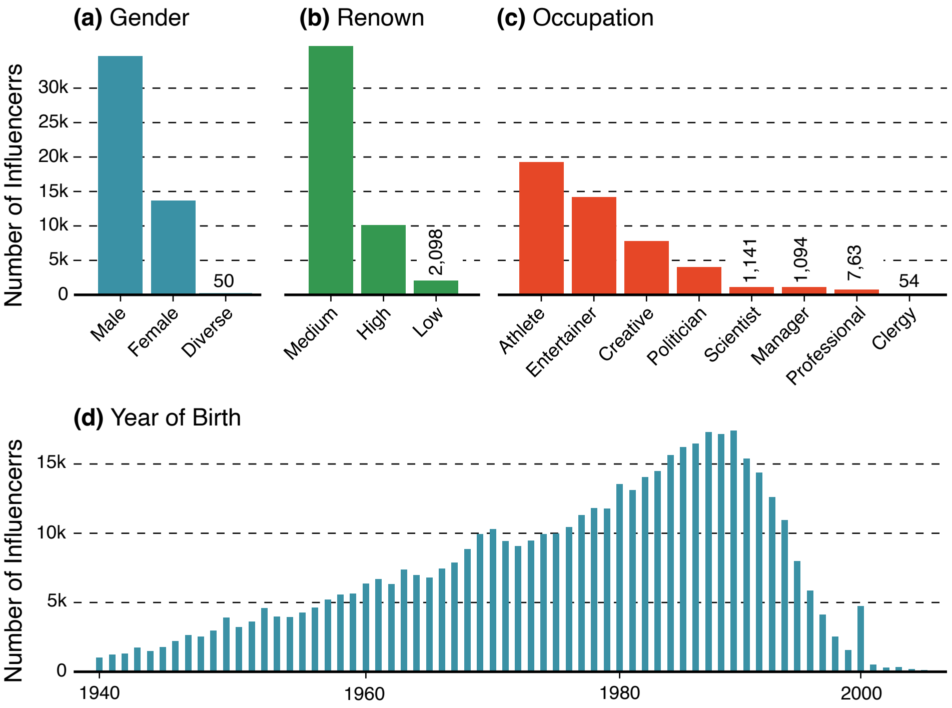


Figure 4.1: Distribution of demographics over the authors in the PAN 2019 dataset.

A total of 92 teams registered for the task, 12 were actively working on submitting a model, and eight made a successful software submission (Section 4.3.2). Performance was measured using *cRank*, the harmonic mean of the macro-averaged multi-class F_1 for gender, renown, occupation, and a leniently calculated F_1 for year of birth. This measure is stricter than average accuracy because it favors consistent results and emphasizes performance on classes reflecting rare demographics. The winning submission achieved an outstanding *cRank* of 0.593. Most submitted systems use word-level features over neural approaches, reporting higher performance of the former in preliminary experiments (Section 4.3.3).

4.3.1 Data and Evaluation

The data used for this task was sampled from the "Webis Celebrity Profiling Corpus 2019" (see Section 4.2), which links the Twitter accounts of influencers with their corresponding Wikidata entries.

To compile the evaluation data for our task, we sampled all influencers for the most widely available demographics, namely gender, occupation, and year of birth. Figure 4.1 shows histograms for each demographic in the sample of the corpus used for this task.[5] Altogether, the evaluation data comprises 48,335 influencers with an average 2,181 tweets.

Wikidata properties have an extreme number of values for certain demographics. To render the prediction tasks feasible, we simplified the labels as follows:

- *Gender:* From the eight different gender-related Wikidata labels, we kept *male* and *female* and merged the remaining six to *diverse*.
- *Renown:* To determine the degree of renown, we calculated the PMF of follower counts, fitted a log-normal distribution, and used its standard deviation to separate the three classes: *low* with less than 1,000 followers, *medium* with between 1,000 and 100,000 followers, and *high* with more than 100,000 followers.
- *Occupation:* The 1,379 different occupations were manually mapped to eight classes by, first, using Wikidata's `subclass of` properties to construct an undirected graph that connects all occupations in the corpus, and then manually identifying the strongly connected sub-structures:
  1. *Athlete* for occupations participating in professional sports.
  2. *Entertainer* for creative activities primarily involving a entertainment artists like acting, TV hosts, and musicians.
  3. *Creative* for creative activities with a focus on creating a work or piece of art, for example, writers, journalists, designers, composers, producers, and architects.
  4. *Politician* for politicians and advocates, lobbyists, and activists.
  5. *Manager* for executives in companies and organizations.
  6. *Scientist* for people working in science and education.
  7. *Professional* for specialist professions like cooks and plumbers.
  8. *Clergy* for professions in the service of a religious group.
- *Year of Birth.* Unlike the profiling literature on age prediction, we did not define a static set of age groups, but used the year of birth between *1940* and *2012* as extracted from Wikidata's `Day of Birth` property.

[5] The high number of influencers for year of birth 2000 is an error in Wikidata that we noticed only at the time of writing. We removed them in our subsequent analyses.

The different demographics in the dataset are not entirely independent. While the correlation of some class combinations like year of birth and renown, and gender and renown are insignificant, others have notable dependencies, since there is 1:2 imbalance between gender and occupation, and occupation and year of birth. Female influencers tend to be younger and more likely to have a performing or creative occupation, while male influencers are more likely to be famous for athletics at a young age, and otherwise for politics and religion. influencers in performing occupations like acting or music tend to be more famous than others.

We split the sampled data 70:30 into a training dataset of 33,836 influencers and a first test dataset of 14,499 influencers, from which we subsampled the second, smaller test dataset of 956 authors.

Evaluation Measures

In previous author profiling tasks at PAN, the performance of an approach was measured by the average accuracies measured for each demographic. However, this measure is unfit for influencer profiling, since the demographics are imbalanced across many different classes. Instead, we macro-average the individual measures of effectiveness for each demographic using the harmonic mean, which rewards systems with a consistent performance across all demographics:

$$\text{cRank} = 4 \div \left(\frac{1}{\text{F}_{1,\text{renown}}} + \frac{1}{\text{F}_{1,\text{occupation}}} + \frac{1}{\text{F}_{1,\text{gender}}} + \frac{1}{\text{F}_{1,\text{birthyear}}}\right).$$

Let T denote the set of classes of a given demographic (e.g., gender), where $t \in T$ is a given class label (e.g., female). The prediction performance for $T \in \{\text{gender, renown, occupation}\}$ is measured using the macro-averaged multi-class F_1-score. This measure averages the harmonic mean of precision and recall over all classes of a demographic, weighting each class equally, to reward the correct prediction of rarer classes:

$$\text{F}_{1,T} = \frac{2}{|T|} \cdot \sum_{t_i \in T} \frac{\text{precision}(t_i) \cdot \text{recall}(t_i)}{\text{precision}(t_i) + \text{recall}(t_i)}.$$

We also use this measure to evaluate prediction performance for the demographic $T =$ year of birth, but change the calculation of true positives: we count a predicted year as correct if it is within a m-window of the true year, where m increases linearly from 2 to 9 years with the true age of the influencer in question:

$$m = (-0.1 \cdot \mathrm{truth} + 202.8).$$

This way of measuring prediction performance for age demographics addresses a shortcoming of the fixed age interval scheme: Defining strict age intervals (i.e., 10-20 years, 20-30, etc.) overly penalizes small prediction errors made at the interval boundaries, such as predicting an age of 21 instead of 20. We also decided not to combine precise predictions with an error function such as mean squared error, because we expect age prediction to become more difficult with increasing age, as people mature and their writing style presumably changes more slowly over the years.

4.3.2 Submitted Systems

Eight participants submitted software for this task, six of which also submitted notebooks describing their system. Five of these six approaches are based on traditional feature engineering, and three also report negative experiments with deep learning models, while only Pelzer [162] used a neural language model (ULMFiT). The most popular algorithm choices are logistic regression and support vector machines (SVM), the most popular features are exclusively based on content, while only Moreno-Sandoval et al. [144] also added grammatical and user-defined features.

To deal with the small classes in the gender and occupation demographics, two participants resorted to oversampling the classes during training, one to downsampling, and one applied class weighting. Three participants grouped the year of birth into eight maximally sized intervals and predicted them instead. The most popular preprocessing steps are replacing or removing hyperlinks, mentions, hashtags, and emojis, while stop words and punctuation are rarely touched. Each approach is described below.

Radivchev et al. [178] uses support vector machines to predict renown and occupation and logistic regression to predict year of birth and gender, using tf·idf vectors of the 10,000 most frequent bigrams of 500 randomly selected tweets per influencer as features. The authors determined class priors to cope with small classes in gender and occupation prediction and grouped the year of births into eight intervals, reversing the window function used for performance measurement. Tweets are preprocessed by removing retweets and all symbols except letters, numbers, @'s, and #'s, replacing hyperlinks with <url> and mentions with <user>, collapsing spaces, and adding a <sep> token at the end of each tweet. The optimal configuration of learning algorithms for each demographic was determined via grid search over several hyperparameter settings for both the SVM and logistic regression. The authors tried multiple alternative approaches, reporting sub-par results for preserving retweets and replacing emojis with <emoji> during preprocessing, using character 3-grams and 4-grams as features, and employing multi-layered perceptrons or a deep pyramid CNN on GloVe embeddings.

Moreno-Sandoval et al. [144] uses logistic regression to predict renown, gender, and year of birth, and a multinomial naive Bayes model to predict occupation, using n-gram features with a minimum frequency of 9 for gender, 6 for year of birth, 3 for occupation, and none for renown, as well as the features average number of emojis, hashtags, mentions, hyperlinks, retweets, words per tweet, word-length, the lexical diversity, the kurtosis and skew of word-length and word-count, respectively, and the number of tweets written in each of the grammatical genders: the first, second, and third person singular and the first and third person plural. Years of birth are combined into eight larger intervals and oversampled. Preprocessing of texts was done for renown, gender, and year of birth in the form of replacing hashtags, mentions, hyperlinks, and emojis with special tokens. The model configurations described above were obtained by testing several combinations of (1) the five algorithms naive Bayes, Gaussian naive Bayes, naive Bayes complement, logistic regression, and random forest, and (2) whether to apply preprocessing, (3) oversampling, and (4) whether to include the features.

Martinc et al. [134] uses logistic regression for all four demographics, with tf·idf vectors of word uni-grams, word-bounded character 3-grams, and 4-character suffix 3-grams of the first 100 tweets per timeline as features. The suffix 3-grams were based on the 10%-80% most frequent words and were weighted with 0.8, the character 3-grams 4-80% with 0.4-weighting, and the word uni-grams 10-80% with 0.8-weighting. No resampling was applied and all years were predicted without regrouping. The text for both trigram features was preprocessed by replacing hashtags, mentions, and hyperlinks with special tokens and the text for the word uni-grams by additionally removing all punctuation and stop words. The authors determined the logistic regression algorithm to be optimal after performing a grid search over different hyperparameter combinations of linear SVMs, SVMs with RBF kernel, logistic regression, random forest, and gradient boosting classifiers. Experiments with BERT-based fine-tuning approaches were reported as non-competitive.

Asif et al. [10] utilizes one model for each combination of the four demographics and the 50 languages the authors detected in the dataset, using the most discriminative words as features. To determine the best learning algorithm for each combination, the authors selected the best-performing one after testing support vector machines, logistic regression, decision trees, Gaussian naive Bayes, random forests, and k-nearest neighbor classifiers. The most discriminative word features for each demographic were determined by aggregating word counts for all users of one class, normalizing these counts by the frequency of the class, and summing the pairwise intraclass distance in relative frequencies. This calculation results in a ranking of words for each demographic, indicating which words are more frequently used by members of one class compared to members of all other classes, where the occurrences of the highest-ranking words were used as features. All tweets are preprocessed by removing hyperlinks, punctuation, stop words, numbers, alphanumeric words, escape characters, #'s, and @'s.

Petrik and Chuda [167] use multiple random forest classifiers with 200 decision trees based on the tf·idf vectors of the top 5,000 1-, 2-, and 3-grams. To train the models, the authors used the synthetic minority oversampling technique in combination with Tomek links to balance the examples for each class. The timeline text is preprocessed by removing mentions and stop words, collapsing letter repetitions, and replacing hyperlinks and emojis with special tokens. Additionally, the authors report on experiments with RCNNs, which did not deliver promising results and were hence discarded.

Pelzer [162] applies a transfer learning strategy by training an ULMFiT instance on the influencer timelines. The classifiers constructed from this instance predicted a class for every tweet in a given timeline and used the majority of all per-tweet predictions to infer the influencer's demographic. The authors further refined their model by regrouping the year of birth into fewer classes and downsampled the examples of all demographics to get a more balanced training dataset. The author reports on slow prediction times of 8 minutes per influencer; this approach was only evaluated on the second, small-scale test dataset.

Baselines Since this is the first edition of the task, the only baselines provided were the common cases of random predictions: (1) Uniform ran-

(**a**) Primary metric cRank and minor F_1 scores for both test datasets.

System	Test dataset 1					Test dataset 2				
	cRank	gender	age	renown	occup	cRank	gender	age	renown	occup
Radivchev	**0.593**	**0.726**	**0.618**	**0.551**	**0.515**	**0.559**	**0.609**	**0.657**	**0.548**	0.461
Moreno	*0.505*	*0.644*	*0.518*	0.388	*0.469*	0.497	0.561	0.516	0.518	0.418
Martinc	0.462	0.580	0.361	*0.517*	0.449	0.465	*0.594*	0.347	0.507	**0.486**
Fernquist	0.424	0.447	0.339	0.493	0.449	0.413	0.465	0.467	0.482	0.300
Petrik	0.377	0.595	0.255	0.480	0.340	0.441	0.555	0.360	*0.526*	0.385
Pelzer	–	–	–	–	–	*0.499*	0.547	*0.518*	0.460	*0.481*
Asif	–	–	–	–	–	0.402	0.588	0.254	0.504	0.427
Bryan	–	–	–	–	–	0.231	0.335	0.207	0.289	0.165
Rand	0.223	0.344	0.123	0.341	0.125	–	–	–	–	–
Uniform	0.138	0.266	0.117	0.099	0.152	–	–	–	–	–
MV	0.136	0.278	0.071	0.285	0.121	–	–	–	–	–

(**b**) F_1 on the first test dataset for each demographic individually.

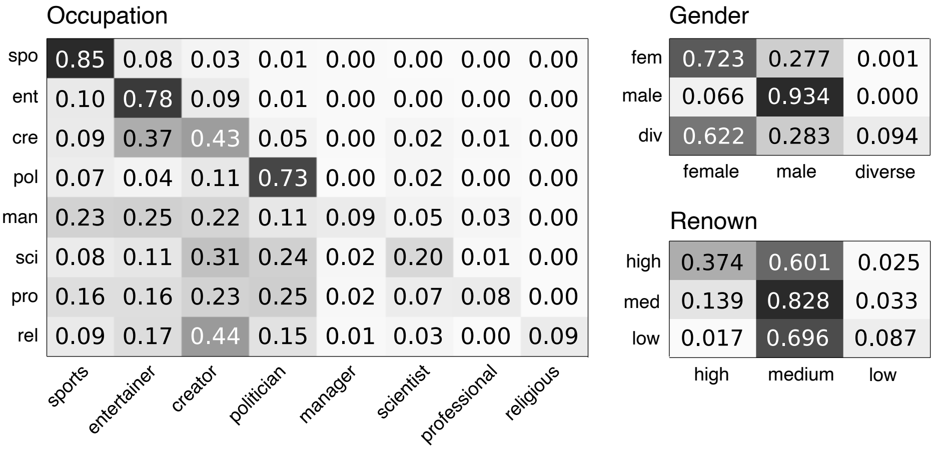
System	Gender			Renown			Occupation							
	fem	male	div	med	high	low	ent	cre	spo	man	pol	sci	pro	cle
Radivchev	**0.88**	**0.95**	**0.30**	**0.87**	0.46	**0.26**	**0.78**	**0.57**	**0.89**	**0.22**	**0.74**	**0.32**	**0.21**	0.27
Moreno	0.82	0.90	0.26	0.48	**0.66**	0	0.79	0.42	0.86	0.22	0.66	0.31	0.17	**0.36**
Martinc	0.81	0.93	0	0.85	0.42	0.16	0.74	0.50	0.87	0.15	0.73	0.27	0.10	0
Petrik	0.85	0.94	0	0.86	0.32	0.09	0.64	0.41	0.78	0	0.62	0	0	0
Fernquist	0.40	0.85	0	0.86	0.26	0.07	0.73	0.44	0.86	0	0.69	0	0	0
Rand	0.27	0.71	0	0.75	0.07	0.06	0.29	0.17	0.39	0.00	0.07	0.01	0.00	0
Uniform	0.29	0.46	0	0	0.28	0	0.16	0.13	0.19	0.03	0.08	0.04	0.03	0
MV	0	0.83	0	0.86	0	0	0	0	0.57	0	0	0	0	0

Table 4.6: Results of the celebrity profiling task. Bold indicates the highest values.

domly draws from a uniform distribution of all classes and reflects the data-agnostic lower bound, (2) Rand randomly selects a class according to the prior likelihood of appearance in the test dataset, and (3) MV always predicts the majority class of the test dataset.

4.3.3 Results and Discussion

Table 4.6a shows the performance of the eight participants who submitted a software to the celebrity profiling task, ranked by the cRank score. The winning approach by Radivchev et al. [178] achieves 0.593 on the first and 0.559 on the second test dataset, closely followed by Moreno-Sandoval et al. [144] with 0.505 on the first, and Pelzer [162] with 0.499 on the second test dataset.



Occupation

	sports	entertainer	creator	politician	manager	scientist	professional	religious
spo	0.85	0.08	0.03	0.01	0.00	0.00	0.00	0.00
ent	0.10	0.78	0.09	0.01	0.00	0.00	0.00	0.00
cre	0.09	0.37	0.43	0.05	0.00	0.02	0.01	0.00
pol	0.07	0.04	0.11	0.73	0.00	0.02	0.00	0.00
man	0.23	0.25	0.22	0.11	0.09	0.05	0.03	0.00
sci	0.08	0.11	0.31	0.24	0.02	0.20	0.01	0.00
pro	0.16	0.16	0.23	0.25	0.02	0.07	0.08	0.00
rel	0.09	0.17	0.44	0.15	0.01	0.03	0.00	0.09

Gender

	female	male	diverse
fem	0.723	0.277	0.001
male	0.066	0.934	0.000
div	0.622	0.283	0.094

Renown

	high	medium	low
high	0.374	0.601	0.025
med	0.139	0.828	0.033
low	0.017	0.696	0.087

Figure 4.2: Averaged normalized confusion matrices for gender, renown, and occupation prediction of the top 5 approaches by cRank.

All submitted approaches beat the baselines, most by a significant margin. The performance measured for our two test datasets is quite similar, comparing participants who submitted runs for both. The scores are less varied on the second test dataset: The leading participant's performance is lower, and Petrik and Chuda's approach improves slightly, overtaking that of Fernquist as fourth in the ranking. These differences can be attributed to the smaller size of the second test dataset and less to the fact that the second dataset contains exclusively English tweets.

Table 4.6b shows the accuracies for all submitted approaches, allowing for a comparison of the general, unweighted correctness of class predictions with the cRank measure. Accuracies are generally higher for all participants, a natural consequence of the imbalanced dataset and the existence of small classes. This can be seen by comparing the results of the baseline-mv, which is almost competitive under accuracy but irrelevant under cRank. The differences in the per-demographic performance can be explained further by inspecting the class-wise F_1 shown in Table 4.6(b). An important observation is that the top three approaches succeed more frequently in predicting small classes correctly, greatly benefiting cRank without notably impacting accuracy. We assume that the good performance on small classes is due to downsampling and the class weighting applied by the top two approaches, whereas models without these strategies mostly fit toward the majority classes. Overfitting toward the majority class is also the likely explanation for the difference in ranking between accuracy and cRank.

Gender. Predicting the binary sex of an author is a widely studied benchmark task for author profiling approaches. All participants achieved a respectable accuracy in predicting influencer gender, frequently surpassing 0.9 accuracy, while F_1 scores are near the 0.6-0.7 accuracy range. Table 4.6(b) shows the class-wise F_1 for all demographics, which explain the achieved performance values on gender prediction: The best approaches are best at predicting diverse gender, while binary gender classification is close to fit for practical use. Interestingly, the averaged confusion matrix for gender in Figure 4.2 shows that diverse influencers are mostly misclassified as female, which is not explained by data imbalance.

Year of Birth. Our approach to age prediction evaluation, departing from fixed-size intervals to a lenient evaluation of year of birth prediction, notably influenced participant systems. Some participants reduce the difficulty of the task by reconstructing intervals and using classification algorithms with notably better performance than the alternatively used strategy of predicting each year individually. No submission tries to solve the prediction with regression algorithms. All models struggle with predicting the year of birth, especially for influencers born before 1980. This is a known difficulty and was addressed by our variable scoring scheme.

Renown. The degree of renown is a particularly imbalanced class, reflected in the accuracy where only four participants could beat the *baseline-mv* on the first test dataset and only three on the second. On the contrary, participants are much better at separating classes correctly as shown by their F_1 scores, although there is a trend toward the majority class as can be seen by the confusion matrix in Figure 4.2. We cannot claim that this task is solved but we have shown that both the most and least famous influencers can partially be distinguished by their writing.

Occupation. As with the other demographics, occupation was predicted far better than the baselines by all participants and the results were highly influenced by the performance on small classes, although not exclusively. All models work better on occupations with a clear topic, like entertainer containing actors and musicians, athletes, and politics. For occupations that cover multiple topics, like creative, manager, professional, and science, all models are rather weak while still beating the baselines. Ignoring the trend toward majority classes, the averaged confusion matrix in Figure 4.2 both show that science is frequently confused with politics and creative, clergy with creative, and creative with entertainer.

In general, all submitted approaches work better for classes with more examples and for classes that can be clearly separated by topic. The final ranking was most influenced by resampling strategies to avoid overfitting to majority classes and by adding grammatical or stylistic features to avoid misclassifying occupations without a topic bias. From a model perspective, we see the most potential for improvement in using all available text data to build influencer representations, instead of just excepts, while still classifying rare classes well, such as few-shot models and prototypical or highway networks. From an author profiling perspective, much is still unclear about the expression of renown, diverse gender, and rare occupations. The best algorithms can partially separate these demographics and errors are systematical rather than random, but a more fundamental understanding of differences in writing is necessary.

4.3.4 Conclusion

The weakly labeled influencer profiling dataset was used in a shared task to demonstrate that it can be used to develop useful and effective profiling technology. In the Celebrity Profiling task at PAN 2019, we invited participants to predict the demographic attributes gender, year of birth, renown, and occupation from 48,335 Twitter timelines of influencers. Eight participants submitted models, and six submitted notebooks describing their approach. Participants found traditional machine learning on content-based features to be the most reliable, with the best performing models adding some style-based features and resampling the training examples to compensate for class imbalance. While much progress has been made on this task, several challenges remain: (1) reliable prediction of rare demographics, such as diverse gender, very young influencers born after 2000, and "rising" stars, (2) the prediction of occupations without clear topical separation, like professional, manager, scientist, and creative, and (3) the classification of authors born before 1980.

Limitations There are some limitations imposed by the task design that could be remedied in the future: reducing the range of years of birth was very broad, e.g. to 1940-2000, omitting the occupations *clergy* and *professional*, and revising the renown boundaries. Having many small classes was a major challenge of this task. We see this as an important aspect of author profiling and especially forensics, since correctly identifying rare demographics is very desirable in practice. Although a certain degree of class imbalance is necessary, the degree of imbalance in all four demographics affected a reliable evaluation and prevented participants from focusing on small classes in particular. To further improve the general robustness and ease of use of our dataset, all non-English tweets and influencers with few textual tweets should be removed.

Besides the prediction of small classes, the *year of birth* has been a major factor influencing algorithm performance. The intention behind our approach was to overcome the inherent weakness of interval-based age prediction and to provide an incentive to participants to develop more fine-grained predictions. Participants did not pick up on this and simply reverse-engineered the scoring function.

4.4 Cross-Population Profiling

The effectiveness of transferring profiling technology to the general population can be evaluated through classification experiments on predicting gender, the most widely studied trait, and predicting influencer occupations. We compare the performance of custom deep learning model with the four best performing models submitted to the most recent PAN author profiling competitions from 2015 to 2018. Instead of retraining the classification models submitted to PAN, we extracted the pre-trained gender inference models from TIRA, where they were originally submitted by the participants of the respective years. Additionally, we train our own baseline gender classification model on influencer profiles. Gender is an appropriate benchmark trait, frequently studied in related work, and a recurring trait prediction task at PAN. We observe a successful model transfer, suggesting that our and PAN's corpora capture the same underlying concept of gender.

4.4.1 Data and Preprocessing

For our experiments, we extracted a subset of 45,475 English-speaking profiles with the traits gender and occupation from the complete dataset and split the subset 70/30 into training and test sets. Table 4.5b compares the size of the PAN datasets and the subset of the influencer dataset, which is by a factor 10 the largest.

Our subset contains 1,379 different occupations, which we manually grouped into eight groups as detailed in 4.3.1: athlete, performer, creative, politics, manager, science, professional, and clergy. We preprocessed

Model	PAN15	PAN16	PAN17	PAN18	Celeb
alvarezcamona15 [7]	**0.859**	–	–	–	0.723
nissim16 [33]	–	0.641	–	–	0.740
nissim17 [15]	–	–	**0.823**	–	0.855
danehsvar18 [51]	–	–	–	**0.822**	0.817
CNN (Celeb)	0.747	0.590	0.747	0.756	**0.861**
CNN (Celeb + PAN15)	0.793	–	–	–	–
CNN (Celeb + PAN16)	–	**0.690**	–	–	–
CNN (Celeb + PAN17)	–	–	0.768	–	–
CNN (Celeb + PAN18)	–	–	–	0.759	–

Table 4.7: Accuracy of (top) the state of the art gender prediction approaches on their respective datasets and transfer performance to celebrities, and (bottom) our baseline deep learning approach, with and without retraining on the PAN datasets.

the text by lowercasing, replacing mentions with `<user>`, hashtags with `<hashtag>`, hyperlinks with `<url>`, number-groups with `<numbers>`, the most frequent emoticons with `<smiley>`, and we removed all punctuation sequences beyond basic English punctuation marks.

4.4.2 Experimental Setting

As baseline models for gender and for occupation prediction, we adapted the convolutional neural network (CNN) for text classification introduced by Kim [102]. Our variant of this model builds on the 100-dimensional GloVe [164] Twitter embeddings, uses four parallel 1D-convolution layers with 128 filters each for 1-, 2-, 3-, and 4-grams, a 64-node dense layer for concatenation after the convolutions, and a final classification layer. The models for occupation and gender only differ in the last classification layer and the loss function used to facilitate the binary classification for gender and the multi-class classification for occupation.

The vocabulary was limited to the most common 100,000 words and padded the word-sequence for each author to 5000 words, which is roughly the average per author word count between ours and the PAN datasets. In the tests on the influencer profiles, this hyperparameter setting achieves more consistent results than fewer or shorter n-gram filters, smaller dense layers, shorter or longer sequence length, or a larger vocabulary. Note that our corpus has labels for more than the two sexes male and female, however, the PAN data did not, so that we excluded profiles with other genders from our experiments, leaving their investigation for future work.

4.4.3 Results and Discussion

Table 4.7 shows all models' transfer performance on gender. In general, all models generalize well to the respectively unseen datasets but perform best on the data they have been specifically trained for. The largest difference can be observed on the smallest dataset PAN15 with less than 1,000 authors, where the model of Álvarez-Carmona et al. [7] suffers a significant performance loss, and PAN16, where the model of Busger op Vollenbroek et al. [33] performs notably better on the influencer data. This was a surprise to us that may be explained by the longer samples of writing per profile in our corpus. This hypothesis is also supported by the large increase in accuracy of the baseline model after retraining for two epochs with the PAN15 and PAN16 training datasets, respectively. The occupation model achieved an accuracy of 0.711.

4.4.4 Conclusion

Overall, the results of our experiments show that profiling models trained on a random sample of people generalize to influencers and vice versa. Our corpus can thus be used for generic author profiling, while providing significantly richer profiles in terms of writing samples and previously unexplored personal attributes. The scale of our corpus allows for the training of deep learning models that, at least on our corpus, outperform the state of the art. We expect that further fine-tuning of the model architecture will yield significant improvements.

4.5 Profiling Influencers based on their Fans' Posts

As profiling technology is ineffective for users who provide little text, we here propose to use the homophily effect of social media to profile a user using the text of connections in the network. We again conducted this analysis as a shared task[6] and extended the "Webis Celebrity Corpus 2019" to include the timelines of fans of the influencers. This section presents the results of the shared task and shows that homophily-based profiling technology far exceed random guessing and, at its peak, reaches the effectiveness of a classifiers trained on and using the original author's texts.

[6] https://pan.webis.de/clef20/pan20-web/celebrity-profiling.html

The goal of the task is to predict three influencer demographics, year of birth, gender, and occupation, given only the original, English tweets of 10 randomly selected, active followers. The training dataset contains the timelines of ten randomly chosen followers per influencer with at least 100 original English tweets for each of the 2,000 influencers, balanced by gender and occupation. Likewise, the test dataset contains another 200 influencers. For consistency, we used the evaluation from the task presented in Section 4.3: the harmonic mean of the macro-averaged multi-class F_1 for gender, occupation, and year of birth.

Three teams submitted a diverse range of models, all outperforming a baseline model trained on the followers' texts, improving strongly above random guessing, and closing in on another baseline trained on the influencers' tweets. We thus demonstrated that the task is, in fact, solvable. An in-depth evaluation reveals similar strengths and weaknesses of the models compared to the previous celebrity profiling task: Topically homogeneous occupations (e.g., *athlete*) are easier to predict than heterogeneous ones (e.g., *creatives*), and younger users are easier to predict than older ones.

4.5.1 Homophily for Author Profiling

All common approaches to author profiling require lots of high-quality text for training from the authors in question. Especially on social media, which is currently the most studied genre in the field, authors with many public, high-quality texts and verified personal demographics are few and far between. With current technology, it is not possible to profile users that write only a few textual posts and only interact by reading, liking, and forwarding the messages of other authors.

Since these passive authors are very frequent on social media, one can profile them only based on other factors. One such factor that provides information about passive authors are the messages posted by other authors who are closely connected to them. Social media theory points out that users with similar demographics and interests form online communities and that online communities develop sociolects (language variation [60]), so inspecting the author's friends, followers, and their social graph relations may also hint at an author's demographics. Since influencers are well-connected, influential, and elevated figures in their communities, they are a suitable subpopulation to study algorithms that profile passive users based on the social graph using the posts of connected authors.

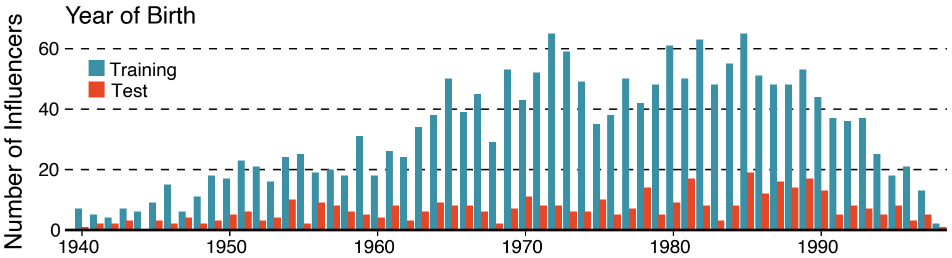


Figure 4.3: Histogram showing the age distribution over both datasets, training and test.

Several studies explore language variation and convergence on social media. Essentially, language variation and convergence explains how groups of people adopt lexical changes and are, together with the psycholinguistic preferences of social groups studied by Pennebaker et al. [163], the reason author profiling is possible. The works that explore language variation have shown, for example, that online language does not convergence to a common "netspeak" but often follows the geographic and demographic [60] similarities of online communities.

Besides real-world factors, a significant impact on lexical variation is attributed to social factors. For example, Pavalanathan and Eisenstein [158] show that lexical variation decreases with the size of the intended audience, which means that social media texts have less lexical variation if they are addressed to a larger audience. Similarly, Tamburrini et al. [213] have shown that an author's words are based on the social identity of the conversion-partner. The specific impact of the network structure on the language variations was studied by [97] who found that language variation is adopted more quickly if individuals are more closely connected.

Based on the related work, it is reasonable to assume that the same linguistic processes of lexical variation and convergence used by Pennebaker et al. [163] to profile individuals based on the individuals' texts also apply to social groups, and it is also possible to profile individuals to a degree based on the social groups' texts.

4.5.2 Evaluation Data

The dataset used in this shared task was again sampled from the corpus presented in Section 4.2. The dataset contains all influencers from the corpus where the year of birth, gender, and occupation are known at the same time. Compared to the previous shared task on on influencer profiling (see Section 4.3), the eligible values for each demographic was reduced to limit the impact of rare labels and scarce text. Influencers with the following demographics were kept:

- *Gender.* From the eight different gender-related Wikidata labels, only *male* and *female* were kept, since all others are rare and too diverse for a meaningful grouping.
- *Occupation.* From the 1,379 different occupation-related Wikidata labels, only those from the *Athlete*, *Entertainer*, *Creative*, and *Politician* categories were kept (see Section 4.3.1).
- *Year of Birth.* Unlike the profiling literature on age prediction, we did not define a static set of age groups, but used the year of birth between *1940* and *1999* as extracted from Wikidata's `Day of Birth` property. Figure 4.3 shows the distribution of the years of birth in the training and test datasets.

To compile the dataset, we added to each selected influencer the Twitter timelines of ten active fans. A fan counted as active with at least 100 original, English tweets and between 10 and 100,000 followers and between 10 and 1,000 followings. The active fans were randomly drawn from the 100,000 most recent followers of each influencer. Finally, the timelines of all remaining followers were downloaded, omitting all retweets, replies, and non-English tweets.

Influencers with less than ten active fans were discarded, which left 10,585 extended influencer profiles to sample training and test data from. From this initial compilation, we selected the largest possible sample of profiles in which all values of occupation and gender are balanced, yielding 2,320 influencers for a balanced training and test split and 8,265 influencers for a supplemental dataset. We split the 2,320 influencer dataset 80:20 into a 1,920-author training dataset and a 400-author test dataset test. The training and supplemental datasets were released to the participants while the test data was kept hidden for evaluation on TIRA.

4.5.3 Evaluation Measures

For comparability with the previous experiments, performance is again measured using the harmonic mean of the per-demographic effectiveness:

$$\text{cRank} = 3 \div (\frac{1}{\text{F}_{1,\text{year of birth}}} + \frac{1}{\text{F}_{1,\text{gender}}} + \frac{1}{\text{F}_{1,\text{occupation}}}).$$

Let T denote the set of classes labels of a given demographic (e.g., gender), where $t \in T$ is a given class label (e.g., female). The prediction performance for $T \in \{\text{gender, occupation}\}$ is measured using the macro-averaged multi-class F_1-score. This measure averages the harmonic mean of precision and recall over all classes of a demographic, weighting each class equally, and thus promoting correct predictions of small classes:

$$\text{F}_{1,T} = \frac{2}{|T|} \cdot \sum_{t_i \in T} \frac{\text{precision}(t_i) \cdot \text{recall}(t_i)}{\text{precision}(t_i) + \text{recall}(t_i)}.$$

We also apply this measure to evaluate the prediction performance for the demographic $T = \text{age}$, but change the computation of true positives: a predicted year is counted as correct if it is within an ε-environment of the true year, where ε increases linearly from 2 to 9 years with the true age of the influencer in question: $\varepsilon = (-0.1 \cdot \text{truth} + 202.8)$.

This way of measuring the prediction performance for the age demographic addresses a shortcoming of the traditional "fixed-age interval scheme:" Defining strict age intervals (e.g., 10-20 years, 20-30, etc.) overly penalizes small prediction errors made at the interval boundaries, such as predicting an age of 21 instead of 20. Furthermore, the precise predictions are not combined with an error function like mean squared error, since predicting the year of birth is more difficult for older users because the writing style becomes more fixed with increasing age and contains less easily identified teen slang.

Baselines Two baselines are used in this task: The baseline N-gram uses n-gram features from the aggregated fan timelines and the baseline Influencer uses n-gram features from the influencer timelines. Both baselines fit a multinomial logistic regression model [159], where the inputs are the tf·idf vectors of the respective tweets. The texts are preprocessed by lowercasing, replacing hashtags, usernames, emoticons, emojis, time expressions, and numbers with respective special tokens, removing all remaining newlines and non-ASCII characters, and collapsing spaces. The tf·idf vectors are constructed from the word 1-grams and 2-grams of all concatenated tweets of the influencers or followers, respectively, with a per-influencer frequency of at least 3. We added special separator tokens to encode the end of a tweet and the end of a follower timeline. Due to the calculation of $F_{1,\text{age}}$, the age prediction was simplified to five classes: 1947, 1963, 1975, 1985, and 1994.

4.5.4 Submitted Approaches

Three participants submitted a software solution to the shared task. Altogether, the submissions were methodologically diverse, covering creative feature engineering, thorough feature selection, and contemporary deep learning methods. As opposed to last year, neither approach is generally superior to the other ones, with each showing individual strengths and weaknesses in some demographics. The overall ranking of the approaches is shown in Table 4.8. The following reviews the submitted systems.

The approach of Hodge and Price [88] utilizes a logistic regression classifier for each individual demographic. The model does not directly use representations of the text, but entirely relies on hand-crafted features as input; specifically: the average tweet length per influencer, the average of all word vectors of the followers' tweets, and the to-token-ratios of the POS-tags, stop words, named entity types, number of links, hashtags, mentions, and emojis. To optimize their model, the authors used 20% of the training dataset for validation in order to pre-evaluate three competing algorithms for each demographic: logistic regression, random forest, and support vector machines. The optimal hyperparameter setting was determined via five-fold cross-validation on the remaining 80% of the training dataset for each evaluated algorithm, where the optimal parameters were determined using the macro-F_1 score. The final model selection on the left-out validation dataset using the official evaluation measures showed that the logistic regression model was best-suited for all demographics.

The approach of Koloski et al. [107] utilizes a logistic regression classifier to predict the age in eight classes, another logistic regression classifier to predict the occupation, and an SVM to predict the gender of the influencers. The model primarily uses lexical representations as features, but limits the input text to 20 tweets per follower and thus 200 tweets in total per influencer. Specifically, the features are computed by (1) preprocessing the text into three versions: the original tweets, the tweets without punctu-

System	cRank	Age	Gender	Occupation
Influencer	0.631	0.500	0.753	0.700
Hodge and Price [88]	**0.577**	**0.432**	0.681	**0.707**
Koloski et al. [107]	0.521	0.407	0.616	0.597
Alroobaea et al. [5]	0.477	0.315	**0.696**	0.598
N-gram	0.469	0.362	0.584	0.521
Random	0.333	0.333	0.500	0.250

Table 4.8: Results of the celebrity profiling task at PAN 2020. Bold scores mark the best overall performance, underlined scores the best performance achieved by a participant.

ation, and the tweets without punctuation and stop words; (2) computing the top 20,000 most frequent character 1-grams and 2-grams and word 1-grams, 2-grams, and 3-grams; and (3) extracting 512 dimensions with a singular value decomposition to be used as features. To optimize their model, the authors first split the training dataset 90:10 into a training and validation set, and used the training split in a five-fold cross-validation to find the optimal n-gram limit, feature dimensionality, and age prediction strategy. Specifically, six alternative feature counts between 2,500 and 50,000 were tested, seven alternative feature dimensions between 128 and 2048, and three different strategies to solve the age prediction task: as a regression task, as a classification task with 60 classes, and as a classification task with eight classes. After optimizing parameters, the authors selected their model based on their performance on the validation dataset, comparing XGBoost, logistic regression, and linear SVMs for each demographic.

The approach of Alroobaea et al. [5] utilizes an LSTM neural network for classification; however, no further details are revealed about its architecture. The model uses exclusively the followers' texts as a tf·idf matrix as input. The text itself is preprocessed by removing links, HTML-style tags, stop words, non-alphanumeric tokens, and typical punctuation marks, replacing mentions with @, and stemming all remaining tokens with NLTK's Snowball stemmer. The authors did not report on any experiments to optimize their model.

4.5.5 Results and Discussion

Table 4.8 shows the results of the participants with successful submissions as well as the performance of the three aforementioned baselines. All par-

(a) Class-wise F_1 and total MAE for year of birth prediction.

Team	1994	1985	1963	1975	1947	MAE
Influencer	0.215	0.632	0.476	0.396	0.129	7.37
Hodge and Price [88]	0.274	**0.463**	0.420	0.319	**0.036**	**9.49**
Koloski et al. [107]	**0.402**	0.389	0.480	0.165	0	11.14
Alroobaea et al. [5]	0	0.111	**0.497**	**0.361**	0	10.89
N-gram	0.362	0.445	0.415	0.226	0	10.12

(b) Class-wise F_1 for gender and occupation prediction.

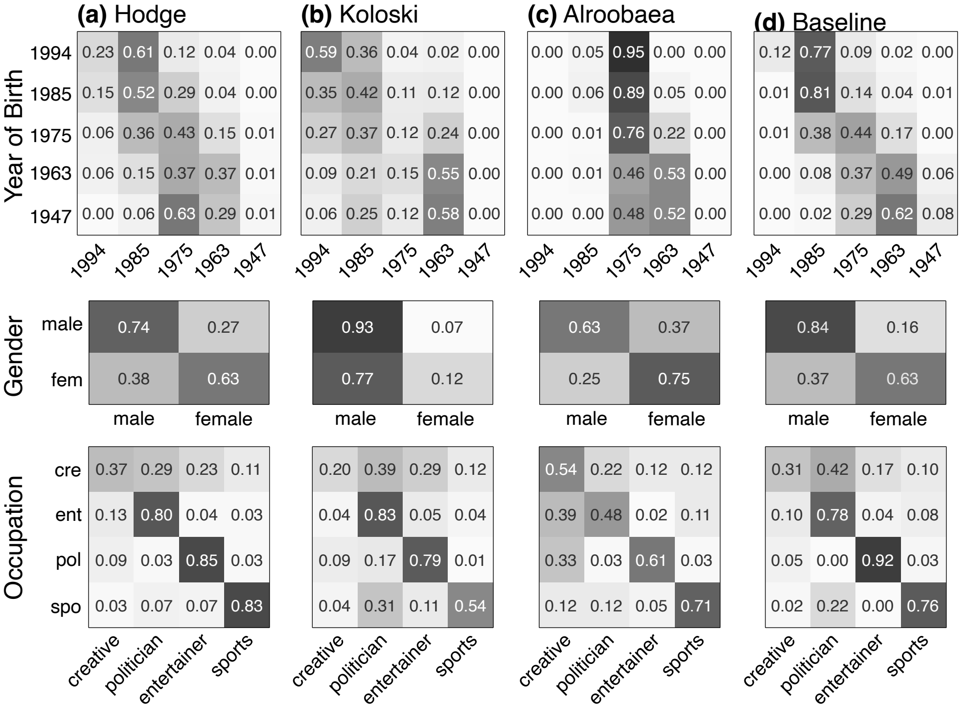
Team	Gender		Occupation			
	female	male	cre	ent	pol	spo
Influencer	0.708	0.762	0.419	0.645	0.864	0.772
Hodge and Price [88]	0.661	**0.697**	**0.457**	**0.731**	**0.776**	**0.830**
Koloski et al. [107]	0.354	0.689	0.292	0.629	0.693	0.632
Alroobaea et al. [5]	**0.712**	0.676	0.454	0.519	0.678	0.721
N-gram	0.434	0.678	0.248	0.578	0.645	0.488

Table 4.9: Class-wise effectiveness for each individual demographic. Listed are the F_1 scores for age, gender, and occupation. For ease of interpretation, the age is evaluated over five classes and the table lists the centroid year of birth of each class together with the mean absolute error (MAE). Bold scores mark the best overall performance, underlined scores the best performance achieved by a participant.

ticipants managed to surpass the Random baseline and improve on the N-gram baseline by up to 0.11 F_1 in the combined metric cRank, in the case of the winning approach. The best performance of the submitted solutions already closes in on the Influencer baseline, which shows that the fans' texts contain noticeable hints about the demographics of the influencer.

Table 4.9 shows the F_1 scores for each individual class. The results show, that, although the submitted approaches are quite diverse, their weaknesses are structural and allow some cautious conclusions about the underlying profiling problem. First, it is easier to predict the age of the youngest influencers from fan tweets than from their own, although age prediction gets increasingly difficult with increasing age. Second, predicting the influencers' gender from their fans' tweets works better for male influencers. Third, predicting the occupation based on fan tweets is competitive with the Influencer baseline.

The best-performing submission by Hodge and Price for predicting the year of birth of the influencers from their fans' tweets achieved an F_1 score



(a) Hodge

Year of Birth	1994	1985	1975	1963	1947
1994	0.23	0.61	0.12	0.04	0.00
1985	0.15	0.52	0.29	0.04	0.00
1975	0.06	0.36	0.43	0.15	0.01
1963	0.06	0.15	0.37	0.37	0.01
1947	0.00	0.06	0.63	0.29	0.01

Gender	male	female
male	0.74	0.27
fem	0.38	0.63

Occupation	creative	politician	entertainer	sports
cre	0.37	0.29	0.23	0.11
ent	0.13	0.80	0.04	0.03
pol	0.09	0.03	0.85	0.03
spo	0.03	0.07	0.07	0.83

(b) Koloski

Year of Birth	1994	1985	1975	1963	1947
1994	0.59	0.36	0.04	0.02	0.00
1985	0.35	0.42	0.11	0.12	0.00
1975	0.27	0.37	0.12	0.24	0.00
1963	0.09	0.21	0.15	0.55	0.00
1947	0.06	0.25	0.12	0.58	0.00

Gender	male	female
male	0.93	0.07
fem	0.77	0.12

Occupation	creative	politician	entertainer	sports
cre	0.20	0.39	0.29	0.12
ent	0.04	0.83	0.05	0.04
pol	0.09	0.17	0.79	0.01
spo	0.04	0.31	0.11	0.54

(c) Alroobaea

Year of Birth	1994	1985	1975	1963	1947
1994	0.00	0.05	0.95	0.00	0.00
1985	0.00	0.06	0.89	0.05	0.00
1975	0.00	0.01	0.76	0.22	0.00
1963	0.00	0.01	0.46	0.53	0.00
1947	0.00	0.00	0.48	0.52	0.00

Gender	male	female
male	0.63	0.37
fem	0.25	0.75

Occupation	creative	politician	entertainer	sports
cre	0.54	0.22	0.12	0.12
ent	0.39	0.48	0.02	0.11
pol	0.33	0.03	0.61	0.03
spo	0.12	0.12	0.05	0.71

(d) Baseline

Year of Birth	1994	1985	1975	1963	1947
1994	0.12	0.77	0.09	0.02	0.00
1985	0.01	0.81	0.14	0.04	0.01
1975	0.01	0.38	0.44	0.17	0.00
1963	0.00	0.08	0.37	0.49	0.06
1947	0.00	0.02	0.29	0.62	0.08

Gender	male	female
male	0.84	0.16
fem	0.37	0.63

Occupation	creative	politician	entertainer	sports
cre	0.31	0.42	0.17	0.10
ent	0.10	0.78	0.04	0.08
pol	0.05	0.00	0.92	0.03
spo	0.02	0.22	0.00	0.76

Figure 4.4: Confusion matrices for gender, occupation, and year of birth.

of 0.432, which is with a distance of 0.07 directly in-between the baselines N-gram and Influencer. Judging by the multi-class F_1 scores shown in Table 4.8, the age prediction task is the most difficult demographic to predict in this dataset. For ease of analysis, we evaluate the age prediction subtask as a five-class problem over the ranges of birth years with the centroids 1994, 1985, 1963, 1975, and 1947.

The results of the multi-class F_1 scores shown in Table 4.8, the class-wise F_1 scores shown in Table 4.9, and the misclassifications depicted in the confusion matrices in Figure 4.4 (top) allow for three observations: First, most submitted models simply perform better on the majority classes. Since no participant employed resampling to balance the training data, this effect may be due to the unbalanced training data. The confusion matrices illustrate this effect, where all models skew towards the center range of birth years, except for the one of Koloski et al., who optimized the age-prediction strategy to achieve the opposite effect: their model skews towards never predicting the center age group. Second, both the N-gram baseline and the model of Koloski et al. significantly outperforms Influencer on age predicting for influencers born between 1990 and 1999.

This observation is not explained by the class imbalance or sampling: Although both, Koloski et al. and the N-gram baseline, resample the age classes from 60 classes down to five or eight, respectively, they still significantly outperform the Influencer baseline, which also reduces the number of age groups to predict. The results do not fully explain this behavior, but it may hint at useful information contained in fan tweets towards better detecting the youngest influencers. However, the increased performance when predicting young influencers does not improve the performance in general, since the Influencer baseline, followed by the model of Hodge and Price, still achieve better multi-class F_1 scores and mean absolute errors. Third, all models poorly predict the oldest influencers born between 1940 and 1955, although, as shown in Figure 4.3, this class has as many subjects as the 1990–1999 year range while covering a broader age spectrum.

The best-performing submission by Alroobaea et al. for predicting the gender of the influencers from fan tweets achieved an F_1 score of 0.696, which is with a distance of 0.057 closer to the Influencer than with 0.112 to the N-gram baseline. Predicting the binary gender has been included as a baseline task since it is very commonly done when predicting demographics, and typically achieves accuracies above the mark of 0.9. Based on the observed results, gender prediction is more difficult for the sampled influencers. The F_1 scores and the confusion matrices, as shown in Figure 4.4 (middle), allow for one observation: The models tend towards predicting a influencer as male rather than as female. This kind of skew is typically explained by imbalanced data or dataset sampling. However, both explanations are unlikely, since our dataset is balanced and has 200 influencers per class, which is usually sufficient to avoid biased data. The best-performing model in this demographic tends to predict female over male, and the Influencer baseline, using the influencers' timelines, does so, too.

The best-performing submission by Hodge and Price for predicting the occupation of the influencers from fan tweets achieved an F_1 score of 0.707, which is marginally better than the Influencer baseline by 0.007. Predicting the occupation is the easiest part of the shared task. Presumably, occupation prediction relies heavily on topic markers in the text, and that these topics are the common ground for discussion between the fans of an influencer. In this respect, it is surprising that the submission supposedly encoding the least lexical but most stylometric features achieved the best performance. The results of the F_1 scores and the confusion matrices shown in Figure 4.4 (bottom) allow for one further observation: Although the class-wise results are mixed between the different submissions, politicians, entertainers, and athletes are consistently predicted well, while creatives are consistently misclassified as either entertainers or politicians. These results are mostly consistent with the results of the 2019 task, albeit, this year, politicians were less frequently misclassified than athletes.

4.5.6 Conclusion

The goal of the task was to determine three demographics of influencers on Twitter based on the tweets of their followers rather than their own: year of birth as a 60-class problem with lenient evaluation, gender as a two-class problem, and occupation as a four-class problem. The models presented are based on a variety of proven methods: feature-based machine learning with stylometric or n-gram features, and LSTMs on tf·idf matrices. The results for individual demographics suggest similar difficulties to those found for the corresponding shared task of 2019: the more thematically diverse *creative* and *entertainer* occupations are harder to profile, as are older authors over younger ones. Our results show that it is possible to profile authors based on their fans' texts almost as well as on their own texts. Technologically, using follower messages to improve author profiling models is a promising future direction.

Our evaluation shows that follower-based profiling models have similar strengths and weaknesses as author-based models for influencer profiling: They work best when the classes are thematically coherent, such as for the *athlete* occupation, but less well in the opposite case, such as for the *creative* occupation. In addition, while predicting the age of influencers is still difficult, the follower-based models tend to predict younger users better than the author-based models in our dataset.

5 Case 3: Trigger Warning Assignment

The third case study is an investigation into assigning trigger warnings to fan fiction documents from Archive of Our Own (AO3) to disclose and advise discretion about emotionally straining or potentially disturbing content not intended for all audiences. The study uses a dataset constructed via weak supervision to assign an appropriate warning from a pre-defined set to a document, based on freeform tags assigned to the works by their authors. The research questions addressed in this chapter are:

Rq 4. Can trigger warnings be effectively assigned to documents via text classification?

Rq 5. How large is the influence of label noise in the dataset on the evaluation of trigger detection models?

For this case study, we use weak supervision to create the "Webis Trigger Warning Corpus 2022" containing 1 million fan fiction works from Archive of Our Own, each labeled with up to 36 warnings from a unified trigger warning taxonomy (see Figure 5.1). The distant knowledge comes from *metadata* knowledge in the form of author-assigned freeform tags and archive warnings, a *database* in the form of relations between the freeform tags across AO3 created by volunteers, and a *curated list* of manually annotated source nodes in this tag graph. Our linking strategy heuristically propagates the labels from the source nodes, along the tag relations, to the 41 million author-assigned freeform tags and reliably labels 1 million documents out of the 9 million in the raw corpus. The strategy is evaluated using both *spot checks* and *weak labels* (see Section 2.2.2).

The first research question is to test document classification for trigger warning assignment. Currently, users must rely on the awareness and rigor of the authors assigning the warnings, as there are no standardized systems for trigger warnings, unlike for movies or video games. We evaluate models in three experiments with varying depth of analysis and breadth of warnings and models covered.

Experiment 1 tests *Violence* and assesses the role of text properties on classification, finding good classification performance (0.94 F_1) in the ideal settings, a limited influence of topic words, which is good for generalization, but also finding that neural classifiers are limited by their input length. Experiment 2 tests four models across all labels and assesses the role of dataset and warning taxonomy properties on classification, finding that the challenges are low recall (false negatives cause more damage than false positives), low effectiveness for rare categories (especially for *Discrimination*), and long document representation. Furthermore, assigning fine-grained warning categories is more desirable but also more difficult than coarse-grained categories. Experiment 3 is a shared task asking participants to submit diverse and specialized systems. The seven submissions confirm previous findings but offer hierarchical models as a promising solution.

The second research question is to assess the influence of label noise in the test data on model evaluation, which is a suspected cause of low model scores. Weakly supervised labeling is likely to introduce noise from errors in the heuristics, for example, when overly cautious authors declare a warning without having substantial textual support for it in the document. We propose a novel LLM-based rank pruning strategy to remove noisy labels from our test dataset, which uses a large language model to quantify how much support for the label is present in each document, and then removes documents with little support. Evaluation on the trigger warning dataset shows that our strategy removes a large number of noisy labels, which in turn reveals model differences that would otherwise be hidden by the noise.

Section 5.1 details how trigger warnings are defined, their significance to online culture, and how they relate to related research in content moderation. Section 5.2 then describes the construction and analysis of the corpus. Section 5.3 describes the three classification experiments and presents the results of the model analysis. Finally, Section 5.4 describes the LLM-based rank pruning method for label noise removal.[1]

[1] The trigger warning corpus and associated code can be found at: `https://github.com/webis-de/ACL-23` and `https://doi.org/10.5281/zenodo.7976807`

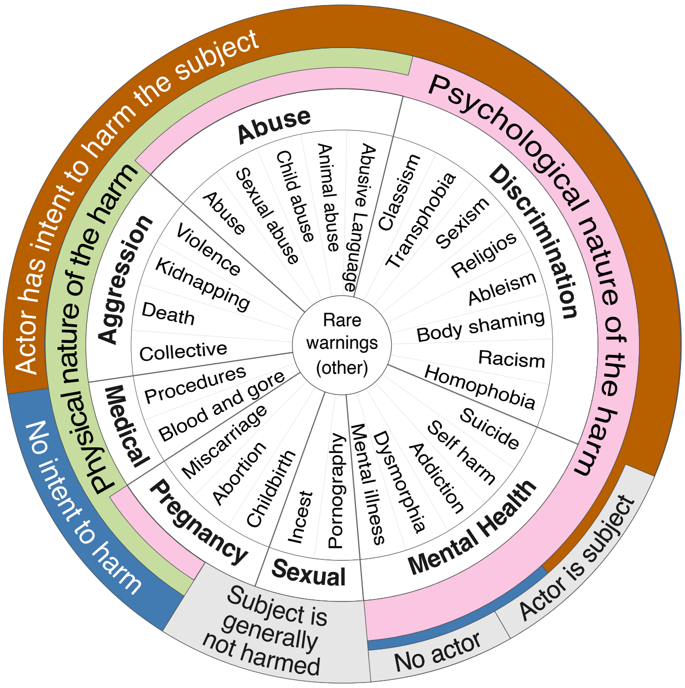


Figure 5.1: Taxonomy of trigger warnings. The inner white ring shows the 29 fine-granular, closed-set trigger warnings, the outer white ring shows the 7 coarse categories. The green and purple ring show the nature of the harm for each warning topic and the blue and orange ring shows the relationship between actor, subject, and the intent to cause harm. The center represents the long tail of the rare triggers, which are omitted for closed-set classification.

5.1 Trigger Warnings in Online Documents

Media of any kind can address topics and situations that trigger discomfort or stress in some people. To help these people decide in advance whether they want to consume such media, so-called content warnings or trigger warnings can be added to them. Trigger warnings were originally used to help patients with post-traumatic stress disorder. But after being picked up by various internet communities to also warn people tending to be "emotionally triggered" by a topic (e.g., to cry), the set of known trauma triggers has grown to include many more, such as abuse, aggression, discrimination, eating disorders, hate, pornography, or suicide. Today, the two terms are often used interchangeably, with "trigger" referring to the semantic cause.

Fiction in particular can make its readers susceptible to triggers. Many readers "lose themselves" in fictional works, identify with their protagonists, and experience their fate with particular intensity. This may partly explain why the community of the fan fiction website Archive of our Own (AO3), where fans write and share stories based on existing characters and worlds from popular media, such as books, movies, or video games, is one of the few where trigger warnings are used proactively and as a matter of course: About 50% of the 7.8 million AO3 works have author-assigned warnings. The other half, however, do not, and neither the AO3 moderators nor the readership seem willing or able to fill that gap.

Related Work

Constructs related to "trigger warnings" have been investigated using computational approaches under different terms and have spanned a broad range of phenomena. Recent research employs terms such as "objectionable content", "objectionable material", "harmful content", "harmful text" [14, 104, 209] as broad terms covering diverse types of content that can potentially evoke negative emotions in the recipient of the material (be it verbal or visual), i.e. cause emotional harm at different degrees of severity. The type of content that is often subsumed under those terms includes violence, sexual content, misguided messages, misinformation, verbal aggression, malice, callousness, or social aggression, among others. And while there is also a clear link to sentiment analysis, phenomena subsumed under "objectionable/harmful content" lie only on one end of the sentiment scale (that of negative sentiment), however, have a finer granularity (cf. range of specific types of content, mentioned above, that may evoke harm).

Now, the notion of "triggering" is equally underspecified (open-ended), but even broader. While most of the objectionable types are indeed unobjectionably harmful—in that they can be linked to *intention to harm*—there may exist concept associations that are triggering to some individuals which, objectively speaking, have little to no link to intention to harm; consider, for instance, that a mention of a thunderstorm may be triggering to a victim of a severe lightning injury. Thus, triggering covers also concepts which would normally be understood to lie at the positive end of a sentiment scale, which can, however, evoke negative associations in some individuals due to their specific traumatic past experience related to the concept. A "trigger warning" just gives a nominal label to the signal that is considered triggering. While we are not aware of prior work on automatic trigger warning assignment nor specifically violence warning assignment, below we outline prior work in NLP and computer science that covers most closely related topics.

Pioneering work on automatic trigger warning assignment is Stratta et al.'s (2020) user study with a browser plugin (DeText) on generic websites. The authors conclude that client-side warnings are feasible and that users respond positively. However, this work is very limited in that *Sexual assault* is the only warning given using a naive dictionary-based approach. Similarly, De Choudhury [54] investigates behavioral characteristics of the anorexia affected population on Tumblr. Analysis of several thousand posts has shown that the platform contains vast amounts of triggering content which may prompt and/or reinforce anorexia-oriented lifestyle choices. Two sub-groups of the anorexia community were identified—pro-anorexia and pro-recovery—with distinguishing affective, social, cognitive, and linguistic properties. Predictive models based on language features extracted from the posts were able to detect anorexia content at 80% accuracy.

Charles et al. [41] recently proposed the Narrative Experiences Online (NEON) taxonomy of multi-media trigger warnings. Its two tiers are synthesized like in ours from 136 guidelines on the web, consisting of 14 top tier categories (versus our 7) and 76 subcategories (versus our 36). However, unlike ours, NEON's subcategories are not explicitly grounded in warnings that are used on a daily basis by millions of people. Moreover, its categories are non-disjoint, not clearly semantically motivated classes with blurred definitions: For instance, compare category "*4. Disturbing content: Content contains imagery, sounds, or effects that may frighten, disgust or scare*" with category "*9. Parental guidance: Content may not be appropriate for children*". Since our two teams worked in parallel, the synthesis of our complementary taxonomies is a fruitful direction for future work.

Harmful Content Trigger warnings can be seen as orthogonal to other harmful content taxonomies, e.g., for violence, hate speech, or toxicity, where some labels overlap but differ in structure and entailment. Banko et al. [14] present a comprehensive taxonomy of harmful online content that has notable overlap with our taxonomy but focuses on online speech. Triggering content, however, can be narrative and does not require an intent to harm to evoke disturbing images. Mollas et al. [143] study the detection of violence and present the ETHOS dataset of YouTube and Reddit comments with crowdsourced multi-label annotations about verbal violence and its target. Based on Wulczyn et al.'s (2017) work, the Toxic Comment Classification Challenge [1] dataset covers different content moderation topics. It contains 223,000 Wikipedia comments (sentence to paragraph level) annotated with six toxicity subtypes.

Multi-label Document Classification Our multi-label classification (MLC) task has (comparably) few labels overall and few labels per document, but it features long documents. The main difference to other MLC datasets is the document genre (fan fiction) and the label domain (trigger warnings). The most similar MLC datasets (with mostly shorter documents) are Reuters RCV1 [114] with 80,000 news articles and 103 topic labels, its predecessor Reuters-21578 with 11,000 news articles and 90 labels, and the Arxiv Academic Paper Dataset (AAPD) [253] with 56,000 abstracts from computer science and 54 labels. Recent meta-studies on long document classification [50, 156] find that sparse-attention transformers, hierarchical models, and input selection methods have little difference in effectiveness to input truncation. Galke and Scherp [73] compare graph and "bag of words" (BoW) methods with transformers, noting that BoW methods are (often) not far behind.

Further multi-label classification datasets cover tasks with very large label sets: EUR-Lex [136] with 15,000 law documents and 4,000 labels, its successor EURLEX57K [39] with 57,000 law documents and 4,300 EUROVOC labels, MIMIC-III [92] with 112,000 clinical reports and 11,600 ICD-9 codes as labels, and the Extreme Labels [28] collection of datasets for product and Wikipedia article classification. Recent work on large label sets addresses label-dependent document representations [251], loss functions for long-tailed label distributions [89], prompt-based few-shot learning for rare labels [254], and sequence labeling with an attention encoder–decoder LSTM for many-label document MLC [253]. Transformer-encoder classifiers are common baselines [39].

Violence and Emotions While affect and emotion recognition in non-fiction text—sentiment analysis more generally—has been long studied in NLP [6], research into interactions between emotions and their triggering cause events was introduced only about a decade ago [113]. Cause events here refer to (verb) arguments or events in the text that are highly correlated with a certain emotion, positive or negative. The goal of the emotion cause extraction task is to identify the emotion's stimulus and the computational methods range from rule-based lexico-syntactic approaches through traditional classifiers to recently also deep learning; see Khunteta and Singh [101] for an overview of the emotion cause extraction area. By contrast the trigger warning assignment task is rather about identifying potentially triggering content which may evoke strongly negative emotions in readers.

Interest in broadly understood verbal violence—although not explicitly referred to as such—has a long history in the NLP community. Waseem et al. [232] and Kogilavani et al. [106] propose taxonomies of abusive and offensive language, respectively; Kogilavani et al. also survey techniques for offensive language detection. Fortuna and Nunes [70] and Schmidt and Wiegand [203] provide an overview on hate speech and Mishra et al. [141] more generally on abuse detection methods with "abuse" defined as "any expression that is meant to denigrate or offend a particular person or group". While not considered from the point of view of triggering, this definition fits the category 'Hateful language" listed in the institutional guidelines. While most work on verbal violence has been carried out in the context of social media (methods ranging from feature engineering to neural networks) it would be useful to extend those systems to cover a broader range of verbal violence, e.g., literary dialogue, in the context of the trigger warning assignment task.

5.2 A Corpus for Trigger Warning Assignment

We constructed a corpus of documents with trigger warnings based on data from Archive of Our Own (AO3), a public online anthology of fan fiction, i.e., amateur writings inspired by existing works of fiction: e.g., novels, cartoons, manga. At the time of corpus creation, AO3 hosted about 8 million works. Aside from basic meta-data, such as title, author, language, statistics (number of words, chapters, etc.), reader reactions, ratings, fandoms (original source(s)/inspiration), and relationships (characters involved in romantic/platonic relationship(s)), crucially for this research, works are labeled with *Archive Warnings* and freeform *Additional Tags*.

A manual examination of a sample of the freeform tags on AO3 showed that a considerable fraction are trigger warnings. Authors often append qualifiers to their warnings, which may indicate the nature of a trigger or its connection to the narrative of their work. These tags are manually liked with tags from a controlled subset, the canonical tags, and the links are determined by community volunteers called "tag wranglers". However, many canonical tags are semantically redundant, extensive, and too sparsely populated with works to use them as our set of trigger warnings to study.

We therefore first synthesize an authoritative hierarchy of 36 trigger labels based on guidelines from relevant institutions, the outcome of which is a two-tier taxonomy, which firmly grounded in real-world trigger warning assignment (see Section 5.2.1).

We then create a large corpus of fan fiction by systematically downloading the works from AO3 with all its metadata (see Section 5.2.2), and then embarked on a semi-automatic mapping of the millions of freeform tags to this condensed set (see Section 5.2.3).

5.2.1 A Taxonomy of Trigger Warnings

While the notion of "trigger warning" in digital media has been around for a decade, none but one recent attempt has been made to propose a "standardized set" [41] due to the open-ended nature of the issue. Most warning labels stem from internet communities, such as social media, gaming, and online-content readers and writers. Not surprisingly, such *community-supplied* labels have all the properties of user-generated content, in particular, heterogeneity and lack of linguistic uniformity, which makes them hardly usable as a set of classes for training classifiers. However, since the arousal of a debate on the use of trigger warnings in educational settings, many universities issued explicit guidelines on their use. We take eight such *institutionally-recommended* guidelines and frequently referenced lists of warnings as authoritative trigger warning sources and consolidate their label sets in a principled way.

Figure 5.1 shows the resulting 36-label taxonomy, consisting of 29 narrowly-defined (closed-set) categories for frequent warnings and 7 more general, higher-level (open-set) labels. The 29 closed-set labels have clear semantics, which is advantageous for classification and practical from the point of view of usability. The 7 open-set labels also match documents that are related to but do not match any of the closed-set labels. This open-set semantics is essential for trigger warnings since traumatic imagery can be evoked by a variety of individually-rare topics (hence the large dimensionality of user-generated warnings). The 7 open-set labels, e.g. *Sexual*, constitute a level of abstraction for the closed-set labels, e.g. *Incest* and *Pornography*; a coarse variant of the label set.

Sources of Trigger Warnings We collected guideline documents on trigger warning assignments from eight universities from the English-speaking world: Cambridge, Manchester, Michigan, Nottingham, Reading, Stanford, Toronto, and York. Table 5.1 illustrates the guidelines, processing, and references. We identified these documents by, first, compiling a list of the top 30 universities according to Times Higher Education [216], QS World University Rankings (2023), and the Russel Group [197] members and, second, searched those universities' domains for combinations of 'trigger', '(sensitive) content', 'warning', 'guide', and 'recommendation'.

Cambridge [38]	**York** [43]	**Stanford** [219]	**Reading** [189]	**Michigan** [127]	**Manchester** [132]	**Toronto** [46]	**Nottingham** [153]	**Merged**	**Labels**
		Discrimination (sexism, racism, homophobia, transphobia)						discrimination	**Discrimination**
		Stereotypes (gender, race, national origin, age)						stereotypes	
	hateful language or behaviour (e.g. racist, sexist, homophobic or transphobic language/behaviour)								
discrimination / bigotry (racism, misogyny, homophobia, transphobia, ableism, anti-Semitism, Islamophobia)								bigotry	
homophobia	*homophobic* lang./behaviour	*homophobia*	Homophobia and heterosexism	Homophobia and heterosexism	Homophobia	homophobia	Homophobia, biphobia, or heterosexism	Homophobia and heterosexism	homophobia
–	–	–	–	–	Heterosexism	–	Heterosexism	Heterosexism	
transphobia	*transphobic* lang./behaviour	*transphobia*	Transphobia and trans misogyny	Transphobia and trans misogyny	Transphobia	–	Transphobia	Transphobia and trans misogyny	transphobia
–	–	–	Classism	Classism	Classism	–	Classism	Classism	classism
–	*sexist* lang./behaviour	*sexism*	Sexism and misogyny	Sexism and misogyny	Sexism	–	Sexism and misogyny	Sexism and misogyny	sexism
misogyny	–	–	–	–	Misogyny	–	–	Misogyny	
–	–	–	*Stereotypes gender*	–	–	–	–	–	
racism	*racist* lang./behaviour	*racism*	Racism and racial slurs	Racism and racial slurs	Racism and racist slurs	depictions of racism or oppression	Racism and racial slurs	Racism and racial slurs	racism
–	–	*Stereotypes race and national origin*	–	–	–	–	–	– Race stereotypes	
–	–	–	Hateful language at religious groups	Hateful language at religious groups	–	–	Hateful language at religious groups (e.g. Islamophobia, antisemitism)	Hateful language at religious groups	religious
Islamophobia	–	–	–	–	–	Islamophobia	*Islamophobia*	islamophobia	
anti-Semitism	–	–	–	–	–	–	*antisemitism*	anti-semitism	
ableism			Mental illness and ableism	Mental illness and ableism	Ableism		Ableism	ableism	ableism
–				*fat phobia*	*fat phobia*			fat phobia	body-shaming
–		*body shaming*						body shaming	
–		*age*						age	–
disordered eating	–	Eating-disord. behavior or body shaming	Eating disorders and body hatred	Eating disorders, body hatred, and fat phobia	Eating disorders, body hatred, and fat phobia	Eating disorders	Eating disorders and body image	Eating disorder, body hate	**Mental-health** dysmorphia
…									

Table 5.1: Creation of the "Discrimination" warnings. Shown are the verbatim statements from the lists, segmented into on concept per row. Terms in multi-term concepts not matching the grouping were removed and *re-inserted* as new concept.

The Structured Set of Warning Labels Since all guidelines follow a different structure (from paragraphs to term lists) and granularity, we manually processed the documents to (1) extract and segment the warnings, (2) align and merge warnings that are closely synonymous (e.g., *Transphobia* with *Transphobia and trans misogyny*) across documents to create the 29 closed-set labels, and (3) group related warnings to form the 7 open-set label groups.

We extract two units: triggering content concepts and concept groups. Concepts are all terms (*Homophobia*) or phrases (*Death or dying*) that refer to a singular semantic field. Concept groups are (structural) groupings of related concepts with a dedicated group name (*Discrimination* (*sexism, racism, homophobia, transphobia*), where *Discrimination* is the group name). We extract concepts from the groups and add them to the list of all concepts. Items of structured lists (same bullet point) or concepts in coordinating conjunctions are not segmented, assuming they belong to the semantic field that defines the warning.

We generally group concepts that are mentioned together in a concept group and use this group's name to determine the open-set label. Concepts are split if a term in a concept did not match the group's intention, e.g. *Body-shaming* is split from *Eating disorders and body shaming* and grouped with *Discrimination*. We create the *Sexual* and *Childbirth* groups and then assign the remaining concepts to the most closely related group. Since we are looking for labels with support ("consensus") across different sources, we ignore concepts with singular occurrence.

Properties of the Warning Labels Four major observations can be made: First, the granularity of triggers is not uniform (e.g., both *Abuse* and the more specific *Child abuse* are included). Second, the set comprises subsets of related concepts which lend themselves to semantic abstraction (e.g., *Sexism*, *Classism* and other *-isms* and *-phobias*). Third, the guidelines are not exhaustive (as they point out themselves) due to the open-set nature of traumatic events and triggering imagery. For this reason, we consider the 7 (coarse-grained) categories as a part of the whole set (instead of just a hierarchy tier): they add the needed open-set semantics (e.g., *Bullying* is discrimination but would not be covered by the closed-set categories). Fourth, the (lexical) semantic field of the labels is not precise enough to be the sole base for document annotation. We developed sharper definitions based on the annotation procedure in Section 5.2.3, which are shown in Table 5.2. Figure 5.1 also shows an additional abstraction of the label definitions in two dimensions: the nature of the harm in the content (physical/psychological) and the relationship between actor, subject, and intent.

Trigger warnings	Definition and Example Tags
Aggression-related	
Violence	Physical violence and destruction. *Manhandling, Slapping, Vandalism, Torture*
Kidnapping	Kidnapping, abduction, and it's consequences. *Captivity, Hostage situations*
Death	Graphic death, murder, and dying characters. *Drowning, Decapitation, Corpses*
Collective-violence	Organized violence by groups. *Terrorism, Civil war, Gang violence*
Other-aggression	*Violent thoughts, Slavery, Cannibalism*
Abuse-related	
Abuse	General abusive treatment. *Domestic Abuse, Bullying, Compulsion, Humiliation*
Sexual-abuse	Abuse and assault with sexual intent. *Rape, Sexual harassment, Voyeurism*
Child-abuse	Abuse of a child. *Child neglect, Pedophilia, Grooming, Child marriage*
Animal-abuse	Mistreatment and death of animals. *Animal Sacrifice, Harm to animals*
Abusive-language	Verbal abuse and strong language. *Threats of rape/violence, Insults, Hate speech*
Other-abuse	*Extortion, Intimidation*
Discrimination-related	
Classism	Discrimination based on social class. *Rich/Poor, Caste divide, Social hierarchies*
Transphobia	Discrimination against transgender persons. *Misgendering, Deadnaming*
Sexism	Discrimination based on gender stereotypes. *Misogyny, Slut shaming*
Religious	Discrimination based on religion. *Islamophobia, Antisemitism, Anti-Catholicism*
Ableism	Discrimination against disabled persons. *Ableist slurs, Ableist language*
Body-shaming	Discrimination based on body properties. *Fat-shaming*
Racism	Discrimination based on race. *Racist Language, Segregation, Xenophobia*
Homophobia	Discrimination against homosexuality. *Homophobic Language, Gay Panic*
Other-discrimination	Discrimination against other or general groups. *Stereotypes, Bigotry*
Mental Health-related	
Mental-illness	Severe mental illness with institutional treatment. *Insanity, Psychosis*
Dysmorphia	Body dissociation. *Dysmorphia, Dysphoria, Eating disorder*
Addiction	Substance or gambling addiction and abuse. *Drug abuse, Withdrawal*
Self-harm	Self-destructive acts or behavior. *Cutting, Self-destruction*
Suicide	Suicide attempt, ideation, conduct, and aftermath. *Suicide*
Other-mental-health	Psychological issues that require help. *Depression, Trauma, Survivor guilt*
Sexual-related	
Pornography	Graphic display of sex, plays, toys, kinks, technique descriptions.
Incest	Sex between family members. *Sibling Incest, Twincest*
Other-sexual	Non-graphic mentions of/ discussions about sex. *Sex shop, Sex education*
Pregnancy-related	
Miscarriage	Death of the unborn and unplanned termination of pregnancy. *Stillbirth*
Abortion	Planned termination of pregnancy. *Abortion*
Childbirth	Being pregnant and giving birth. *Pregnancy, Childbirth*
Other-pregnancy	Fertility, recovering from pregnancy, and issues with newborn. *Fertility Issues*
Medical-related	
Blood and gore	Display of gore. *Blood, Open wounds*
Procedures	Medical procedures. *Amputation, Stitches, Surgery*
Other-medical	Illnesses and injuries. *Cancer, Hanahaki disease*
Other-content-warning	Crime, Police, Weapons, Needles, Prisons, Fluff, Politics, …

Table 5.2: The 36 warnings with *example canonical tags*. Since trigger warnings are an open-set problem, some other verbatim warnings (see Table 5.7) on Archive of Our Own are not part of our taxonomy.

Corpus size	
Words	58B
Total works	7.9M
- with closed-set warnings	2.8M
- with open-set warnings	281K
- without warnings	4.7M

Filter criteria	
Non-English language	751K
Publication pre-2009	246K

Filter criteria	
More than 100 chapters	3K
More than 93k words (top 1%)	79K
Less than 50 words (bottom 1%)	122K
More than 66 tags (top 1%)	8K
More than 10% unclean tags	4.7M
Less than 3 tags (conf. thresh.)	2.3M
Less than 5 kudos (popularity	632K
Less than 100 hits threshold)	751K
Duplicates	8K

Table 5.3: Selection of corpus statistics of the Webis Trigger Warning Corpus 2022.

5.2.2 A Corpus of Fan Fiction Documents

Our inspiration for operationalizing trigger warnings is based on finding "hidden in plain sight" a large collection of fictional works with millions of manually assigned warnings that have accumulated for years on the widely known fan fiction website Archive of our Own (AO3), which to our knowledge have not previously been used as a basis for automating a task. We therefore first compile a near-complete corpus of AO3 fan fiction (i.e., fanfics, documents) and its metadata, namely language, length, comments, hits (i.e., reads), kudos (likes), (chapter) publication date(s), and, notably the freeform *Additional tags* and fixed-set *Archive Warnings*.

The *Archive Warnings* are a set of six content warnings pre-defined by AO3. Authors must actively assign at least one to each of their works. The labels are:

1. *Major Character Death* when the death of a character is part of the story.
2. *Underage* when works contain sexual activity by characters younger than 18.
3. *Rape/Non-Con* when non-consensual sexual activity is described.
4. *Graphic Depictions of Violence* when gory, explicit violence is described.
5. *Creator Chose Not To Use Archive Warnings* when the work describes content that may warrant a warning, but the author chose to omit the warning to avoid any spoilers.
6. *No Archive Warnings Apply* when the work has no triggering content.

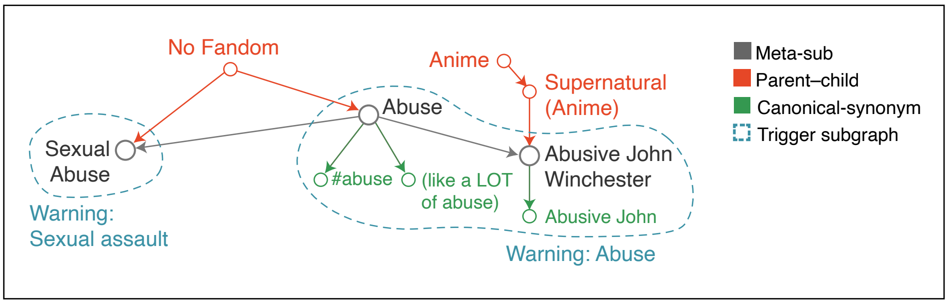


Figure 5.2: Excerpt of AO3's tag graph. The edges are the three relations added by tag wranglers, connecting three subgraphs of freeform tags to gray canonical tags.

The *Additional tags* allow authors to define open-set, freeform content descriptors, which are used as keywords for search and browsing, like *romance, slow burn, fluff, and jealousy*, but also to assign additional trigger warnings like *abandonment, monsters, blood drinking*. Additional Tags are heterogeneous, user-generated content but frequently used tags are "canonized" by volunteer "tag wranglers". The use of canonized tags is encouraged and supported by the web interface.[2]

The corpus also contains the tag graph spanned by the author-assigned freeform tags. Illustrated in Figure 5.2, the tag graph defines three relations between tags: canonical-synonym, parent–child (i.e., fandom and media-type relations), and meta-sub relations which form a hierarchy of meanings. All relations form acyclic digraphs where canonical tags form a controlled subset to connect the freeform tag subgraphs. Tag relations are manually created and maintained by volunteer community experts (the so-called "tag wranglers") following specific guidelines [217]. We consider this data a highly reliable basis for our subsequent distant-supervision annotation of trigger warnings. The final corpus contains about 8 million works totaling 58 billion words. Table 5.3 shows selected corpus statistics.

Scraping the Works The scraper collected all public works from AO3 using each work's unique URL, which is based on its permanent and unique ID. The IDs were identified using the AO3-search: it returns all works created within a time range when passing a `created_at:`DATE-RANGE query parameter but no query terms. Individual searches were started for each day since the site's creation (August 13, 2008, and August 09, 2021) to concur with AO3's crawling limits. URLs which were not publicly accessible, redirected to external sites or yielded HTTP errors were omitted. The

[2] `https://archiveofourown.org/wrangling_guidelines/2`

Sample	Nr. tags in set (% of all)				Warnings (% of set)			
	Tag occurrence		Unique tags		Closed		Open	
0-2k	27.6M	(51.98)	2K	(0.02)	538	(26.71)	82	(4.07)
10-11k	0.3M	(0.56)	1K	(0.01)	127	(12.70)	19	(1.90)
Tag graph	41.0M	(77.18)	2M	(20.17)	241K	(12.30)	33K	(1.68)
All tags	53.1M		9.7M		–		–	

Table 5.4: Number of AO3 free-form tags that can be annotated with a trigger warning by different methods. The samples 0-2k and 10-11k contain manually and Tag graph distantly supervised annotations.

complete crawl contains 7,866,512 works and the most active day yielded about 10,000 works. Finally, the scraper archived the web pages' HTML in WARC files using ChatNoir Resiliparse [24] and parsed the HTML using Scrapy to extract each work's text and metadata.

In addition to the works, the relevant section of the tag graph were also collected by scraping and parsing the HTML page of each tag that was used in one of the works. A tag's page lists all relations of that tag so that the relevant section of the tag graph can be reconstructed.

Deduplication The deduplication removed 8,011 full and near duplicates from the crawl. The 4,249 full duplicates were identified using SHA-256 fingerprinting. Near-duplicates include pairs of works whose text differs only to a very small extent so that neither the meaning and especially not the assigned warning labels changed. We identified them by applying MinHash [206] with 8 buckets and considered resulting pairs as near-duplicates if their Jaccard similarity exceeded 0.6 or if their cosine similarity exceeded 0.875. This approach favors precision over recall and ultimately identified 3,762 near-duplicates.

5.2.3 Linking the Freeform Tags to Trigger Warnings

To determine the trigger warnings for each work, we created a table that maps all freeform tags to all semantically matching trigger warnings from our taxonomy. Creating this mapping table is a three-step process:

1. Manually annotate the 2,000 most common tags. This is feasible and greatly boosts the reliability of the dataset, since the 2,000 most common tags cover about 50% of all tag occurrences (see Table 5.4)

2. Automatically annotate 2.0 million unique freeform tags via weak supervision by identifying substructures of the tag graph so that each node in the substructure maps to the same trigger warning. This way, only one tag must be initially mapped and all others can be automatically mapped.

3. Merge both results while giving priority to the manual annotations.

If this process fails to map a freeform tag to a warning, we assume that the tag does not describe harmful content and discard it.

Annotating Common Tags Two samples of freeform tags were manually annotated: the 2,000 most frequent tags (`0-2k`), which cover just over 50% of tag occurrences, and the 10,000th-11,000th most frequent tags (`10-11k`) used to evaluate the weak supervision approach.

We used an iterative annotation process with two annotators that jointly developed annotations and annotation guidelines (see Table 5.5). First, two annotators individually annotated each tag by assigning it a trigger from the taxonomy based on the guidelines, or, initially, their own understanding of the problem. Then, both annotators discussed and resolved every disagreement and updated the annotation guide (see Table 5.5). These steps were repeated for two more rounds until disagreement was negligible.

The first annotated sample `0-2k` contains 538 tags annotated with one of the 29 closed-set triggers and another 82 open-set ('other') triggers. The ratio of tag-to-trigger assignments reduces by about half for less frequent tags and stabilizes at 9–16%. Table 5.2 shows the resulting definitions and example tags for each label.

Identifying Substructures of the Tag Graph The tag graph was split into "trigger graphs", which are rooted subgraphs where all tags belong to a related concept that maps to the same trigger warning as the only source node, i.e., its root, and all relations are directed. Figure 5.2 shows an excerpt of such trigger graphs and the different relations between the nodes for the example case of *Abuse*.

The source nodes were manually annotated and the respective warnings were assigned to all successors of the source. Trigger graphs were identified in five steps:

1. Group all tags with a synonym relation and identify the canonical tag. In every set of synonyms, one tag is marked by the tag wranglers as

General Guidelines

- Exclude general trigger tags without topic specification: *Triggers, additional warnings in author's note, additional warnings apply, other: see story notes, ...*
- Exclude ambiguous (triggering and non-triggering) tags: *Stuffing, hardcore, kinky, crazy, coping.*
- Annotate ambiguous (different topics) tags with all options: *Asphyxiation is sexual and death.*
- Exclude tropes: *Whump, hurt-comfort, ...*
- Exclude tags that declare the setting of the work: *Post-world war 2.*
- Annotate explicit warnings, not implied or associated: *weapons, safehouses is not violence.*
- Annotate fantasy concepts like the inspiration: *Male pregnancy is pregnancy, Hanahaki disease is medical, species dysphoria is mental-health.*

Aggression

- Aggression is only physical: *Psychological violence is abuse, threats of violence is abusive-language.*
- Execution devices are Death: *Guillotine, electric chair is death.*
- Weapons are violence if the tag mentions violence: *Gun violence is violence.*
- Annotate loss and grief as mental-health, even if death is implied.
- Annotate potential or uncertain death as death: *Possible character death is death.*
- Annotate intended deadly violence as death: *Murder, assassination.* If graphic violence is directly indicated, annotate death and violence: *Fight to the death.*
- Exclude tags where the death is a descriptor of the setting or a character: *Dead Link.*
- Organized violence is collective-violence: *Acts of war, organized crime, drug-related crime.*
- *Human trafficking is kidnapping.*

Pregnancy

- Lactation and fertility (issues) and interactions/issues with newborns are *Pregnancy.*

Table 5.5: Guidelines for annotating the freeform tags. The general principles take effect unless there is a label-specific exception.

the canonical version. All other synonyms are terminal nodes the synonym graph and direct successors of the canonical tag.

Abuse

- Forcing others to act is abuse, including fantasy concepts: *Slavery, mind-control, compulsion*.
- If the forced action is sexual in nature, annotate sexual-abuse: *Non-consensual ...*.
- Annotate the more specific abuse label (sexual, child, animal) instead of *Other-abuse*.
- Stalking, voyeurism, and rape is *Sexual-abuse*.
- Sexual abuse of children is *Child-abuse*.
- Hate-speech, threads, and intimidation are abusive-language. Hate speech towards a group is both, abusive-language and discrimination: *Racist slurs are racism and abusive-language*.

Mental-health

- Annotate mental-illness if the affliction requires stationary treatment: *Schizophrenia, psychosis, ...*
- Annotate mental-health otherwise: *Depression, anxiety attacks*
- Exclude stress, angst, or anxiety.
- Substance abuse is addiction. Exclude recreational, non-abusive substance use.
- Highly addictive drugs (heroin, ...) are always addiction.
- Exclude medical drug use, unless 'self-medication' is stated.
- Annotate (sex/gender/species) dysphoria and eating-disorder as dysmorphia.

Sexual

- Annotate all tags as pornography if they indicate a sex act without intent to harm.
- Sex toys are pornography.
- Sexual position preference (*Top, Bottom*) are pornography.
- Kinks are pornography if the kink is impossible to practice without sex.
- Kinks that do not require a sexual act are other-sexual: *Size kink, Praise kink, Plushophilia*

Medical

- Annotate medical if there is no intent to harm. Acts of harmful mutilation by others are aggression or abuse.
- Injuries and (chronic) illnesses are *Medical*, but exclude mild afflictions like *Allergies*.
- Exclude equipment: *Band-Aids, Needles*
- Wounds and open injuries are blood-gore.

Table 5.6: Guidelines for annotating the freeform tags (cont.). The general principles take effect unless there is a label-specific exception.

Sample	Prec	Rec	F_1	Acc
Fine-grained				
`0-2k`	0.94	0.94	0.94	0.94
`10-11k`	0.96	0.96	0.96	0.96
Coarse-grained				
`0-2k`	0.95	0.95	0.95	0.95
`10-11k`	0.96	0.96	0.96	0.96

Verbatim warnings	Tag	Unique
	occur.	tags
Total	62,316	27,694
Classified as warning	34,806	9,595
- of all wrangled	0.86	0.79
- of all free-form	0.56	0.35

Table 5.7: Effectiveness of the distantly supervised classification on two manually annotated tag sets (left). Number of verbatim warnings (e.g., 'warning', 'tw:', ...) annotated as a warning by our method (right).

2. Identify source nodes in the meta–sub graph. Meta–sub are directed relations that link canonical tags (Step 1) and indicate a directed lexical entailment between them. They have a typical depth of 2–4.

3. Identify candidate source nodes of the trigger graphs: Meta-sources (Step 2) that are also direct successors of the *No Fandom* node in the parent–child graph. All terminal nodes in the parent–child graph are canonical tags and all predecessors are either a fandom, a media type, or *No Fandom*. The latter is added as a parent to tags that apply independently of the fandom, which includes trigger warnings but also, for example, holidays and languages. This yields about 5,000 candidate sources.

4. Identify sources of the trigger graphs: Manually annotate all candidate sources (Step 3) and discard nodes that do not map to a warning.

5. Segment the tag graph into trigger graphs: Starting from each trigger graph source (Step 4) and traverse the tag graph depth-first along the meta–sub relation. If a successor does not match the trigger warning assigned to its predecessor, the connecting edge is removed and the successor becomes a new trigger graph source, and annotated with a new trigger.

By following this linking strategy, it is possible to map the 41 million nodes in the tag graph to the respective trigger warnings with a very limited number of manual annotations.

5.2.4 Evaluation of the Supervision Method

First, we evaluate how effectively our distant-supervision approach annotates the freeform tags by comparing the inferred annotations with the two manually annotated tag sets `0-2k` and `10-11k` across the four different trigger warning sets. As shown in Table 5.7 (left), our approach scores well above 0.9 in accuracy and weighted average F_1. There is little difference between evaluating the fine-grained labels and their coarse equivalent.

Second, we evaluate how complete the set of all freeform tags can be annotated by our method. As shown in Table 5.4, due to the long-tailed distribution of the freeform tags, 52% of all occurrences can be manually annotated with high reliability and another 25% with an accuracy of about 0.95. This method maps all tags assigned to a work for more than half of all works in the corpus. The other half of the works are only partially annotated because tags are only wrangled, i.e. added to the tag graph, if they occur thrice. Accordingly, the method only annotates about 20% of the unique tags and misses all freeform tags with only a single occurrence.

Third, we evaluate how many freeform tags that contain a verbatim 'warning' are annotated with a warning from the taxonomy. Table 5.7 (right) shows that about 80% of verbatim warnings (that are part of the tag graph and can hence be annotated by the method) are also annotated with a taxonomy category. The other 20% are almost exclusively warnings that do not match any category, such as *Politics*, *Fluff*, *Police*, ... This ratio is lower for rare freeform tags which are not wrangled and thus not part of the tag graph. A verbatim tag contains one of the tokens 'tw(:)', 'cw(:)', or 'trigger(s)'.

5.3 Trigger Detection as Classification

To assess if trigger detection can be effectively solved via classification, we conduct three experiments that vary the depth of model analysis and the breadth of models and warnings covered. The first experiment investigates only a single warning, *Violence*, to analyze how the properties of the text influence the classification. The second experiment investigates multi-label classification across all warnings to analyze how the properties of the dataset and the warning taxonomy influence the classification. The third experiment also investigates multi-label classification, but analyzes differences between models and features.

Sample	Violent					Not Violent				
	Works	Words	Kudos	Hits	FF	Works	Words	Kudos	Hits	FF
Corpus	571,525	5,732	40	782	8	4,4 M	1,847	52	758	5
Random	10,000	6,773	51	1,088	8	10,000	1,869	74	1,074	5
Popularity	10,000	16,810	238	4,706	11	10,000	2,859	224	3,155	6
Rigor	10,000	7,161	60	1,255	9	10,000	2,127	84	1,235	6

Table 5.8: Descriptive statistics of corpus and sample datasets. Shown are number of works and median numbers of words, kudos, hits, and freeform tags (FF). The median is reported due to the long-tailed nature of the measures; the mean is about 2-4 times higher.

5.3.1 Violence Classification

To asses the influence of text properties, this experiment trains classification models to find documents that contain violence, as indicated by the *Archive Warning: Graphic Depictions of Violence*. We sample a balanced evaluation dataset from the trigger warning corpus (see Section 5.2), train an SVM, a BERT, and a Longformer, and investigate the influence of text length, popularity, rigor of the author, and the most discriminative features.

Evaluation Dataset

Because AO3 works do not include any annotations below document level—that is, we do not know the extent of violent content nor where in the text it can be found—our goal was to build a corpus with high-confidence examples of texts with and without violence. We apply three sampling strategies with varying reliability criteria: random sampling to represent the corpus, popularity-based sampling to exclude low-effort works, and rigor-based sampling to exclude works that are not thoroughly tagged by the author, i.e. work where the *Archive Warning* is possibly incorrect. Table 5.8 gives an overview of the three sampled datasets.

All sampling strategies randomly select 10,000 violent works (tagged with *Graphic Depictions of Violence*) and 10,000 non-violent works (tagged with *No Archive Warnings Apply* but not with *Graphic Depictions of Violence*). Before selecting the examples, we discarded all works with less than 100 words and works written in a non-English language. The random sample then draws the examples uniformly at random. The popularity-based sample first discards all works with less than 1,000 hits and less than 100 kudos and then draws uniformly at random. The rigor-based sample discards

Sample	SVM				BERT				Longformer			
	F_1	Prec	Rec	Acc	F_1	Prec	Rec	Acc	F_1	Prec	Rec	Acc
Random	0.86	0.86	0.87	0.86	0.79	0.75	0.83	0.78	0.86	0.84	0.88	0.86
Popularity	**0.89**	**0.88**	**0.91**	**0.89**	0.80	0.81	0.78	0.80	0.86	0.82	**0.91**	0.86
Rigor	0.86	0.88	0.85	0.87	0.79	0.70	0.90	0.76	0.85	0.83	0.87	0.84

Table 5.9: Experiment 1. Violence classification effectiveness on the test set for all three dataset samples; reported are micro and macro-averaged F_1 score, precision, recall, and accuracy.

all works with less than 10 *Additional Tags* (including characters and relationships) and then draws uniformly at random.

Table 5.8 shows the meta-data of the entire corpus and the three samples, extended by Table 5.10. The random and rigor-based samples are highly similar to the overall corpus; the popularity-based sample diverts by having longer (esp. violent) documents with more freeform tags.

Experiment Setting

We evaluate the four labeled datasets in a text classification setting by building classification models to assign trigger warnings at the document level.

Models We use three long-document classification baselines for our experiments: SVM, BERT, and Longformer. First, we use support vector machines (SVM) [91] since they are often used for text classification, are easily interpretable, and are not limited by the input sequence length. Second, we use a BERT transformer [56] as the go-to classification baseline; we used the pretrained `bert-base-uncased` checkpoint with 12 layers and 110M parameters, fine-tuned on our classification task. Third, we use a sparse-attention Longformer [17] as the state-of-the-art in many long document classification tasks [156]. We used the `allenai/longformer-base-4096` pretrained checkpoint, fine-tuned on our classification task.

Text Preprocessing For the SVM, we remove HTML tags, URLs, emojis, numbers, punctuation, and special characters and apply the Porter Stemmer [169]. For BERT and Longformer, we only remove HTML tags, URLs, numbers, and special characters, while punctuation is retained. For both neural models, the inputs are truncated at (and padded to) the maximum sequence length.

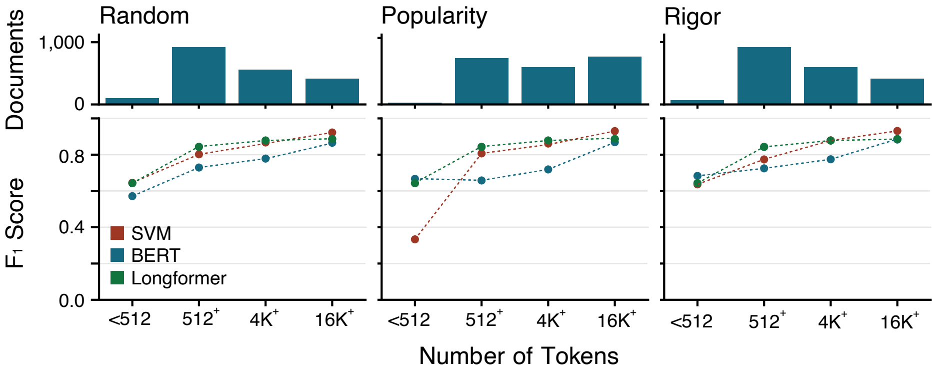


Figure 5.3: Classification effectiveness in terms of F_1 on the sample datasets over intervals of number of tokens.

Classification Setup The preprocessed data are split into 90:10 training and test sets via stratified sampling to maintain the class distribution.

As features for the SMV we use binary, uni- and bigram bag-of-word document vectors obtained from the lowercased preprocessed text; we keep only each dataset's 100,000 most frequent features. Maximum sequence lengths of 512 tokens for BERT and 4,096 tokens for Longformer are used.

Results

For each sample and model, we train a model on the training set and evaluate on the test set, the results of which are reported in Table 5.9. It can be seen that the SVM reaches overall best scores except for recall. Across the three sample datasets, the models achieve best F_1 on the popularity-based sample, followed by the random and the rigor-based sample. Recall is higher than precision for most neural models and vice versa for the SVM.

Figure 5.3 shows the effectiveness of the models on subsets of documents of varying lengths over input length. If the documents are shorter than the model's maximum input length, the SVM almost always performs worse (in terms of F_1) than the neural models and vice versa.

Meta-data (Tag) Differences Between Classes Table 5.10 shows the effect of topic on classification effectiveness. We list the relative count difference between all works D_i with an *Additional Tag* i (rating, freeform, characters) between violent v and non-violent nv documents defined as:

	Random	Pop.	Rigor
Rating			
Δ_{Mature}	0.551	0.492	0.537
$\Delta_{\text{Not Rated}}$	0.140	0.211	0.167
Δ_{Explicit}	0.275	0.206	0.231
Character Tags			
$\|D_i\|$	27,320	22,036	28,974
$\Delta_i > \ \ 0.75$	193	346	199
$\Delta_i > \ \ 0.25$	946	1,154	993
$\Delta_i < -0.25$	173	205	184
$\Delta_i < -0.75$	26	28	22
Most violent	*Original Characters*		(430)
	Original Female C.		(298)
	Original Male C.		(237)
	Harry Potter		(126)
Least violent	*Katsuki Yuuri*		(-54)
	Victor Nikiforov		(-56)
	Sherlock Holmes		(-143)
	Victor Nikiforov		(-148)

	Random	Pop.	Rigor
Rating			
$\Delta_{\text{Teen+}}$	-0.047	-0.141	-0.058
$\Delta_{\text{All Audiences}}$	-0.790	-0.840	-0.826
Freeform Tags			
$\|D_i\|$	64,961	80,364	71,767
$\Delta_i > \ \ 0.75$	333	504	357
$\Delta_i > \ \ 0.25$	922	1268	961
$\Delta_i < -0.25$	252	299	345
$\Delta_i < -0.75$	30	27	41
Most violent	*Angst*		(976)
	Violence		(967)
	Torture		(554)
	Drama		(534)
Least violent	*Fluff*		(-1174)
	Estab. Relationship		(-365)
	Drabble		(-184)
	Humor		(-155)

Table 5.10: Differences in the Meta-data frequency between violent and non-violent documents. Shown are the Δ_i as well as the absolute distance for the example tags split by ratings, characters (as indicator of fandom and plot), and freeform tags as content descriptors.

$$\Delta_i = \frac{|D_i^{\text{v}}| - |D_i^{\text{nv}}|}{|D_i^{\text{v}} \cup D_i^{\text{nv}}|}.$$

A $\Delta_i = 1$ indicates that all occurrences of the tag were assigned to violent documents and $\Delta_i = -1$ indicates the opposite.

Discussion

The final result (the SVM beats both neural models) is unexpected and can be (partially) explained by the influence of document length and topic.

Document Length Although the SVM has no contextual semantic information, it covers the tokens of the whole document through the bag-of-words representation, while BERT and Longformer are limited to a fixed input sequence (512/4,096 tokens respectively), which is only a fraction of the documents (see Table 5.8). Our analysis of the relation between text length and effectiveness (see Figure 5.3) reveals that neural models perform better than the SVM on documents shorter than their input limit; on longer documents, the violence might not have been part of the truncated input.

Topic Another possible explanation for the SVM's effectiveness is that the classes are separable by topic words (characters, fandom concepts) due to co-occurrence with (non-)violent documents; hence the classifier could not learn the more complex concept of violence. Our analysis shows that some fandoms are more violent than others (between 5–30% of works) and that about 5% of tagged characters and 2% of freeform tags are strongly associated with violent documents (strongly non-violent ones are rare). Conversely, the top SVM features (see Table 5.11) contain hardly topic words but mostly words clearly associated with violence. We hypothesize that topic impacts our violence classifier, but the evidence is not conclusive, warranting deeper analysis.

Class Distribution The classification model are effective with F_1 scores ranging from 0.837 to 0.939. While these results are promising, the task is far from solved. Due to the skewed class distribution in the fan fiction corpus (ca. 13% of works are violent; likely more extreme for other genres), a low false positive rate is crucial for a model to be transferable to real-world applications. Otherwise, a high recall would be more relevant as not to miss a warning label.

5.3.2 Analyzing Multi-label Trigger Detection

The second experiment investigates trigger detection as multi-label classification across all warnings to analyze how the properties of the dataset and the warning taxonomy influence the classification. To asses the influence of dataset properties and the warning taxonomy, this experiment trains four models for multi-label classification across all 36 warnings in the taxonomy. We sample several evaluation datasets from the trigger warning corpus (see Section 5.2) to investigate the influence of label set granularity (coarse vs. fine), open-endedness of the label set, document length, and support for the label from the freeform tags.

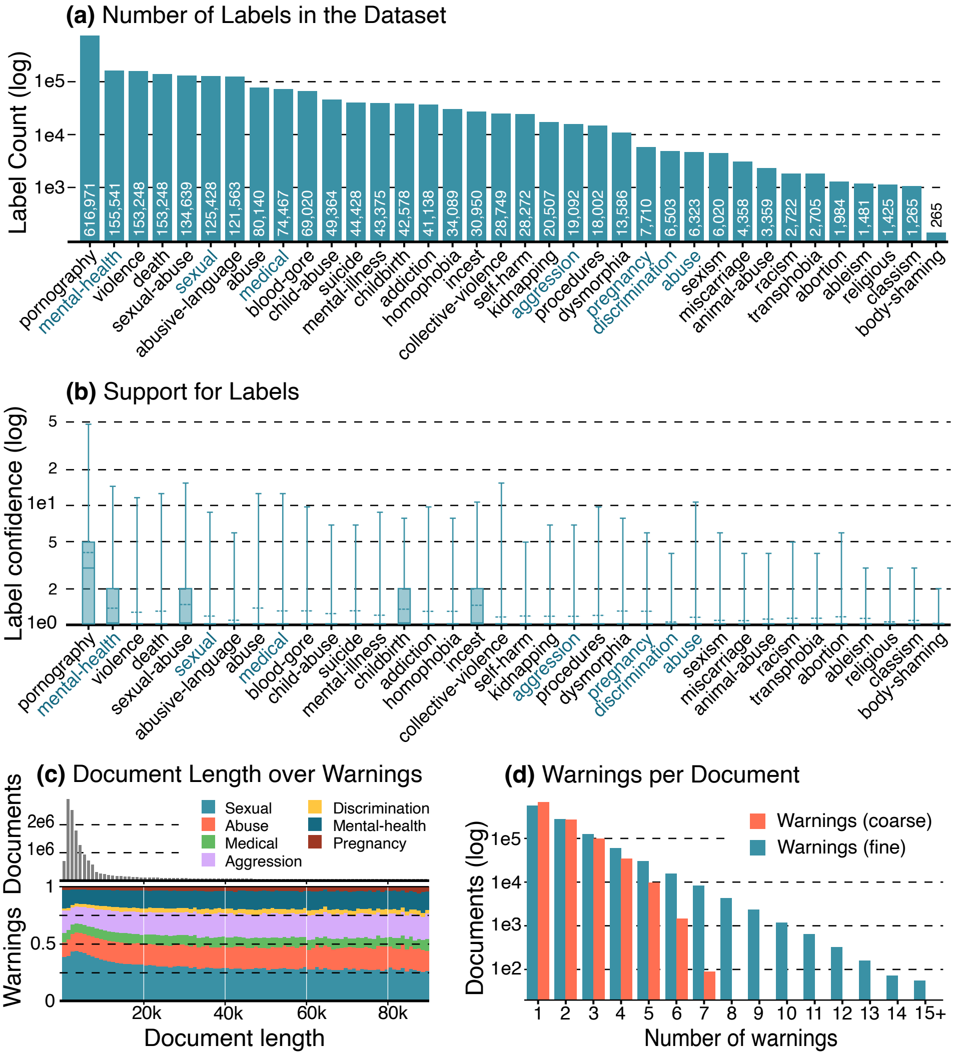


Figure 5.4: (**a**) Distribution of the fine-grained warnings over works in the dataset. Open-set warnings are highlighted. (**b**) Distribution of the label confidence for each (fine-grained) label. The label confidence for a warning of one work is the number of freeform tags assigned to that work, which are annotated with the respective warning. Dashed lines indicate the mean. (**c**) Distribution of document length in the dataset (top, log-scale) and distribution of all coarse-grained warnings split by text length (bottom). (**d**) Distribution of the number of documents that have a certain number of fine- and coarse-grained warning labels assigned. Document count is log-scaled.

Random		Popularity		Rigor	
Features indicating violence					
4.65	blood	3.82	blood	4.54	blood
2.40	dead	2.32	screams	2.62	dead
2.37	kill	2.02	scream	2.23	screams
2.33	screams	1.94	dead	2.13	pain
1.99	screamed	1.91	kill	2.03	bloody
1.95	flesh	1.89	pain	1.96	scream
1.89	screaming	1.89	killed	1.93	bleeding
1.86	scream	1.84	bloody	1.93	blade
1.79	pain	1.81	bleeding	1.91	kill
1.77	killed	1.75	blade	1.87	killed
	⋮		⋮		⋮
0.91	hannibal (84)	0.55	sith (341)	0.97	hannibal (67)
Features indicating non-violence					
-1.67	kiss	-1.16	kiss	-1.86	kiss
-1.07	managed	-0.96	embarrassing	-1.00	teasing
-1.01	ridiculous	-0.91	halfway	-0.93	spent
-0.92	admit	-0.90	experience	-0.92	demanded
-0.91	teasing	-0.90	surprised	-0.90	hadn
-0.91	shoulders	-0.87	close	-0.89	fin
-0.89	snorted	-0.82	dance	-0.89	flushed
-0.89	curled	-0.81	teasing	-0.87	imagined
-0.88	weekend	-0.80	ridiculous	-0.85	ridiculou
-0.88	surprised	-0.80	kissing	-0.84	carefully

Table 5.11: Most discriminative SVM features for both classes and all three sample datasets. The upper row group also lists the first topic (fandom-specific) feature, it's score, and position in the list (rank). It should be noted that there are almost no topic features in the top 1000 features which we inspected manually.

Evaluation Dataset

As a basis for the computational study of trigger warning assignment and our evaluation, we sampled a densely-annotated (excluding works without labels) dataset with 1,092,322 works from the previously constructed corpus. The sampling has two step: First, filtering out works from the corpus that do not match reliability criteria. Second, creating stratified standard splits that preserve label balance. Table 5.12 and Figure 5.4 show the descriptive statistics.

Dataset Properties	
Mean no. words	8K
Median no. words	3K
90pct no. words	21K
Mean no. chapters	3.0
Median no. chapters	1
Fine warnings	2.1M
Coarse warnings	1.7M

Dataset Properties	
Works with fewer than 512 words	56K
Works with fewer than 4,096 words	645K
Works with only closed warnings	728K
Works with only open warnings	94K
Works with open and closed warning	271K
Total Works	1.1M

Table 5.12: Properties of the dataset sampled for the multi-label trigger detection analysis. This sample only contains works with at least one trigger warning.

Sampling Method Works are filtered out according to the following criteria (see Table 5.3):

- Works without trigger warnings, which removes about 4.7M works.
- Works not written in English. Note that sampling a multi-lingual dataset is feasible in a few-shot scenario.
- Works published before AO3's release in 2009. Works with an earlier data were migrated from other archives and pre-dated to reflect their original data of publication. They were excluded because their tagging is not reliable.
- Works with atypical properties, which includes works with more than 100 chapters, more than 93,000 words (the top percentile), less than 50 words (which are usually placeholders for links or non-text media), and more than 66 tags (the top percentile).
- Works with less than 3 tags. It is very uncommon for works with 1 or 2 tags to indicate trigger warnings. Those tags more often indicate tropes or meta-information, so these works were removed to reduce label noise.
- Works with less than 5 kudos, i.e. likes, and less than 100 hits, i.e. reads, which are usually low-quality writing.
- works where less than 90% of the tags could be mapped, i.e. works with many unique, non-reviewed tags. This criterion filters works whose tags could not be thoroughly annotated with the weak supervision method, i.e. it is not clear if the tags indicate a trigger warning

> and this risks false negatives. However, we allow 10% of the tags to be non-annotated because this doubles the number of works with rare warnings while adding only about 70,000 works overall.

The remaining works were split 90:5:5 into training, validation, and test dataset. The balance of warning labels was preserved by iterating works with certain warnings from the least to the most common, adding a random work into either test or validation until they contained the targeted number of works with that label and then adding the remaining works into the training set.

Properties of the Experimental Dataset

We analyze five properties of the dataset to characterize trigger warnings in fan fiction and as foundation for the evaluation.

Warning Label Distribution Figure 5.4(a) shows that warnings follow a long-tailed distribution, which is common in multi-label settings: *Pornography* warnings are extremely common since sexual exploration is a relevant part of fan fiction. The open-set *Mental-health* warning is also common since it collects topics of strong anxiety and depression. Conversely, *Discrimination* warnings are rare. The number of works with rare labels is sufficient to train standard classification models.

Document Length Table 5.12 and Figure 5.4(c) show most works to be short (median about 3,000 words) and that longer works are often split into short chapters (90th percentile chapter length about 5,000 words). This exceeds BERT's input length (512 tokens) but comes close to that of a small Longformer (4,096 tokens). The label distribution is largely robust across document length, except for short documents having more *Sexual* content.

Warnings per Work Figure 5.4(d) shows an exponential decay of documents over number of warnings. A single warning is assigned to about half the works, while more than 10,000 have five or more labels, even in the coarse-grained 7-label setting.

Support per Warning Figure 5.4(b) show that most warnings have a median support of one freeform tag (mean 1.2–1.5). Most labels have rarely more than one, except for *Incest, Childbirth, Sexual-abuse*, and *Mental-health*. Again, *Pornography* is an outlier with a median of 3 and mean of 4 supporting tags. Authors tag sexual practices, kinks, and toys in great detail.

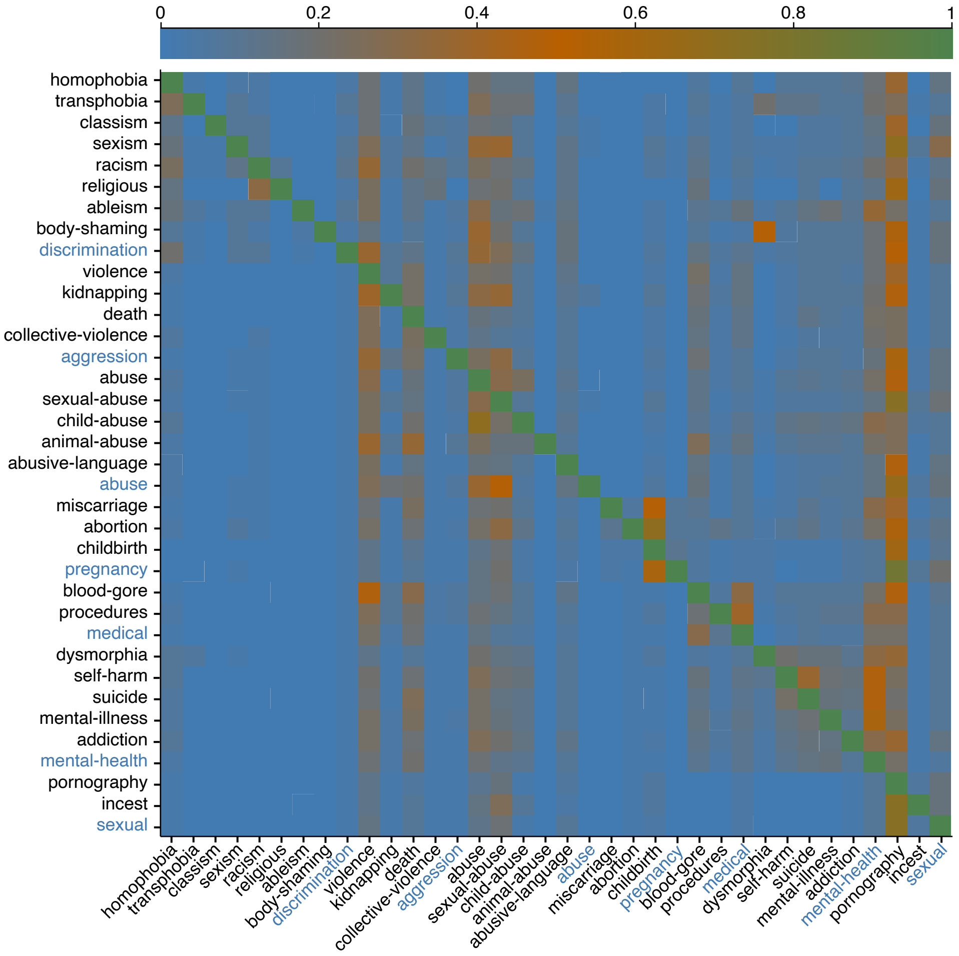


Figure 5.5: Co-occurrences between labels. Fields show which fraction of the row label also occurs with the column label. Labels are ordered by label group (as in the taxonomy visualization).

Co-occurrences between Warnings Figure 5.5 shows that warning co-occurrences are common with frequent tags, so that most labels co-occur with *Pornography* 20–40% of the time and 10-30% with *Violence*, *Mental-health*, *Abuse*, and *Death*. Furthermore, labels from the same group tend to weakly (about 10%) co-occur more with each other (especially in *Medical* and *Pregnancy*). Besides, some labels co-occur more frequently: *Pregnancy*, *Sexual-abuse*, and *Sexism* co-occurs with *Pornography* about 60% of the time.

Model	Labels	Sample	Features	Parameters
SVM	fine	10k	1–3-grams, χ^2	$C = 2$
	coarse	10k	1–3-grams, χ^2	$C = 2$
XGB	fine	10k	1-grams	max_depth $= 4$, lr $= 0.25$
	coarse	10k	1–3-grams, χ^2	max_depth $= 4$, lr $= 0.25$
BERT	fine	69k	–	epochs $= 10$, lr $= 2e - 5$
	coarse	69k	–	epochs $= 5$, lr $= 2e - 5$
LF	fine	10k	–	epochs $= 2$, lr $= 2e - 5$
	coarse	69k	–	epochs $= 3$, lr $= 2e - 5$

Table 5.13: Optimal parameters for SVM, XGBoost (XGB), RoBERTa, (BERT) and Longformer (LF) according to macro-averaged F_1 on the validation split.

Religious co-occurs with *Racism* about 30% of the time, as does each, *Body-shaming* and *Transphobia*, with *Dysmorphia* since the latter includes eating disorders and (gender) dysphoria.

Experiment Setting

To study the impact of label granularity, open-endedness, document length, and support on trigger detection, we evaluated the effectiveness of four models on the evaluation dataset described above. Each model was tested with the optimal parameters found by a parameter search. We optimized (1) the training data by undersampling the dataset to three different size thresholds, (2) the features for SVM and XGB by testing four different feature sets, and (3) all common model parameters. The optimal configurations are shown in Figure 5.13.

All models were trained once with 36 target labels (fine-grained) and once with 7 target labels (coarse-grained), where both variants were ablated individually. All ablation was done via grid search. The best configuration was selected by macro F_1 on the validation dataset. Model training was done on a single A100 GPU.

Models Four models were selected based on their use in recent comparative studies on long-document classification [50, 73, 156]: a Support Vector Machine SVM, XGBoost [42] (XGB), RoBERTa [125] (BERT), and Longformer [17] (LF). Each model was trained once for predicting the 36-label fine-grained warning set and once for the 7-label coarse-grained label set with identical input documents.

The SVM is a well-established traditional baseline in text classification [91] which is computationally cheap and serves as a good point of reference. The SVM is a linear SVM in one-vs-rest mode from scikit-learn [160] with TF-IDF document vectors of the word 1–3-grams with a minimum document frequency of 5 as features, tokenized by the `bertbase-uncased` tokenizer from Hugging Face. XGBoost, as opposed to the linear SVM, expresses non-linear partitioning of the feature space. The XGB model is a histogram-optimized tree construction from the XGBoost library [42] with the same features as the SVM.

Engineered feature spaces are (still) competitive in long-document classification since positional information is less significant than the input size limitation of transformer models. The experiments of Dai et al. [50] and Park et al. [156] suggest that RoBERTa and Longformer with truncation are comparative to the state-of-the-art and as efficient as models that take longer contexts into account. The BERT model is a `roberta-base` checkpoint from the Hugging Face with input padding and truncation to 512 tokens. The LF model is an `allenai/longformer-base-4096` checkpoint from Hugging Face with input padding and truncation to 4,096 tokens. For all models, the text was lower-cased and HTML formatting as well as non-alphanumeric symbols except `.,!?"'` removed.

Undersampling Since the training dataset is very large and skewed towards a few very common labels, the training dataset was undersampled in three variations: to the 25% quartile (10,000 works/label), the 50% quartile (28,000 works/label), and the 75% quartile (69,000 works/label).

The sampling strategy started with the rarest label and randomly added works with this label until, either, the threshold was reached, or, all documents with that label were added. Previously added documents (with multiple labels) counted towards the threshold. We ignored the occasional over-drawing of labels (when a high-frequency label was already sampled over the threshold by sampling the lower-frequency labels alone) since this behavior is difficult to avoid for multi-label datasets and did rarely occur.

SVM and XGBoost Features All feature sets used tf-idf vectors of token n-grams (using the `bert-base-uncased` tokenizer) with a minimum document frequency of 5. We ablated the four feature sets: (1) token 1-grams, (2) token 3-grams, (3) token 1–3-grams and χ^2-feature selection, and (4) token 1–5-grams and χ^2-feature selection. For SVM, we selected the best 50,000 (**a**) Fine (36 labels)

	Macro-avg.			Micro-avg.		
	Prec	Rec	F_1	Prec	Rec	F_1
SVM	**0.47**	0.18	0.25	**0.75**	0.37	0.49
XGB	0.44	**0.25**	**0.30**	0.72	**0.40**	**0.52**
BERT	0.36	0.19	0.23	0.56	0.37	0.45
LF	0.26	0.23	0.21	0.45	0.30	0.36

(**b**) Coarse (7 labels)

	Macro-avg.			Micro-avg.		
	Prec	Rec	F_1	Prec	Rec	F_1
SVM	0.59	**0.54**	**0.56**	0.71	**0.61**	**0.66**
XGB	**0.65**	0.51	**0.56**	**0.77**	0.58	**0.66**
BERT	0.45	0.52	0.46	0.53	0.54	0.53
LF	0.44	0.48	0.43	0.50	0.47	0.49

Table 5.14: Test Data Results of SVM, XGBoost, RoBERTa, and Longformer.

features. For XGB, we selected the 20,000 best features. Preprocessing and tokenization were identical for all approaches.

Model Parameters All models were ablated on all three input data samples, except for Longformer with fine-grained labels and XGB which were not trained on the 69,000 works sample due to resource limitations. For SVM, we ablated the regularization parameter $C \in \{0.1, 0.2, 0.5, 1.0, 2.0\}$. For XGB, we ablated the tree depth max_depth $\in \{2, 3, 4\}$ and the learning rate $\in \{0.25, 0.5, 0.75\}$ with 100 estimators and early stopping at 10 rounds. For BERT, we ablated the number of epochs $\in \{3, 5, 10\}$ and the learning rate $\in \{1e-4, 5e-5, 2e-5, 1e-5\}$ with a batch size of 32. For LF, we ablated the number of epochs $\in \{2, 3, 5\}$ and the learning rate $\in \{1e-4, 5e-5, 2e-5, 1e-5\}$ with a batch size of 4.

Results

Table 5.14 shows the (micro- and macro-averaged) effectiveness of the four models when trained once for a 36-label and once for a 7-label setting. The best model has a micro-F_1 of 0.52 on the fine-grained dataset, lower than the scores on comparable datasets reported on Papers with Code: 0.91 [89] on Reuters-21578 and 72.8 [122] on AAPD.

The overall most effective model is XGB with 0.3 macro- and 0.52 micro-F_1 on the fine-grained label set, followed by SVM and BERT. Precision is generally higher than recall by about 0.2–0.3. Micro-averaged scores are higher than macro-averaged scores by about 0.2 (fine-grained), which is not uncommon for strong label imbalance. The label-wise analysis (see Table 5.15) shows that the models are most effective on the very common warnings (about 0.88 on *Pornography*) and least effective on the rare warn-

Warning	SVM	XGB	BERT	LF
coarse-grained (7 labels)				
sexual-content	0.87	**0.88**	0.71	0.63
aggression	**0.55**	0.53	0.53	0.52
abuse	**0.47**	0.42	0.40	0.36
mental-health	**0.58**	0.52	0.52	0.47
medical	**0.53**	0.52	0.40	0.39
pregnancy	0.61	**0.67**	0.43	0.42
discrimination	0.33	**0.37**	0.24	0.21
fine-grained (36 labels)				
pornography	0.86	**0.88**	0.76	0.66
violence	0.30	**0.33**	0.27	0.23
mental-health	0.34	**0.35**	0.29	0.33
death	0.24	0.26	**0.27**	0.25
sexual	0.09	0.12	**0.25**	0.07
sexual-abuse	0.33	**0.39**	0.34	0.25
abuse	0.23	**0.26**	0.24	0.23
medical	0.32	0.37	**0.41**	0.33
blood-gore	0.28	**0.34**	0.32	0.25
abusive-language	0.09	0.11	**0.21**	0.12
suicide	0.26	0.32	**0.34**	0.27
child-abuse	0.22	0.25	**0.31**	0.27
childbirth	0.55	**0.63**	0.47	0.44
mental-illness	0.11	**0.16**	**0.16**	0.15
fine-grained cont.				
addiction	0.22	**0.33**	0.26	0.27
incest	0.52	**0.53**	0.50	0.37
homophobia	0.31	**0.39**	0.27	0.21
self-harm	0.37	**0.41**	0.33	0.29
kidnapping	0.26	**0.36**	0.25	0.23
aggression	0.33	**0.38**	0.31	0.26
collective-violence	0.35	**0.36**	0.32	0.20
procedures	0.26	**0.30**	0.17	0.17
dysmorphia	0.41	**0.44**	0.34	0.23
pregnancy	0.37	**0.44**	0.21	0.23
abuse	0.20	**0.21**	0.11	0.08
sexism	**0.14**	**0.14**	0.01	0.05
discrimination	0.06	0.06	0.00	0.05
racism	0.10	**0.17**	0.06	0.12
miscarriage	0.18	**0.35**	0.18	0.16
animal-abuse	0.08	**0.17**	0.11	0.14
transphobia	0.14	**0.34**	0.17	0.20
abortion	0.17	**0.32**	0.02	0.18
ableism	0.00	0.06	0.00	**0.07**
religious-disc.	0.10	**0.12**	0.04	0.09
classism	**0.10**	0.05	0.00	0.04
body-shaming	0.00	0.00	0.00	0.00

Table 5.15: Classification effectiveness of SVM, XGBoost, RoBERTa, and Longformer on the test dataset. Shown are the micro F_1 scores for each label.

ings (0.0–0.2). These rare warnings are often *Discrimination*. XGB is often more effective for rare labels than the others (about +0.25 on *Abortion* and *Transphobia*). BERT is more effective on seven of the more frequent labels but is about 0.1 less effective on the most frequent labels (*Pornography, Violence, Mental-health*), resulting in reduced total effectiveness. LF failed to generalize to the test data and is weaker than BERT; on the validation data, LF outperforms BERT by about 0.1.

Granularity Table 5.14 shows the difference between predicting coarse (7) and fine-granular (36) labels. The models are consistently more effective on the coarse-grained label set: recall is higher by about 0.2–0.3 and precision by up to 0.2. The macro-average effectiveness improves more than the micro-averaged one since coarse labels are more frequent and the rare *Discrimination* labels are combined, which reduces their impact on the average. Consequently, the difference between the macro- and micro-average is also lower (from about 0.2 to 0.1). The difference between precision and (**a**) Macro F_1

Set	Total	Length				Open-endedness		Confidence
		512	4k	16k	16k+	Open	Closed	
fine-grained								
SVM	0.25	0.21	0.27	0.24	0.18	0.24	0.25	0.28
XGB	**0.30**	0.19	**0.29**	**0.31**	**0.30**	**0.28**	**0.30**	**0.30**
BERT	0.23	**0.30**	0.27	0.20	0.15	0.23	0.23	0.23
LF	0.21	0.29	0.24	0.17	0.14	0.19	0.21	0.15
coarse-grained								
SVM	**0.56**	0.51	**0.57**	0.56	0.53	–	–	0.57
XGB	**0.56**	0.37	0.54	**0.57**	**0.59**	–	–	**0.60**
BERT	0.46	**0.52**	0.48	0.43	0.43	–	–	0.40
LF	0.43	**0.52**	0.45	0.39	0.39	–	–	0.37

(**b**) Micro F_1

Set	Total	Length				Open-endedness		Confid.	Section		
		512	4k	16k	16k+	Open	Closed		Top	Mid	Bot
fine-grained											
SVM	0.49	0.40	**0.53**	0.49	0.39	0.26	0.54	**0.82**	**0.43**	**0.55**	0.50
XGB	**0.52**	0.36	**0.53**	**0.53**	**0.49**	**0.29**	**0.56**	**0.82**	0.42	0.52	**0.55**
BERT	0.45	**0.53**	0.49	0.41	0.34	**0.29**	0.48	0.61	0.37	0.54	0.45
LF	0.36	0.46	0.41	0.31	0.26	0.24	0.38	0.52	0.26	0.48	0.46
coarse-grained											
SVM	**0.66**	**0.59**	**0.68**	0.66	0.59	–	–	0.78	–	–	–
XGB	**0.66**	0.48	0.66	**0.67**	**0.66**	–	–	**0.81**	–	–	–
BERT	0.53	**0.59**	0.57	0.50	0.48	–	–	0.57	–	–	–
LF	0.49	0.58	0.52	0.44	0.43	–	–	0.5	–	–	–

Table 5.16: Test effectiveness of SVM, XGBoost, RoBERTa, and Longformer, split by various characteristics. **Total** indicates the overall F_1 scores. **Length** indicates the scores on documents in the length (of tokens) intervals 50—512, 512–4,096, 4,096–16,000, and 16,000–93,000 (16k+). **Open-endedness** indicates the scores on open vs. closed classes. Label confidence (**Confid.**) indicates the scores on all works that have at least 2 free-form tags as support for each assigned warning. **Section** indicates the average scores of only the 12 most common tags (top 33%), and equivalently the middle and bottom third.

recall is also lower (from about 0.25 to 0.1) since recall improves more than precision. Micro-averaged precision is independent of granularity.

Open-endedness Table 5.16 shows the average effectiveness of the open and closed-set (fine-grained) warnings. The difference in macro-F_1 is negligible, however, the closed-set labels are more effective by 0.1–0.3 in micro-

F_1 since it is strongly affected by the high scores of *Pornography*. Table 5.15 shows no notable difference between open and closed-set labels.

Document Length Table 5.16 also shows assignment effectiveness depending on a work's length. The neural models are more effective for works that are shorter than their input length limit. BERT is the most effective model on works with less than 512 tokens by 0.1 macro and 0.2 micro-F_1 over XGB. However, BERT becomes less effective the longer the documents are (XGB is more effective by 0.15 for works with more than 16,000 tokens). Longformer behaves the same.

Support Table 5.16 also shows the effectiveness on works that have at least two freeform tags supporting each annotated warning label. The support has no impact on macro-F_1 but the micro-F_1 is higher for the set of works with a minimum support of 2, likely because *Pornography* is often supported by multiple freeform tags and strongly impacts the micro-average.

Discussion

There are five key observations from the results: First, there is no notable difference in effectiveness between labels with open and closed-set semantics, which speaks for the inclusion of open-set warnings in the future. Second, learning and predicting from the full text is essential and more important for trigger warnings than for other multi-label classification datasets. Models with short input length are less effective because they rely on truncated text and so only see the beginning of a work, while the relevant passages for the classification decision are often contained in later passages. Third, recall is worse than precision, which is a key issue. Trigger warning assignment is a high-recall task since false negatives correspond to warnings not given, which cause more harm than false positives, i.e. superfluous warnings that did not come to pass. Fourth, models are less effective for rare labels, which is common for multi-label classification problems. Fifth, models are more effective on coarse-grained labels. However, predicting fine-grained labels with high reliability can greatly reduce the number of documents that a reader may want or need to skip to be safe. Future work should focus on improving the fine-grained prediction performance.

5.3.3 Shared Task Evaluation

The third experiment, similarly to the second, investigates trigger detection as multi-label classification across all warnings, but with a focus on analyzing the differences between model architectures and features. As with the experimental validation for the second case study of this thesis in Section 4.3, we organized a shared task to find the most effective systems.[3,4] The "Trigger Detection" task at PAN 2023 asked the question:

> *Given a fan fiction document, determine all required trigger warnings from the given label set.*

Evaluation Dataset

The *PAN23-trigger-detection* evaluation dataset is also sampled from the corpus presented in Section 5.2 and contains 341,246 fan fiction works from Archive of our Own (AO3) annotated with 32 trigger warnings. No documents were sampled for the rarest labels from the corpus because decisions on individual documents would have a disproportional impact on the macro-average scores.

As for the second experiment, the evaluation dataset was sampled by applying several filtering criteria. The criteria are identical, but the individual thresholds are more strict to get a smaller and cleaner dataset that is easier to use with limited resources. We filtered out all works without assigned warnings, published pre-2009, with non-mapped (unique) freeform tags, non-English content, works outside of 50–6,000 words and 2–66 freeform tags, and works with less than 1,000 hits or 10 kudos.

The dataset was split via stratified sampling into 90:5:5 training, validation, and test sets; i.e., we kept the label distribution equal across the three splits. The training dataset with ca. 300,000 works is large enough to train deep neural classifiers. The datasets contain ca. 5% very short documents (<512 words) that can be used by a BERT-based system without truncation and ca. 85% medium-sized documents (<4,096 words) that can be used by a sparse-attention model. The most frequent label is *Pornography* and occurs in ca. 77% of the documents. Most labels are less common, between ca. 10% for *Sexual-assault* and 6e-4% for *Animal-cruelty*. Documents have 1–13 labels per document, ca. 71% with a single label, 20% with two, and 6% with three.

[3] Baseline and Evaluators: github.com/pan-webis-de/pan-code

[4] Evaluation Dataset: zenodo.org/record/7612628

Evaluation Measures

The submissions are evaluated using the established multi-label classification metrics: F_1 and Accuracy. In addition, we assess (1) the effectiveness of individual labels, (2) the effectiveness in relation to document metadata, and (3) the effectiveness of voting-based ensembles of the best submissions.

The primary metrics are precision, recall, and F_1 at both micro- and macro average, and subset accuracy, which measures accuracy on a per-sample basis (i.e., if all labels of one example are set correctly). The evaluation favors the macro over the micro F_1 scores due to the label imbalance, and the evaluation favors recall over precision, since trigger warning assignment is considered a high-recall task where false negatives cause more harm than false positives. Additionally, the average precision and recall is evaluated for the most frequent warning, *Pornography*, the 15 next-most common labels (*Sexual-assault–Dissection*), and the 16 least common labels. We also compute the number of classes with either zero or a very low (<0.1) precision and recall to check for high-frequency label bias.

As metadata-based metrics, we compare micro and macro F_1 for the document subsets that fall within certain metadata thresholds as follows:

- **Document length:** Short (less than 500), medium, or long (more than 4,000). We assume that short works are easier to classify since models can capitalize more directly on BERT (which has a short input size).
- **Tag count:** Few (less than 5), medium, or many (more than 20). We assume that works with many freeform tags are easier to classify because many tags suggest that authors took greater care with annotating their works and the resulting higher label quality leads to better effectiveness.
- **Rating:** *Explicit*, *Mature*, or neither. We assume that explicit or mature works contain more markers and are thus easier to classify.
- **Archive warnings:** Has an archive warning (*Graphic Depictions Of Violence, Major Character Death, Rape/Non-Con, Underage*), has no warning (*No Archive Warnings Apply*), or does not specify a warnings (*Choose Not To Use Archive Warnings*). We assume that works with a warning are easier to classify and works without specified warning are the hardest, since authors hide warning tags within spoilers and might therefore less diligently annotate freeform warnings.

Participant	Model	Features	Length	Imbalance
Sahin et al. [200]	RoBERTa/LSTM	CLS embedding	Hierarchical cls.	Weighted loss
Su et al. [211]	RoBERTa/CNN	Context embeddings with 1D convolution and mean-pooling	Hierarchical cls.	–
XGBoost baseline	XGBoost	TF·IDF	Document features	Undersampling
Cao et al. [35]	RoBERTa	CLS token	Voting	Re-sampling
Cao et al. [34]	RoBERTa	CLS token	Voting	Re-sampling + separate classifiers
Felser et al. [68]	MLP	Aggregate word emb. + topic model	Document features	Weighted loss
Shashirekha [205]	LSTM	GloVE	(LSTM)	–

Table 5.17: Overview of the submitted methods. Listed is the (dominant) model architecture, the feature representation, the method used to handle long documents, and the sampling strategy used to handle the skewed label distribution.

- **Popularity:** Low (less than 50 comments or 60 bookmarks or more than 450 kudos or 8,500 hits), medium, or high (less than 280 comments or 330 bookmarks or more than 1,850 kudos or 5,000 hits). We assume that popular works are also easier to classify because authors are more diligent when tagging works that gain much attention.

Finally, four ensembles were constructed from the submitted results, where the assignment of a true label is decided by voting to surpass a threshold τ. The *Top-3* ensemble uses the three best submissions with $\tau = 2$, the other ensembles use all submissions with $\tau = \{3, 5, 7\}$.

Baselines The baseline was a XGBoost [42] classifier adapted from the previous experiments, with word-1–3-gram features encoded as TF·IDF document vectors with a minimum document frequency of 5 and χ^2 selection of the top 10,000 features. The dataset was undersampled uniformly at random to 1,000 samples per label. As parameters, we used a `max depth` or 3, a `learning rate` of 0.25, and 300 estimators with 10-round early stopping. All features, parameters, and sampling thresholds were determined via grid search as previously described for experiment 2.

Submissions

The 6 submissions to the "PAN 2023 Trigger Detection" task employed a broad set of techniques, from hierarchical transformer structures to strategic feature engineering. Table 5.17 shows an overview of the different strategies used by the participants. All participants used a form of a neural network as a model, where RoBERTa was most common and most successful as a classifier or pre-trained model to produce a strong input encoding. Most submissions also focused on improving the long document aspect of the task (most documents are longer than the input size of the state-of-the-art classification models) by using hierarchical classifiers (chunks are encoded, and prediction is based on a combination of encodings), or voting-based approaches (chunks are labeled individually, document labels are aggregated over chunk labels). The submissions cope with the label imbalance (the most common label (*Pornography*) is an order of magnitude more common than the other labels) through over- and undersampling or by changing class-weights in the loss function, so that misclassifying a rare class increases the error more than a common label.

Sahin et al. [200] submitted a hierarchical transformer architecture that achieved the top macro F_1 score (by a slim margin of 0.002) and came in second in micro F_1 and accuracy, while having a relatively high recall within the top approaches. The approach first segments the document into chunks (200 words with 50 words overlap) and then pre-trains a RoBERTa transformer on the chunks to learn the genre. The architecture then embeds all chunks of a document using the pre-trained transformer, followed by an LSTM for each label (in a one-vs-all setting), predicting the class from a sequence of chunk-embeddings (RoBERTa's `[CLS]` token). To cope with label imbalance, the approach assigns positive weights in the loss function to the rare half of the labels.

Su et al. [211] submitted a siamese transformer that achieves the second-best macro F_1 score (by a slim margin of 0.002) and the top scores in micro F_1 and accuracy, while notably favoring precision over recall. The approach segments the documents into 505-word chunks, encodes the first and last chunk using RoBERTa, mean-pools the contextual embeddings (ignoring the `[CLS]` token), and classifies based on the pooled embeddings using a 1D convolutional neural network.

Cao et al. [35] submitted a voting-based transformer that favors recall over precision. The approach segments the training documents into chunks, assigns each chunk the labels from its source document, and trains a single RoBERTa-based classifier on each chunk. To make predictions, the documents are again chunked, the labels for each chunk are predicted, and a label is assigned to the document if it is assigned to more than half of the chunks. The training data was dynamically over- and undersampled:

Participant	**Macro**			**Micro**			**Acc**
	Prec	Rec	F_1	Prec	Rec	F_1	
Sahin et al. [200]	0.37	0.42	**0.352**	0.73	0.74	0.74	0.59
Su et al. [211]	**0.54**	0.30	0.350	0.80	0.71	**0.75**	**0.62**
XGBoost baseline	0.52	0.25	0.301	**0.88**	0.57	0.69	0.53
Cao H. et al. [35]	0.24	0.29	0.228	0.43	0.79	0.56	0.18
Cao G. et al. [34]	0.28	0.22	0.225	0.58	0.66	0.62	0.32
Felser et al. [68]	0.11	**0.63**	0.161	0.27	**0.82**	0.40	0.27
Shashirekha et al. [205]	0.10	0.04	0.048	0.82	0.50	0.63	0.52
Ensemble (Top 3)	**0.56**	0.30	0.36	0.88	0.68	**0.77**	**0.63**
Ensemble ($\tau = 3$)	0.38	0.42	**0.37**	0.65	0.80	0.72	0.52
Ensemble ($\tau = 5$)	0.55	0.20	0.26	0.88	0.65	0.75	0.60
Ensemble ($\tau = 7$)	0.39	0.07	0.10	**0.97**	0.50	0.66	0.53

Table 5.18: Participant scores at the shared task on trigger detection. Shown are the core metrics, sorted by macro F_1. Bold indicates the leading approach for each metric. Scores of the voting-based ensembles are bold when they are better than the leading submission.

pornography was undersampled to 5,000 examples and other labels to 2,000 examples. Examples with rare labels were replicated 8-10 times.

Cao et al. [34] also submitted a voting-based transformer that achieved very balanced results, neither favoring macro over micro scores nor precision over recall. The approach chunks and votes similarly to Cao et al. [35] but builds two different models to overcome the data imbalance, one for pornography and one for the other 31 classes. The pornography model was trained on a random selection of 40,000 works with and 40,000 works without the pornography warning. The second model removes works with only the pornography warning, undersamples frequent classes to 3,000 examples, and oversamples rare labels by replicating works 4-6 times.

Felser et al. [68] submitted a 1-vs-rest multi-layer perceptron based on two features: fasttext-based document embeddings and superclass probabilities. This approach achieved the top micro and macro recall, at the cost of precision on the test dataset. Document embeddings were created by training a fasttext model from the training data, generating the embeddings for each unique word in a document, scaling them by term frequency, and adding and normalizing the scaled word vectors over the document. The superclass probabilities were determined by grouping the 32 labels semantically into 6 superclasses, bootstrapping a seeded LDA with the 50 most relevant bi-grams of each group (determined through a TF·IDF-like approach for n-gram weighting, which downgrades pornographic terms), and training a classifier to predict the superclass based on the topic model outputs based on class probabilities. Label imbalance was addressed via class penalties in the loss function, with a higher penalty in the the MLP-2 variant.

Lastly, **Shashirekha et al.** [205] present an LSTM-based approach using GloVE-embeddings, which is third in micro F_1 with very high precision but rather weak in macro average scores.

Results

Table 5.18 shows the evaluation results for the primary metrics ordered by macro F_1. Here, the hierarchical classifiers are the most effective by a large margin, followed by the XGBoost baseline. The most effective approach by macro F_1 is the one by Sahin et al. with 0.352, a small margin before that of Su et al. with 0.350. The best approach by micro F_1 and subset accuracy is the one by Su et al.. The XGBoost baseline is only beaten by these two top approaches. The models score very differently in precision and recall, depending on the architecture. Four models score generally higher in recall, the other 4 in precision. There is no obvious relationship between effectiveness and preference for precision or recall. The ensembles (top 3 and $\tau = 3$) beat the submissions but by a very small margin of ca. 0.02.

Table 5.19 shows the evaluation results for the extended metrics. Unsurprisingly, all submissions score very high on *Pornography* and notably lower on all rare labels, which explains the difference between macro and micro F_1. There is a clear decrease in efficiency with decreasing label frequency. It also becomes more obvious that models tend to be good in either precision or recall with large differences between them. Combining the strength of the high-recall and high-precision approaches is a potential way forward, albeit our basic ensemble exploits that only marginally.

Table 5.20 shows the evaluation results based on document subsets with common metadata values. Regarding the document length, the macro F_1 scores are mixed: Models that use the complete work as single examples during training (Sahin et al. [200], the baseline, and Felser et al. [68]) are slightly (0.05–0.1) less effective on short texts; models that use only a section of the document (Su et al. [211], Cao, G. et al. [34]) are slightly (0.05–1.0) less effective on long texts. On micro F_1, all models tend to perform worse on shorter texts. This contradicts our assumption (and prior evidence [245]) that models will be generally better on short texts which can

Participant	Porn.		Mid		Bot		Zero P/R		<0.1 P/R	
	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec
Sahin et al. [200]	0.95	0.96	**0.62**	0.48	0.12	0.51	3	3	10	6
Su et al. [211]	0.90	0.97	0.61	0.43	**0.57**	0.19	6	6	6	9
XGBoost baseline	**0.98**	0.87	**0.62**	0.21	0.38	0.24	**2**	**2**	**2**	9
Cao H. et al. [35]	0.86	**0.98**	0.22	0.61	0.16	0.12	5	5	7	13
Cao G. et al. [34]	0.97	0.88	0.29	0.42	0.24	0.09	4	4	8	15
Felser et al. [68] (MLP1)	0.97	0.91	0.18	**0.72**	0.03	**0.64**	7	5	24	**5**
Felser et al. [68] (MLP2)	0.97	0.91	0.26	0.45	0.03	0.31	13	13	22	13
Shashirekha et al. [205]	0.93	0.91	0.18	0.04	0.00	0.00	23	24	25	30
Ensemble (Top 3)	0.96	0.96	**0.72**	0.38	0.50	0.25	5	5	5	7
Ensemble ($\tau = 3$)	0.94	0.97	0.43	0.61	0.28	0.36	4	4	4	6
Ensemble ($\tau = 5$)	0.97	0.93	0.68	0.33	0.76	0.10	9	9	9	16
Ensemble ($\tau = 7$)	**0.98**	0.87	0.82	0.08	**0.91**	0.02	18	20	18	25

Table 5.19: Participant scores at the shared task on trigger detection. Shown are the extended metrics: precision and recall for *Pornography*, the more common half of labels excluding pornography (**Mid**) and the rare half (**Bot**) as well as the number of classes where precision and recall is either **zero** or below **0.1**. Participants are sorted by total macro F_1 (cf. Table 5.18).

fully capitalize on BERTs strength on short inputs. An alternative hypothesis is that shorter documents are simply less clear and have fewer of the markers that the classifier expects to make a positive prediction.

Regarding the tag count, the top models are slightly (0–0.1) less effective when there are many freeform tags. There is no difference between the less effective models. This also contradicts our assumption that models with many tags are easier to classify due to higher label reliability.

Regarding popularity, there is no notable difference in micro F_1. On macro F_1, models are slightly (0.04–0.14) more efficient on high popularity works than on low popularity works. This agrees with our assumption that labels of popular works are more reliable.

Regarding the archive warnings, there is no notable difference between works with or without warnings. However, the most effective models are slightly (ca. 0.05 macro, ca. 0.15 micro) less effective on works with undeclared warnings than on others. This agrees with our assumption that these works are less diligently tagged by their authors (e.g. as a spoiler tag).

Lastly, regarding the rating, models are more (ca. 0.2–0.3 micro F_1) effective on explicit works, which is likely an artifact from the very effective classification of the *Pornography* label. On macro F_1, contrary to the micro (**a**) Macro F_1

Participant Team	**Length**			**Tag count**			**Popularity**			**AO3 Warning**			**Explicit**	
	Short	Med	Long	Few	Med	Many	Low	Med	High	w/	w/o	Unk	Yes	No
Sahin et al. [200]	0.28	0.35	**0.34**	0.36	**0.35**	**0.30**	**0.30**	**0.35**	**0.35**	**0.35**	**0.35**	**0.31**	**0.31**	0.37
Su et al. [211]	**0.39**	**0.36**	0.27	**0.37**	0.34	0.25	0.22	0.33	**0.35**	**0.35**	**0.35**	0.29	0.28	**0.38**
XGBoost baseline	0.24	0.30	0.29	0.31	0.30	0.22	0.16	0.28	0.30	0.30	0.30	0.25	0.28	0.30
Cao H. et al. [35]	0.23	0.23	0.22	0.21	0.23	0.23	0.19	0.25	0.22	0.24	0.19	0.24	0.21	0.23
Cao G. et al. [34]	0.22	0.23	0.18	0.23	0.22	0.19	0.20	0.25	0.22	0.23	0.21	0.20	0.18	0.25
Felser et al. [68] (1)	0.13	0.16	0.17	0.14	0.17	0.20	0.17	0.16	0.16	0.16	0.14	0.18	0.16	0.15
Felser et al. [68] (2)	0.12	0.15	0.16	0.14	0.16	0.17	0.15	0.15	0.15	0.15	0.14	0.16	0.15	0.10
Shashirekha et al. [205]	0.04	0.05	0.04	0.05	0.05	0.04	0.04	0.04	0.05	0.04	0.04	0.05	0.05	0.04
Ensemble (Top 3)	0.32	**0.37**	0.33	**0.41**	0.35	0.25	0.23	**0.36**	0.36	0.36	**0.35**	0.31	0.31	**0.39**
Ensemble ($\tau = 3$)	0.31	**0.37**	**0.36**	0.38	**0.37**	**0.33**	0.25	0.35	**0.37**	**0.37**	**0.35**	**0.35**	**0.34**	0.38
Ensemble ($\tau = 5$)	0.27	0.27	0.21	0.29	0.25	0.18	0.20	0.28	0.26	0.26	0.25	0.21	0.23	0.27
Ensemble ($\tau = 7$)	0.09	0.10	0.07	0.10	0.10	0.07	0.08	0.08	0.10	0.10	0.10	0.08	0.09	0.10

(**b**) Micro F_1

Participant Team	**Length**			**Tag count**			**Popularity**			**AO3 Warning**			**Explicit**	
	Short	Med	Long	Few	Med	Many	Low	Med	High	w/	w/o	Unk	Yes	No
Sahin et al. [200]	0.73	0.74	**0.72**	0.75	0.74	**0.66**	0.76	0.75	0.73	0.74	0.79	0.62	0.79	**0.59**
Su et al. [211]	**0.76**	**0.76**	**0.72**	**0.77**	**0.75**	**0.66**	**0.77**	**0.77**	**0.75**	**0.76**	**0.80**	**0.64**	**0.81**	**0.59**
XGBoost baseline	0.58	0.69	0.70	0.72	0.68	0.59	0.70	0.72	0.68	0.70	0.76	0.52	0.77	0.41
Cao H. et al. [35]	0.52	0.56	0.57	0.54	0.57	0.58	0.63	0.58	0.55	0.55	0.56	0.56	0.60	0.45
Cao G. et al. [34]	0.58	0.62	0.61	0.61	0.62	0.59	0.69	0.65	0.61	0.62	0.63	0.56	0.66	0.47
Felser et al. [68] (1)	0.31	0.40	0.42	0.38	0.41	0.44	0.43	0.43	0.40	0.39	0.42	0.38	0.50	0.25
Felser et al. [68] (2)	0.45	0.54	0.56	0.54	0.54	0.52	0.57	0.57	0.53	0.54	0.59	0.44	0.66	0.31
Shashirekha et al. [205]	0.60	0.63	0.61	0.67	0.61	0.50	0.58	0.64	0.62	0.63	0.72	0.39	0.73	0.28
Ensemble (Top 3)	**0.78**	**0.77**	**0.75**	**0.80**	**0.76**	0.66	**0.78**	**0.78**	**0.77**	**0.77**	**0.82**	**0.64**	**0.82**	**0.61**
Ensemble ($\tau = 3$)	0.68	0.72	0.72	0.72	0.72	**0.68**	0.76	0.74	0.71	0.71	0.76	**0.64**	0.77	0.58
Ensemble ($\tau = 5$)	0.74	0.75	0.73	0.78	0.74	0.64	0.76	0.76	0.74	0.75	0.80	0.61	0.81	0.55
Ensemble ($\tau = 7$)	0.62	0.67	0.65	0.71	0.65	0.54	0.64	0.68	0.66	0.67	0.75	0.44	0.75	0.31

Table 5.20: Participant scores at the shared task on trigger detection. Shown are scores of examples with certain properties based on different document lengths, number of freeform tags (tag confidence), popularity confidence (hits, kudos, comments, bookmarks), works with, without, and with unspecified AO3 *Archive Warning*, and works with or without and explicit or mature rating. Participants are sorted by total macro F_1 (see Table 5.18).

score, the submissions are slightly (0–0.1) less effective on explicit works. This also contradicts our assumptions that explicit or mature works are easier to classify.

Discussion

Several factors affect the system effectiveness. First, encoding and training on the full documents is important for good scores on long documents, and hierarchical models seem to work best in this respect. The key is to find triggering passages that appear only in some parts of the document and that influence the classification decision, rather than finding the topic or style that is also present at the beginning. Surprisingly, short documents appear to be much harder to classify, so models with a strong encoding for short texts (BERT) are important and document vectors are less effective as features. None of the top models manage to be good at both short and long document effectiveness, leaving room for improvement. The effect sizes for all metadata comparisons are small (about 0.05–0.15).

Second, all submissions are much less effective on rare labels and very effective on very common labels. The triggering concept goes beyond what can be observed from the passages in the training data, so the models cannot connect the triggers in the test data to the learned concept.

Third, the submissions are more effective on popular fan fiction and less effective on those with a *Choose Not To Use Archive Warnings* declaration. Authors' diligence in annotating freeform tags varies widely, so some works are under-tagged (i.e., authors want to avoid spoilers), and authors are more diligent in assigning warnings to popular works. However, we also find that submissions are less effective for works with many freeform tags, so the assumption the over-tagging decreases label reliability also has merit.

5.3.4 Conclusion

This section presents three extensive experiments on trigger warning classification. We find that triggers warnings can very effectively, with more than 0.9 F_1, be assigned through classification: for common warnings and on a coarse granularity, as with *Violence* and *Pornography*. On the flip side, the classifiers often have a low recall for many warnings, which is a problem since missing a warning can harm readers. We identify three larger areas where future work can improve trigger detection: coping with long documents, rare warnings, and label noise.

Regarding long documents, all three experiments are conclusive in that the need for a warning is often not apparent at the beginning of a documents and that effective classifiers must consider the complete documents. This is essential, as the best current technology, transformer models, are very limited in their input length (512 tokens for BERT) and become either very expensive or lose performance when extending the input length. Consequently, the best performing models either use count vectors that aggregate whole documents, or use some form of hierarchical classification.

Regarding rare warnings, we find that models perform worse on rare labels than on common ones. There are multiple potential explanations: Firstly, the common labels, notably *Pornography* from experiment 3 is much easier to detect based on unambiguous keywords and *Violence* from experiment 1 is more rigorously annotated. Second, as is typical for classification, it is harder to generalize when there are fewer examples to learn from.

Regarding label noise, the shared task evaluation in experiment three shows that the classifiers are more effective on popular works and less effective on works with an *Choose Not To Use Archive Warnings* declaration. A convincing explanation for this finding is that fan fiction authors are not always rigorously annotating trigger warnings via freeform tags, especially since this annotation is not supported or organized and there are no guidelines besides a common understanding in the community. As a consequence, some works are not tagged while others are tagged liberally, which, from a model evaluation perspective, results in label noise. We study this label noise in the following Section 5.4.

Limitations It should be noted that our contributions to trigger warnings and trigger detection are limited to fan fiction documents. Models trained on our datasets might not transfer to other online content like news articles, websites, or social media posts. Particularly social-media texts are shorter and contain fewer descriptions and more verbal expressions, which is a substantial-enough shift to warrant models explicitly trained in the genre. Similarly, the conclusions of our experiments are limited by the models we used, as well as the genre of the text. Furthermore, the trigger warning scheme we used is a simple structure. Further research should investigate more detailed trigger (warning) typologies with a more rich semantics.

Impact We hypothesize that an automatic assignment of trigger warnings can help reduce the impact of distressing content on vulnerable groups. They would solve the problem that most social media providers are unwilling[5] or unable to integrate trigger warnings into their platforms, as users could have them automatically assigned by their respective devices before they see disturbing content.

Another potential positive impact of analyzing trigger warnings, such as those voluntarily used by social media users, is that this data can partially if not completely relieve the burden on human content moderators who are otherwise constantly confronted with extreme content. This is especially relevant to the recent news that OpenAI has outsourced content moderation for ChatGPT's output to Kenyan workers.[6] This news follows earlier reports that major social media platforms have done or are still doing the same thing to Filipino workers.[7] Any technology that helps make this type of manual moderation obsolete is very welcome. The labels obtained from manual moderation by these workers will of course be used by OpenAI and the social media providers to develop specific moderation models for ChatGPT or their platforms. We are not currently in a position to analyze whether a domain transfer from fan fiction to these moderation tasks is possible, nor do we know whether web data labeled with trigger warnings are already being used for these purposes in the aforementioned companies, but found insufficient for their purposes. Nor are fan fiction sites likely to cover all aspects of distressing content generated by large language models or found on social media. Nor does any of this absolve companies of their currently largely neglected duty to take responsibility for the welfare of their (external) workers.

Regarding potential negative impacts of this work, first, the presented data contains annotated, potentially distressing content, like violence or rape, in sufficient quantities to train generative models. This calls for taking measures to ensure one's personal health of body and mind when conducting manual data analyses with a focus on such distressing content, as exemplified by the above moderation example. Second, some content on AO3 might border on legality in some countries, and dependent on who owns it for what purposes, in particular regarding descriptions of underage sexuality and pedophilia, where what is considered underage differs from country to country. Some works might have meanwhile been removed from the

[5] E.g., for fear of possible backlash from the community: `https://www.lbc.co.uk/news/universities-backlash-trigger-warnings-on-english-literature-texts/`

[6] `https://time.com/6247678/openai-chatgpt-kenya-workers/`

[7] `https://www.washingtonpost.com/technology/2019/07/25/social-media-companies-are-outsourcing-their-dirty-work-philippines-generation-workers-is-paying-price/`

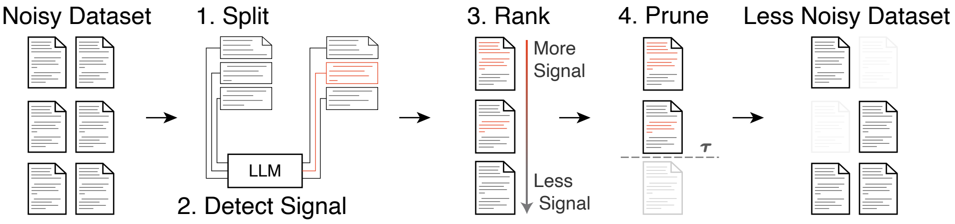


Figure 5.6: Overview of the proposed method of pruning documents with a label depending on how strong the signal for this label is according to an LLM classifier.

platform but are still included in our dataset. As a precaution, we do not release the works' text in our datasets. Instead, we release only work IDs and utilities to scrape the text from AO3. We further maintain an archived version for reproducibility and ongoing research. Third, some of the stories are written about real, living humans and may include details about them. Additionally, some stories might contain information about the author. Lastly, we used the data only partially compliant with its intended use: The AO3 tags are intended as trigger warnings, and the fan fiction stories are intended to be read.

5.4 De-Noising the Trigger Detection Dataset

One of the key findings across all classification experiments in the previous sections is that label noise in the dataset is likely to degrade the evaluation results. Here, we present a three-step de-noising strategy (see Figure 5.6) to specifically remove false-positive labeled examples from the dataset: Given the labeled documents, we use large language models to estimate "how strong the signal within a document is in the direction of its class label," rank all documents according to their estimated signal strength, and drop documents below a certain threshold.

We evaluate our proposed method using three well-performing models (XGBoost, RoBERTa, and Longformer) on the multi-label trigger detection dataset from Section 5.2, which provides some organic information about label reliability. The results show that our method increases the ratio of noisy to reliable documents in the benchmark from 1:1 to 1:6, that models tested on de-noised data score up to 0.15 F_1 higher than when tested on "noisy" documents, and that models can score the same on noisy data but significantly different on the de-noised dataset.[8]

[8] The datasets and code can be found at: `https://github.com/webis-de/CLEF-24`

5.4.1 Classification Benchmarks Degrade under Noise

There are text classification tasks for which providing a sufficient amount of labeled data is difficult. The difficulty may be due to the subjectivity of the task (Is this text a product *description* or a product *advertisement*?), a high number of classes (Which of the 188 cognitive biases occur in this text?), a missing dichotomy since only one class can be characterized (Does this text has an enticing writing style?), the need for expert knowledge (Is argument A more convincing than argument B?), or a combination of these characteristics. For such tasks, LLMs have shown great performance, even in zero-shot settings, but especially with in-context learning and chain-of-thought reasoning [233].

But, just as powerful as LLMs are in this respect, they are obviously not a panacea: Time, cost, and latency are among their main limiting factors, especially for classification tasks that require ad hoc decisions and high throughput. Consider, for example, the generation of a search engine result page (SERP) on which documents containing product advertising, undesirable prejudices, or sarcasm are to be filtered out. The practical and efficient approaches, instead, fine-tune neural networks based on dense document representations, such as BERT or RoBERTa [125]. Their limiting factor, however, is the knowledge acquisition bottleneck, i.e. the lack or the quality of labeled data. This lack of labeled data is often countered by collecting data from weakly-supervised sources. One example of this is the extraction of trigger warnings from online blogs, where authors signal if their work contains harmful content. One example of this is the extraction of debate portals, where online debaters rate the persuasive power of arguments.

However, weakly-supervised data acquisition leads to noisy data due to errors or inconsistencies in the distant knowledge source. The use of noisy data to benchmark classification models (which is the focus of this chapter) is problematic: model performances may be underestimated, model differences may be smaller or vanish, or, in the worst case, leaderboard rankings change. Or the other way around: reducing label noise in benchmark data increases model scores and may increase the performance difference between models, which makes it easier to assess which model is actually better and by how much.

5.4.2 Related Work

Although current (pre-trained) deep learning models are somewhat robust to label noise given sufficient training data [194, 260], reducing label noise is still essential when training non-neural models [71, 148] or with limited training data. Most related work focuses on training data de-noising neural classifiers [123, 258], especially with semi-supervised methods like adapting the loss function [157, 201], by over-parameterization [124], or by rank pruning [152] via predicted probabilities. Some related works also use weak supervision methods to estimate label reliability [188, 190] from (multiple) external sources. For our work, we adapt the rank pruning idea but use an external source (an LLM) instead of a semi-supervised signal. However, the most notable difference of our work is that we do not focus on de-noising the training data to improve the model but the test data to improve the benchmark reliability, which is why we study organic noise instead of only injecting synthetic noise like the related work (e.g., on TREC question-type and AG-News datasets [74]).

5.4.3 Finding and Pruning Noisy Documents

Our label de-noising procedure assumes the following: First, the input dataset contains a set of documents, and each document has one or more labels from a finite set. Second, each reliable document with a true positive label contains a signal above a confidence threshold τ (i.e., a piece of set) that justifies the label. Third, there are a number of noisy documents that have been assigned a positive label where the signal with respect to that label is weaker than τ. Our pruning strategy, illustrated in Figure 5.6, attempts to find and remove documents that are noisy with respect to a particular label by determining the signal strength of that label.

To do this, we rank all documents independently for each label according to the strength of the signal of this label and then determine τ as a threshold. The de-noising scheme consists of four steps for each label: (1) Splitting of documents into smaller chunks, i.e. several consecutive sentences, where the chunk size is a hyperparameter. (2) Determine whether a chunk carries a signal for the label using a prompt-based binary classification, where LLM and prompt are hyperparameters that depend on the task and the label. (3) Ranking of the documents on the basis of the absolute number of signals, i.e. the positively classified chunks. (4) Pruning of the documents with the lowest rank up to a rank or signal strength threshold τ.

Warning	**Source Data**		**Sample used in this Work**			**Length**	
	Unknown	Reliable	Unknown	Flipped	Reliable	Mean	Std
Death	124,958	1,579	600	200	200	3,351	2,717
Violence	119,684	1,736	600	200	200	4,021	2,853
Homophobia	22,688	558	600	200	200	4,125	2,809
Self-harm	23,029	1,343	600	200	200	3,478	2,688

Table 5.21: Number and length of the source and the evaluation documents.

5.4.4 De-Noising Trigger Warning Assignment

LLM-based de-noising is evaluated on a multi-label classification task by measuring the noise ratio and model effectiveness at different τ.

Dataset

We use evaluation data from the Webis Trigger Warning Corpus (WTWC) [245], which was used in the 2023 shared task on trigger detection [244]. The WTWC is well suited as it contains organic false positive and negative labels that emerge from human authors (sensitive human authors assign warnings for weak signals) and from weakly supervised labeling (which assigns warnings for loosely related or implied concepts). The dataset also contains additional reliability information in the "author notes" prepended to some chapters.

The test data is a sample of 4,000 WTWC documents balanced between 4 warning labels *Death*, *Violence* (the two most common warnings, excluding Pornography as outlier), *Homophobia* and *Self-harm* (the two closest to median frequency with sufficient *Reliable* documents) as our evaluation dataset (cf. Table 5.21), which is large enough to test our method. For each label, we first sample 200 *Reliable* documents where the author note mentions either `tw, cw, trigger(s), content warning` within 20 tokens of a warning term (e.g. `homophobia`). Then, we sample 800 non-*Reliable* documents and create a subset of 200 known falsely labeled data by *Flipping* the documents' label to a different one. The reliability of the remaining 600 documents was marked as *Unknown*. We adopted all other sampling criteria from experiment 3 presented in Section 5.3.3 (English documents with 50-10,000 words).

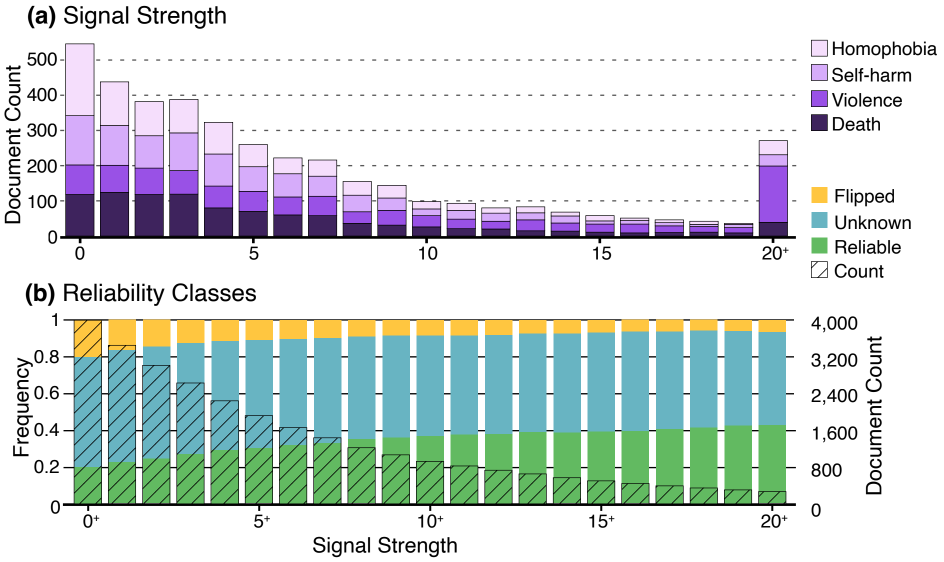


Figure 5.7: (**a**) Signal strength distribution: Number of documents with a certain amount of positively classified 5-sentence chunks by label. (**b**) Number of documents and document reliability in the pruned corpus at different thresholds.

De-Noising Implementation

The de-noising technique is applied using 5 consecutive, non-overlapping sentences as chunks and `Mixtral-8x7B-v0.1` from Huggingface as large language model. We use a binary classification prompt derived from Mistral's prompting guide:

```
You are a text classification model.  You determine if a given
text contains death, graphic display of death, murder, or dying
characters.  If the given text contains intense, explicit, and
graphic death, you answer:  Yes.  If the text contains mild or
implicit death or no death at all, you answer:  No.
```

Chunks are classified by predicting the next-token probabilities, given the above prompt, and comparing the logits of the `Yes` and `No` tokens. Documents are ranked and pruned according to the absolute number of positively (`Yes` > `No`) classified chunks per document, i.e. at a τ of 5^+ all documents with less than 5 positive chunks will be pruned.

Experiments and Evaluation

To evaluate the hypotheses we conduct three experiments across three baseline classification models. First, pruning the complete dataset (with τ from

0^+ to 20^+) and observing the ratio of reliability classes. Second, splitting the data 80:20 into training and test and only pruning the test dataset (with τ from 0^+ to 5^+)[9] while training the models on the complete training data. Third, pruning the complete dataset (with τ from 0^+ to 5^+) before the train-test split and also training the models with pruned data. Decreasing scores in this last experiment would indicate that our method also removes (many) difficult cases, leading to both, poor models and a poor benchmark.

All three models are trained for multi-label classification: a fine-tuned `FacebookAI/roberta-base` and `allenai/longformer-base-4096` [17] and a feature-based `XGBoost` [42] classifier (the baseline of the shared task [244]) with the top 10,000 $\mathrm{tf} \cdot \mathrm{idf}$ word 1–3-gram features selected via χ^2. The RoBERTa input was truncated to 512 tokens and the Longformer input to 4,096 tokens. We report the micro-averaged multi-label F_1 via a 5-fold Monte Carlo cross-validation and the 95% t-estimated confidence intervals.

RoBERTa was trained for 10 epochs with a learning rate (LR) of 2e-5 and Longformer for 7 epochs and 2e-5. XGBoost was trained with 50 estimators, 3 maximum depth, and 0.5 LR. All training parameters were tuned using grid search on an independent split. We tested `{7, 10, 20}` epochs and `{5e-4, 1e-5, 2e-5, 5e-5}` LR for the neuronal models and `{2, 3, 4}` maximum depth and `{0.25, 0.5, 0.75}` LR for XGBoost.

Results and Discussion

The first hypothesis is that the de-noising method removes noise from the dataset if, with increasing τ, the proportion of *Reliable* documents increases and of *Flipped* documents decreases. Figure 5.7(b) shows that the proportion of *Reliable* documents increases from 0.2 to 0.41 and decreases for *Flipped* documents from 0.2 to 0.05. Note that the proportion changes are strongest for smaller τ.

The first hypothesis is that de-noising improves the benchmark when the models' test scores increase with increased de-noising (for train-test and test-only pruning) and when the relative difference between models' test scores changes. Figure 5.8(a) shows that the F_1 of all models increases by 0.05–0.1 with $\tau = 5^+$ when pruning only the test data. The effect is strongest for XGBoost and weakest for RoBERTa (where the input documents are strongly truncated). Figure 5.8(a) also shows that XGBoost and RoBERTa score evenly without pruning but XGBoost improves more strongly and is

[9] At $\tau = 5^+$, half the dataset has been pruned.

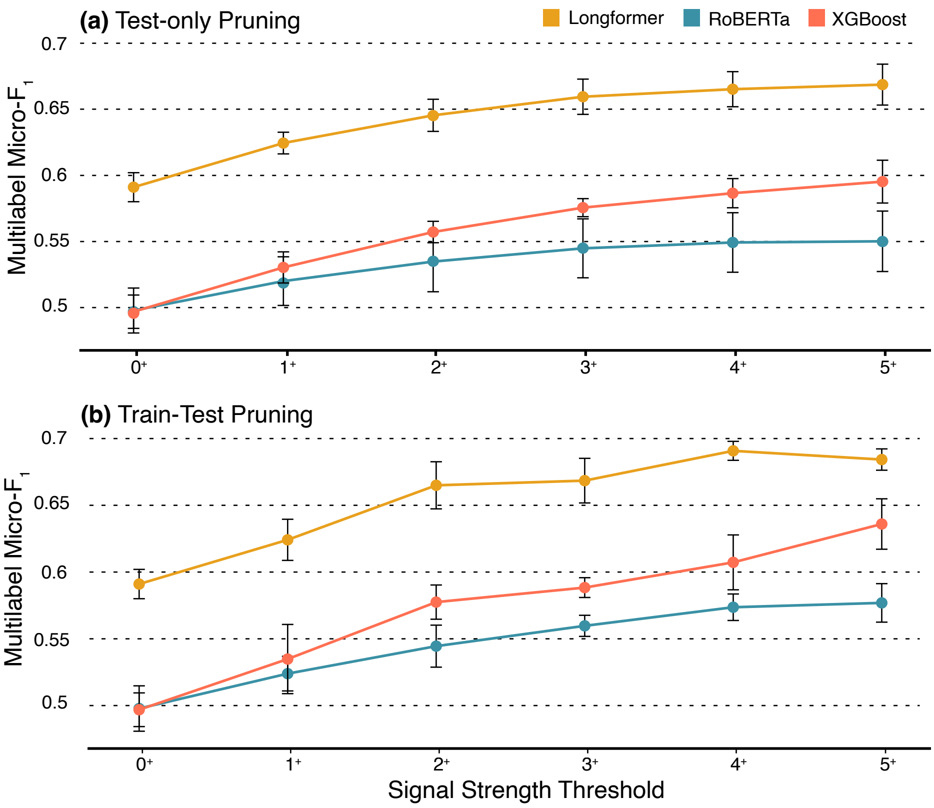


Figure 5.8: Model F_1 with confidence intervals of three classification models at different pruning thresholds when (a) only test data and (b) training and test data are pruned.

significantly more effective with $\tau = 5^+$. This shows that de-noising can reveal model differences that are otherwise hidden by the noise. Figure 5.8(b) shows that the F_1 of all models increases when pruning all data and more strongly than when only pruning the test data.

5.4.5 Conclusion

In this section, we investigate using rank-based pruning based on an LLMs classification signal to de-noise a document-level trigger warning classification dataset. We present a new, LLM-based de-noising strategy and show that it doubles the relative number of reliably labeled documents and halves the noisily labeled ones. We further show that the de-noising strategy increases the model scores and the differences between models, hence we assume that the de-noised dataset is more suited as a benchmark.

6
Conclusion

This dissertation reveals the conceptual principles of using weak supervision to create large and novel datasets based on user-generated content. Our framework describes how to make weak labels accessible for a variety of tasks by identifying appropriate sources of distant knowledge and strategies for linking data and knowledge. Our framework can provide standard solutions to labeling problems, inspire discovering new sources of knowledge, and developing linking strategies for specific problems.

We compile the strength and robustness of these principles through three case studies that focus on the use of three novel datasets to answer relevant research questions. This chapter presents the main contributions and results of each case study, as well as the takeaways for future research in weakly supervised data labeling.

6.1 Main Contributions and Findings

A main conceptual contribution of this work, presented in Chapter 2, is the study of the principles of weak supervision for generating large labeled datasets. Section 2.1 presents our unified view of supervised learning strategies used to solve the label scarcity problem, providing comparative definitions, etymological notes, and describing what distinguishes weak supervision from semi-supervised and self-supervised learning. We find that the main differences lie in how the missing relationship between labels and data points is handled: while semi-supervised learning uses knowledge from labeled data points to make assumptions about unlabeled ones, and self-supervised learning derives labels directly from the data object, weak supervision determines labels using some form of distant knowledge.

Section 2.2 presents a comprehensive search, the selection of 35 high-quality publications, and their systematic review: We determine the established design parameters, i.e., tasks, platforms, data types, and types of distant knowledge, the linking strategies used for weakly supervised data labeling, and the available means to evaluate the effectiveness of the supervision strategy. We find that classification tasks such as misinformation or location prediction are common, but also language processing tasks such as text normalization and information extraction, and computational social science tasks such as trend prediction or bot detection. Twitter has been a popular data source, especially for post- or user-level data, due to its popularity among users and ease of data access. The review identified seven types of distant knowledge, curated lists, databases, web data, metadata, distant metadata, computed metadata, and classifiers, as well as five types of evaluation strategies, spot checks, weak labels, annotated data, and models. The choice of the appropriate labeling and evaluation strategies depends on the accessible distant knowledge for the data to be labeled and the task to be solved. However, if the knowledge is "very distant" from the data points, connecting them introduces label noise, which allows for a more comprehensive evaluation.

We build on these concepts by constructing three large, novel datasets needed to investigate several relevant research questions, organized into three case studies that are concluded in the following sections. For each, we apply weak supervision to new domains and platforms, such as fan fiction documents on Archive of Our Own, and develop new linking strategies that combine multiple distant knowledge sources.

6.1.1 Persuasiveness of Debaters on Reddit

Chapter 3 presents our case study on the persuasiveness of debaters on Reddit. The dataset created for this study includes 3,801 debaters from Reddit's ChangeMyView debate forum, each debater's debate contributions, whether it was persuasive or not, and the debater's persuasiveness over their active period.

We use this dataset to investigate why some debaters are more persuasive than others by modeling debater effectiveness in ChangeMyView and analyzing their behavior and argumentative style choices. We find that persuasiveness improves over time for the average debater, that the distribution of "frames" in debaters' arguments can play an important role in persuasiveness, and that characteristics based on the presence of certain types of arguments in debaters' text do not sufficiently indicate persuasiveness.

Although our work significantly advances the understanding of persuasiveness, several questions remain. First, despite general agreement on what is persuasive, there are differences in the evaluation of persuasiveness based on the positions of the evaluators, which are largely ignored by us and related work. In addition, there are still opportunities to better model features based on argumentative units that were ineffective in the experiments, to add deeper behavioral features such as experience and the dynamics of debater interaction, and to use more sophisticated models such as conditional random fields to model the cumulative effect of persuasion in ChangeMyView discussions, or newer transformer-based models such as large language models.

6.1.2 Profiling Influencers on Twitter

Chapter 4 presents our case study on influencer profiling on Twitter. The dataset created for this study, the "Webis Celebrity Corpus 2019", contains 71,706 influencers, each influencer's full Twitter timeline, and up to 239 attributes from Wikidata. Many of these attributes are not available in general population datasets, and our dataset allows their study for the first time. We developed a new linking strategy to link Twitter accounts to the corresponding Wikidata entities, using a set of heuristics to generate candidate matches based on names and titles, and then filtering the likely mismatches based on Wikidata properties. Using a weak label evaluation strategy, which is one of the most reliable evaluation strategies for weak supervision, we can validate that the method achieves a high precision (0.994) with a reasonable recall (0.723).

We demonstrate the usefulness of the dataset by organizing a shared task with the goal of developing author profiling technology, where eight participants submitted models to profile gender, year of birth, fame and occupation. The most reliable models used traditional machine learning with a combination of content and style-based features. Remaining challenges include (1) predicting rare demographics, such as diverse gender, very young influencers born after 2000, and identifying influencers of little renown, (2) predicting occupations without clear topical separation, like professional, manager, science, and creative, and (3) the discrimination of authors born before 1980. We also show that profiling models can be trained on the influencer dataset and applied to general population datasets with an acceptable loss of effectiveness of about 0.05 F_1.

Moreover, we use the dataset to investigate a novel research question: whether influencers on Twitter can be profiled using only their fans' posts. For this study, we extend the influencer dataset to include the complete timelines of 10 followers for 2,320 influencers, and organize another shared task in which four participants submitted models to profile gender, year of birth, and occupation. The best submitted models again use proven methods, feature-based machine learning with stylometric and n-gram features. The evaluation shows similar strengths and weaknesses between follower-based and author-based profiling models: They work best when the occupational classes are topically coherent, and they profile younger authors more accurately than older ones. However, the results impressively show that it is possible to profile authors based on their fans' texts almost as well as on their own texts. Using follower messages to improve author profiling models is a promising future direction.

6.1.3 Trigger Warning Assignment

Chapter 5 presents our case study on assigning trigger warnings to fan fiction works on Archive of Our Own. The dataset created for this study, the "Webis Trigger Warning Corpus 2022", contains about 1 million documents labeled with trigger warnings. To create the dataset, we developed a 36 warning taxonomy and determined the trigger warnings of each document by combining distant knowledge from the free-form content descriptors added by authors, as well as the relationships between content descriptors added by site moderators, and developed a novel linking strategy using multiple heuristics and graph propagation rules to translate the content descriptors into trigger warnings. Because this linking strategy is complex, we also evaluate the data using both spot checks, which show an F_1 of 0.95, and weak labels, which show a recall across all tags of 0.86, the two most reliable evaluation strategies applicable to weak supervision.

With this dataset, we investigate whether trigger warnings can be assigned to documents with sufficient quality, first by classifying only *Violence*, second by varying data and label set parameters via multi-label classification, and third, by organizing a shared task and analyzing the seven submitted models. We find that trigger warnings can be effectively classified for common warnings and at a coarse granularity, such as with *Violence* and *Pornography*, and that models that consider the entire documents and not just the beginning perform best, which sets trigger detection apart from other multi-label classification datasets. However, we also find that the classifiers often have low recall for many warnings, which is problematic for applications in content moderation where false negatives can be harmful, and we find that classifiers have a low precision on rare labels and in open-set situations. In addition, the experimental results suggest that the dataset contains unmitigated label noise, which degrades the model scores because authors declare harmful topics even though there is little textual support for them.

Following the previous study, we investigate the influence of label noise in the dataset on the evaluation of trigger detection models and present "LLM-based rank pruning" as a method to reduce label noise in document classification datasets. We evaluate this method by de-noising a sample of our trigger warning dataset and find that we can reduce noisy labels by 75%, thereby increasing model scores and revealing hidden model differences.

6.2 Future Directions for Weak Supervision

In this dissertation, we discuss and demonstrate a number of ways to use weak supervision to construct large and novel datasets for a variety of tasks centered on user-generated content. Our conceptual and applied work also highlights the limitations of weak labeling in accessibility, evaluation, and noise reduction.

Regarding accessibility, since weak supervision strategies are generally constrained by availability and access to data and labels about the problem at hand, we find that constructed dataset are often limited in domain and genres, which introduces a strong bias in these datasets. In particular, data labeling was constrained by the format and permissions to access social media data. For example, the linking strategy and distant knowledge of our influencer profiling dataset would work for any platform that highlights its influencers in a technical way, for example using a Twitter-like verification check mark. Since this form of highlighting is rare, our dataset is limited to microblog timelines. Similar limitations apply to the other two datasets. This introduces a strong bias into the datasets and into all analysis results. These biases prevent the use of models trained on the data in broader contexts, as the content created on different platforms is very different.

Another accessibility issue is the platform monoculture in all the fields of study we considered: most of the papers used Twitter as a basis for their research because of the amount of data and the ease of access. With the recent restrictions on data access on these and most other platforms, collecting new datasets of user-generated content with the established strategies is an open problem. One possible direction for methodological research is to transfer or adapt the linking strategies to different data or data from different platforms, as is being attempted at the time of writing with the microblogging platform Bluesky.

Regarding evaluation, we find that effective and affordable evaluation of weak labeling strategies is an open research problem, where much work in the field does not or only insufficiently evaluate the used linking strategies nor validate the resulting datasets, because reliable evaluation either depends on the availability of a weak label or, in the case of spot checks, is a high additional cost. The development of reliable evaluation strategies or data validation tools is a promising direction.

Finally, regarding noise reduction, we find that even well-functioning weak labeling strategies can introduce label noise and degrade data quality, especially for test data used to evaluate model performance. This is evident in our case study on trigger warnings. Quantifying label noise in test data, assessing its impact on model rankings, and developing automated tools to reduce label noise are relevant future research directions.

Bibliography

[1] C. J. Adams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukierski. 2017. Toxic comment classification challenge.

[2] Thomas Aichner, Matthias Grünfelder, Oswin Maurer, and Deni Jegeni. 2021. Twenty-five years of social media: A review of social media applications and definitions from 1994 to 2019. *Cyberpsychol Behav Soc Netw*, 4(24):215–222.

[3] Khalid Al-Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. 2020. Exploiting personal characteristics of debaters for predicting persuasiveness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7067–7072. Association for Computational Linguistics.

[4] Nikolaos Aletras and Benjamin Paul Chamberlain. 2018. Predicting twitter user socioeconomic attributes with network and language information. In *Proceedings of the 29th on Hypertext and Social Media*, pages 20–24.

[5] Roobaea Alroobaea, Ahmed H. Almulihi, Fahd S. Alharithi, Seifeddine Mechti, Moez Krichen, and Lamia Hadrich Belguith. 2020. A Deep Learning Model to Predict Gender, Age and Occupation of the Celebrities based on Tweets Followers. In *CLEF 2020 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

[6] Nourah Alswaidan and Mohamed El Bachir Menai. 2020. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, 62(8):2937–2987.

[7] Miguel A. Álvarez-Carmona, A. Pastor López-Monroy, Manuel Montes y Gómez, Luis Villaseñor-Pineda, and Hugo Jair Escalante. 2015. INAOE's participation at PAN'15: Author Profiling task—Notebook for PAN at CLEF 2015. In *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. CEUR-WS.org.

[8] David Alvarez-Melis and Martin Saveski. 2016. Topic modeling in twitter: Aggregating tweets by conversations. In *Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20, 2016*, pages 519–522. AAAI Press.

[9] Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2009. Automatically Profiling the Author of an Anonymous Text. *Commun. ACM*, 52(2):119–123.

[10] Muhammad Usman Asif, Naeem Shahzad, Zeeshan Ramzan, and Fahad Najib. 2019. Word Distance Approach for Celebrity profiling—Notebook for PAN at CLEF 2019. In *CLEF 2019 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

[11] Abinew Ali Ayele, Nikolay Babakov, Janek Bevendorff, Xavier Bonet Casals, Berta Chulvi, Daryna Dementieva, Ashaf Elnagar, Dayne Freitag, Maik Fröbe, Damir Korenčić, Maximilian Mayerl, Daniil Moskovskiy, Animesh Mukherjee, Alexander Panchenko, Martin Potthast, Francisco Rangel, Naquee Rizwan, Paolo Rosso, Florian Schneider, Alisa Smirnova, Efstathios Stamatatos, Elisei Stakovskii, Benno Stein, Mariona Taulé, Dmitry Ustalov, Xintong Wang, Matti Wiegmann, Seid Muhie Yimam, and Eva Zangerle. 2024. Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 15th International Conference of the CLEF Association (CLEF 2024)*, Lecture Notes in Computer Science, Berlin Heidelberg New York. Springer.

[12] Adam Badawy, Emilio Ferrara, and Kristina Lerman. 2018. Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In *IEEE/ACM 2018 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018, Barcelona, Spain, August 28-31, 2018*, pages 258–265. IEEE Computer Society.

[13] Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 2020. What was written vs. who read it: News media profiling using text analysis and social media context. *arXiv preprint arXiv:2005.04518*.

[14] Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A unified taxonomy of harmful content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms, WOAH 2020, Online, November 20, 2020*, pages 125–137. Association for Computational Linguistics.

[15] Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim. 2017. N-GrAM: New Groningen Author-profiling Model—Notebook for PAN at CLEF 2017. In *CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Dublin, Ireland*. CEUR-WS.org.

[16] Hila Becker, Mor Naaman, and Luis Gravano. 2010. Learning similarity metrics for event identification in social media. In *Proceedings of the Third International Conference on Web Search and Web Data Mining, WSDM 2010, New York, NY, USA, February 4-6, 2010*, pages 291–300. ACM.

[17] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

[18] Shane Bergsma, Matt Post, and David Yarowsky. 2012. Stylometric analysis of scientific articles. In *HLT-NAACL*, pages 327–337. The Association for Computational Linguistics.

[19] Janek Bevendorff, Ian Borrego-Obrador, Mara Chinea-Ríos, Marc Franco-Salvador, Maik Fröbe, Annina Heini, Krzysztof Kredens, Maximilian Mayerl, Piotr Pęzik, Martin Potthast, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, Benno Stein, Matti Wiegmann, Magdalena Wolska, , and Eva Zangerle. 2023. Overview of PAN 2023: Authorship Verification, Multi-Author Writing Style Analysis, Profiling Cryptocurrency Influencers, and Trigger Detection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023)*, Lecture Notes in Computer Science, pages 459–481, Berlin Heidelberg New York. Springer.

[20] Janek Bevendorff, Xavier Bonet Casals, Berta Chulvi, Daryna Dementieva, Ashaf Elnagar, Dayne Freitag, Maik Fröbe, Damir Korenčić, Maximilian Mayerl, Animesh Mukherjee, Alexander Panchenko, Martin Potthast, Francisco Rangel, Paolo Rosso, Alisa Smirnova, Efstathios Stamatatos, Benno Stein, Mariona Taulé, Dmitry Ustalov, Matti Wiegmann, and Eva Zangerle. 2024. Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification. In *Advances in Information Retrieval. 46th European Conference on IR Research (ECIR 2024)*, Lecture Notes in Computer Science, Berlin Heidelberg New York. Springer.

[21] Janek Bevendorff, BERTa Chulvi, Gretel Liz De La Peña Sarracén, Mike Kestemont, Enrique Manjavacas, Ilia Markov, Maximilian Mayerl, Martin Potthast, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, Benno Stein, Matti Wiegmann, Magdalena Wolska, and Eva Zangerle. 2021. Overview of PAN 2021: Authorship Verification, Profiling Hate Speech Spreaders on Twitter, and Style Change Detection. In *Advances in Information Retrieval. 43rd European Conference on IR Research (ECIR 2021)*, volume 12036 of *Lecture Notes in Computer Science*, pages 567–573, Berlin Heidelberg New York. Springer.

[22] Janek Bevendorff, Berta Chulvi, Elisabetta Fersini, Annina Heini, Mike Kestemont, Krzysztof Kredens, Maximilian Mayerl, Reyner Ortega-Bueno, Piotr Pezik, Martin Potthast, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, Benno Stein, Matti Wiegmann, Magdalena Wolska, and Eva Zangerle. 2022. Overview of PAN 2022: Authorship Verification, Profiling Irony and Stereotype Spreaders, and Style Change Detection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 13th International Conference of the CLEF Association (CLEF 2022)*, volume 13186 of *Lecture Notes in Computer Science*, pages 382–394, Berlin Heidelberg New York. Springer.

[23] Janek Bevendorff, Bilal Ghanem, Anastasia Giachanou, Mike Kestemont, Enrique Manjavacas, Ilia Markov, Maximilian Mayerl, Martin Potthast, Francisco Rangel, Paolo Rosso, Günther Specht, Efstathios Stamatatos, Benno Stein, Matti Wiegmann, and Eva Zangerle. 2020. Overview of PAN 2020: Authorship Verification, Celebrity Profiling, Profiling Fake News Spreaders on Twitter, and Style Change Detection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 10th International Conference of the CLEF Initiative (CLEF 2020)*, volume 12260 of *Lecture Notes in Computer Science*, pages 372–383, Berlin Heidelberg New York. Springer.

[24] Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2018. Elastic ChatNoir: Search Engine for the ClueWeb and the Common Crawl. In *Advances in Information Retrieval. 40th European Conference on IR Research (ECIR 2018)*, Lecture Notes in Computer Science, Berlin Heidelberg New York. Springer.

[25] Janek Bevendorff, Matti Wiegmann, Martin Potthast, and Benno Stein. 2022. The Impact of Online Affiliate Marketing on Web Search. In *4th International Symposium on Open Search Technology (OSSYM 2022)*. International Open Search Symposium.

[26] Janek Bevendorff, Matti Wiegmann, Martin Potthast, and Benno Stein. 2024. Is Google Getting Worse? A Longitudinal Investigation of SEO Spam in Search Engines. In *Advances in Information Retrieval. 46th European Conference on IR Research (ECIR 2024)*, Lecture Notes in Computer Science. Springer.

[27] Janek Bevendorff, Matti Wiegmann, Martin Potthast, and Benno Stein. 2024. Product Spam on YouTube: A Case Study. In *ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR 2024)*. ACM.

[28] K. Bhatia, K. Dahiya, H. Jain, P. Kar, A. Mittal, Y. Prabhu, and M. Varma. 2016. The extreme classification repository: Multi-label datasets and code.

[29] Christian Biemann. 2007. *Unsupervised and knowledge-free natural language processing in the structure discovery paradigm*. Ph.D. thesis, Leipzig University, Germany.

[30] Su Lin Blodgett, L. Green, and Brendan T. O'Connor. 2016. Demographic dialectal variation in social media: A case study of african-american english. *ArXiv*, abs/1608.08868.

[31] Erin E Buckels, Paul D Trapnell, and Delroy L Paulhus. 2014. Trolls just want to have fun. *Personality and individual Differences*, 67:97–102.

[32] John D. Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating Gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309. ACM.

[33] Mart Busger op Vollenbroek, Talvany Carlotto, Tim Kreutz, Maria Medvedeva, Chris Pool, Johannes Bjerva, Hessel Haagsma, and Malvina Nissim. 2016. GronUP: Groningen User Profiling—Notebook for PAN at CLEF 2016. In *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal*. CEUR-WS.org.

[34] Guiyuan Cao, Zhongyuan Han, Haojie Cao, Ximin Huang, Zhengqiao Zeng, Yaozu Tan, and Jiyin Cai. 2023. A dual-model classification method based on RoBERTa for Trigger Detection. In *Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum*. CEUR-WS.org.

[35] Haojie Cao, Zhongyuan Han, Guiyuan Cao, Ruihao Zhu, Yongqi Liang, Siman Liu, and Minhua Huang. 2023. Trigger Warning Labeling with RoBERTa and Resampling for Distressing Content Detection. In *Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum*. CEUR-WS.org.

[36] Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 438–444. The Association for Computer Linguistics.

[37] Miguel Ángel Álvarez Carmona, Estefanía Guzmán-Falcón, Manuel Montes-y-Gómez, Hugo Jair Escalante, Luis Villaseñor Pineda, Verónica Reyes-Meza, and Antonio Rico Sulayes. 2018. Overview of MEX-A3T at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. In *IberEval@SEPLN*, volume 2150 of *CEUR Workshop Proceedings*, pages 74–96. CEUR-WS.org.

[38] Cambridge Centre for Teaching and Learning CCTL. 2023. When to use content notes. `https://www.cctl.cam.ac.uk/content-notes/how-use/when-use`. Last accessed: May 10, 2023.

[39] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy*, pages 6314–6322. Association for Computational Linguistics.

[40] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. 2006. *Semi-Supervised Learning*. The MIT Press.

[41] Ashleigh Charles, Laurie Hare-Duke, Hannah Nudds, Donna Franklin, Joy Llewellyn-Beardsley, Stefan Rennick-Egglestone, Onni Gust, Fiona Ng, Elizabeth Evans, Emily Knox, et al. 2022. Typology of content warnings and trigger warnings: Systematic review. *PloS one*, 17(5):e0266722.

[42] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

[43] Claire Childs, Department of Language, and Linguistic Science. 2021. LLS Departmental Guidance on Content Warnings. `https://www.york.ac.uk/media/abouttheuniversity/equality/documents/LLS-Departmental-Guidance-on-Content-Warnings-2021.pdf`. Last accessed: May 10, 2023.

[44] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *ICWSM*. The AAAI Press.

[45] Morgane Ciot, Morgan Sonderegger, and Derek Ruths. 2013. Gender inference of twitter users in non-english contexts. In *EMNLP*, pages 1136–1145. ACL.

[46] University of Toronto, Centre for Teaching and Learning CTL. 2021. Teaching Sensitive Materials. Last accessed: May 10, 2023.

[47] A. Culotta, N. Kumar, and Jennifer Cutler. 2015. Predicting the demographics of twitter users from website traffic data. In *AAAI Conference on Artificial Intelligence*, pages 72–78.

[48] Walter Daelemans, Mike Kestemont, Enrique Manjavacas, Martin Potthast, Francisco Rangel, Paolo Rosso, Günther Specht, Efstathios Stamatatos, Benno Stein, Michael Tschuggnall, Matti Wiegmann, and Eva Zangerle. 2019. Overview of PAN 2019: Bots and Gender Profiling, Celebrity Profiling, Cross-domain Authorship Attribution and Style Change Detection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 10th International Conference of the CLEF Initiative (CLEF 2019)*, volume 11696 of *Lecture Notes in Computer Science*, pages 402–416, Berlin Heidelberg New York. Springer.

[49] Jifeng Dai, Kaiming He, and Jian Sun. 2015. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1635–1643.

[50] Xiang Dai, Ilias Chalkidis, Sune Darkner, and Desmond Elliott. 2022. Revisiting transformer-based models for long document classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 7212–7230. Association for Computational Linguistics.

[51] Saman Daneshvar and Diana Inkpen. 2018. Gender Identification in Twitter using N-grams and LSA—Notebook for PAN at CLEF 2018. In *CLEF 2018 Evaluation Labs and Workshop – Working Notes Papers, 10-14 September, Avignon, France*. CEUR-WS.org.

[52] D. Davidov, Oren Tsur, and A. Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *International Conference on Computational Linguistics*, pages 241–249.

[53] D. Davidov, Oren Tsur, and A. Rappoport. 2010. Semi-supervised recognition of sarcasm in twitter and amazon. In *Conference on Computational Natural Language Learning*, pages 107–116.

[54] Munmun De Choudhury. 2015. Anorexia on Tumblr: A Characterization Study. In *Proceedings of the 5th International Conference on Digital Health 2015*.

[55] Jan Deriu, Aurélien Lucchi, V. D. Luca, Aliaksei Severyn, Simon Müller, Mark Cieliebak, Thomas Hofmann, and Martin Jaggi. 2017. *Leveraging Large Amounts of Weakly Supervised Data for Multi-Language Sentiment Classification*. ACM.

[56] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.

[57] Cícero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 189–194. Association for Computational Linguistics.

[58] Ryo Egawa, Gaku Morio, and Katsuhide Fujita. 2019. Annotating and analyzing semantic role of elementary units and relations in online persuasive arguments. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 2: Student Research Workshop*, pages 422–428. Association for Computational Linguistics.

[59] Jacob Eisenstein. 2018. *Natural Language Processing*. MIT Press.

[60] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLOS ONE*, 9(11):1–13.

[61] Jacob Eisenstein, Brendan T. O'Connor, Noah A. Smith, and E. Xing. 2010. A latent variable model for geographic lexical variation. In *Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287.

[62] Theresa Elstner, Frank Loebe, Yamen Ajjour, Christopher Akiki, Alexander Bondarenko, Maik Fröbe, Lukas Gienapp, Nikolay Kolyada, Janis Mohr, Stephan Sandfuchs, Matti Wiegmann, Jörg Frochte, Nicola Ferro, Sven Hofmann, Benno Stein, Matthias Hagen, and Martin Potthast. 2023. Shared

Tasks as Tutorials: A Methodical Approach. In *Thirteenth AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI 23)*, pages 15807–15815. EAAI.

[63] Chris Emmery, Grzegorz Chrupala, and Walter Daelemans. 2017. Simple queries as distant labels for predicting gender on twitter. In *NUT@EMNLP*, pages 50–55. Association for Computational Linguistics.

[64] Dominique Estival, Tanja Gaustad, Son Pham, Will Radford, and Ben Hutchinson. 2007. Author profiling for english emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING'07)*, pages 263–272. City University of Hong Kong.

[65] Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford, and Ben Hutchinson. 2007. TAT: an author profiling tool with application to arabic emails. In *ALTA*, pages 21–30. Australasian Language Technology Association.

[66] Golnoosh Farnadi, Jie Tang, Martine De Cock, and Marie-Francine Moens. 2018. User profiling through deep multimodal fusion. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 171–179.

[67] Mehwish Fatima, Komal Hasan, Saba Anwar, and Rao Muhammad Adeel Nawab. 2017. Multilingual author profiling on facebook. *Inf. Process. Manage.*, 53(4):886–904.

[68] Jenny Felser, Christoph Demus, Dirk Labudde, and Michael Spranger. 2023. FoSIL at PAN?23: Trigger Detection with a Two Stage Topic Classifier. In *Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum*. CEUR-WS.org.

[69] Mylynn Felt. 2016. Social media and the social sciences: How researchers employ big data analytics. *Big data & society*, 3(1):2053951716645828.

[70] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

[71] Benoît Frénay and Michel Verleysen. 2014. Classification in the Presence of Label Noise: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25:845–869.

[72] Maik Fröbe, Matti Wiegmann, Nikolay Kolyada, Bastian Grahm, Theresa Elstner, Frank Loebe, Matthias Hagen, Benno Stein, and Martin Potthast. 2023. Continuous Integration for Reproducible Shared Tasks with TIRA.io. In *Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023)*, Lecture Notes in Computer Science, pages 236–241, Berlin Heidelberg New York. Springer.

[73] Lukas Galke and Ansgar Scherp. 2022. Bag-of-words vs. graph vs. sequence in text classification: Questioning the necessity of text-graphs and the surprising strength of a wide MLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4038–4051. Association for Computational Linguistics.

[74] Siddhant Garg, Goutham Ramakrishnan, and Varun Thumbe. 2021. Towards Robustness to Label Noise in Text Classification via Noise Modeling. *30th ACM International Conference on Information & Knowledge Management*.

[75] Anastasia Giachanou, Paolo Rosso, and Fabio Crestani. 2019. Leveraging emotional signals for credibility detection. In *Proceedings of the 42nd International ACM SIGIR*, pages 877–880. ACM.

[76] Eric Gilbert and Karrie Karahalios. 2010. Widespread worry and the stock market. In *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*. The AAAI Press.

[77] Matej Gjurkovic and Jan Snajder. 2018. Reddit: A gold mine for personality prediction. In *PEOPLES@NAACL-HTL*, pages 87–97. Association for Computational Linguistics.

[78] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

[79] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.

[80] Weiwei Guo, Hao Li, Heng Ji, and Mona T. Diab. 2013. Linking tweets to news: A framework to enrich short text data in social media. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 239–249. The Association for Computer Linguistics.

[81] Zhen Guo, Zhe Zhang, and Munindar P. Singh. 2020. In opinion holders' shoes: Modeling cumulative influence for view change in online argumentation. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 2388–2399. ACM / IW3C2.

[82] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*. Springer Series in Statistics. Springer.

[83] Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING*, pages 539–545.

[84] Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strotgen, and D. Klakow. 2020. A survey on recent approaches for natural language processing in low-resource scenarios. In *North American Chapter of the Association for Computational Linguistics*, pages 2545–2568.

[85] Stefan Helmstetter and Heiko Paulheim. 2018. Weakly supervised learning for fake news detection on twitter. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 274–277.

[86] Christopher Hidey and Kathleen R. McKeown. 2018. Persuasive Influence Detection: The Role of Argument Sequencing. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5173–5180. AAAI Press.

[87] Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining, ArgMining@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*, pages 11–21. Association for Computational Linguistics.

[88] Abigail Hodge and Samantha Price. 2020. Celebrity Profiling using Twitter Follower Feeds—Notebook for PAN at CLEF 2020. In *CLEF 2020 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

[89] Yi Huang, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. 2021. Balancing methods for multi-label text classification with long-tailed class distribution. In *EMNLP (1)*, pages 8153–8161. Association for Computational Linguistics.

[90] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. 2013. *An introduction to statistical learning*, volume 112. Springer.

[91] Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, ECML'98, pages 137–142, Berlin, Heidelberg. Springer-Verlag.

[92] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

[93] Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing. Third Edition Draft of January 2023.* Prentice Hall series in artificial intelligence. Prentice Hall.

[94] David Jurgens. 2013. That's what friends are for: Inferring location in online social media platforms based on social relationships. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013*. The AAAI Press.

[95] Jurgita Kapociute-Dzikiene, Andrius Utka, and Ligita Sarkute. 2015. Authorship attribution and author profiling of lithuanian literary texts. In *BSNLP@RANLP*, pages 96–105. INCOMA Ltd. Shoumen, BULGARIA.

[96] Kawaljeet Kaur Kapoor, Kuttimani Tamilmani, Nripendra P Rana, Pushp Patil, Yogesh K Dwivedi, and Sridhar Nerur. 2018. Advances in social media research: Past, present and future. *Information Systems Frontiers*, 20:531–558.

[97] Daniel Kershaw, Matthew Rowe, and Patrick Stacey. 2016. Towards Modelling Language Innovation Acceptance in Online Social Networks. In *Proceedings of the Ninth ACM WSDM*, pages 553–562. ACM.

[98] Jens Kersten, Anna Kruspe, Matti Wiegmann, and Friederike Klan. 2019. Robust Filtering of Crisis-related Tweets. In *16th International Conference on Information Systems for Crisis Response And Management (ISCRAM 2019)*, pages 814–824.

[99] Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Martin Potthast, and Benno Stein. 2020. Overview of the Cross-Domain Authorship Verification Task at PAN 2020. In *Working Notes Papers of the CLEF 2020 Evaluation Labs*, volume 2696 of *CEUR Workshop Proceedings*.

[100] Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Benno Stein, and Martin Potthast. 2021. Overview of the Cross-Domain Authorship Verification Task at PAN 2021. In *Working Notes Papers of the CLEF 2021 Evaluation Labs*, volume 2936 of *CEUR Workshop Proceedings*, pages 1743–1759.

[101] Arunima Khunteta and Pardeep Singh. 2021. Emotion cause extraction—a review of various methods and corpora. In *Proceedings of the 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC)*, pages 314–319. IEEE.

[102] Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*, pages 1746–1751. ACL.

[103] Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu,

Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler C. Murray, Christopher Newell, Smita R Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. 2023. The semantic scholar open data platform. *ArXiv*, abs/2301.10140.

[104] Hannah Rose Kirk, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022. Handling and presenting harmful text in nlp research. *CoRR*, abs/2204.14256.

[105] Konstantin Kobs, Martin Potthast, Matti Wiegmann, Albin Zehe, Benno Stein, and Andreas Hotho. 2020. Towards Predicting the Subscription Status of Twitch.tv Users – ECML-PKDD ChAT Discovery Challenge 2020. In *ECML-PKDD 2020 ChAT Discovery Challenge on Chat Analytics for Twitch*, volume 2661 of *CEUR Workshop Proceedings*.

[106] S. V. Kogilavani, S. Malliga, K. R. Jaiabinaya, M. Malini, and M. Manisha Kokila. 2021. Characterization and mechanical properties of offensive language taxonomy and detection techniques. *Materials Today: Proceedings*.

[107] Boško Koloski, Senja Pollak, and Blaž Škrlj. 2020. Know your Neighbors: Efficient Author Profiling via Follower Tweets—Notebook for PAN at CLEF 2020. In *CLEF 2020 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

[108] Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, 17(4):401–412.

[109] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15):5802–5805.

[110] Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-annotated corpus of hindi-english code-mixed data. In *LREC*. European Language Resources Association (ELRA).

[111] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 957–966. JMLR.org.

[112] Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. *Proceedings of the AAAI Conference on Artificial Intelligence*.

[113] Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. 2010. A text-driven rule-based system for emotion cause detection. In *Proceedings of the 2010 NAACL-HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 45–53.

[114] David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397.

[115] Jialu Li, Esin Durmus, and Claire Cardie. 2020. Exploring the role of argument structure in online debate persuasion. *CoRR*, abs/2010.03538.

[116] Jiwei Li, Alan Ritter, and E. Hovy. 2014. Weakly supervised user profile extraction from twitter. In *Annual Meeting of the Association for Computational Linguistics*, pages 165–174.

[117] Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2023. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. *ArXiv*, abs/2308.12032.

[118] Shen Li, João Graça, and B. Taskar. 2012. Wiki-ly supervised part-of-speech tagging. In *Conference on Empirical Methods in Natural Language Processing*, pages 1389–1398.

[119] Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! numersense: Probing numerical commonsense knowledge of pre-trained language models. In *Conference on Empirical Methods in Natural Language Processing*, pages 6862–6868.

[120] Tatiana Litvinova, Pavel Seredin, Olga Litvinova, and Olga Zagorovskaya. 2017. Differences in type-token ratio and part-of-speech frequencies in male and female russian written texts. In *Proceedings of the Workshop on Stylistic Variation*, pages 69–73. Association for Computational Linguistics.

[121] Fei Liu, F. Weng, Bingqing Wang, and Yang Liu. 2011. Insertion, deletion, or substitution? normalizing text messages without pre-categorization nor supervision. In *Annual Meeting of the Association for Computational Linguistics*, pages 71–76.

[122] Han Liu, Caixia Yuan, and Xiaojie Wang. 2020. Label-wise document pre-training for multi-label text classification. In *Natural Language Processing and Chinese Computing - 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14-18, 2020, Proceedings, Part I*, volume 12430 of *Lecture Notes in Computer Science*, pages 641–653. Springer.

[123] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. 2020. Early-Learning Regularization Prevents Memorization of

Noisy Labels. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

[124] Sheng Liu, Zhihui Zhu, Qing Qu, and Chong You. 2022. Robust Training under Label Noise by Over-parameterization. *ArXiv*, abs/2202.14026.

[125] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv*, abs/1907.11692.

[126] Liane Longpre, Esin Durmus, and Claire Cardie. 2019. Persuasion of the Undecided: Language vs. the Listener. In *Proceedings of the 6th Workshop on Argument Mining*, pages 167–176, Florence, Italy. Association for Computational Linguistics.

[127] University of Michigan, College of Literature, Science, and the Arts LSA. 2023. An Introduction to Content Warnings and Trigger Warnings. `https://sites.lsa.umich.edu/inclusive-teaching-sandbox/wp-content/uploads/sites/853/2021/02/An-Introduction-to-Content-Warnings-and-Trigger-Warnings-Draft.pdf`. Last accessed: May 10, 2023.

[128] Kelvin Luu, Chenhao Tan, and Noah A. Smith. 2019. Measuring Online Debaters' Persuasive Skill from Text over Time. *Trans. Assoc. Comput. Linguistics*, 7:537–550.

[129] Veronica Lynn, Niranjan Balasubramanian, and H. Andrew Schwartz. 2020. Hierarchical Modeling for User Personality Prediction: The Role of Message-Level Attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

[130] Matteo Magnani and Luca Rossi. 2011. The ml-model for multi-layer social networks. In *International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2011, Kaohsiung, Taiwan, 25-27 July 2011*, pages 5–12. IEEE Computer Society.

[131] Aibek Makazhanov and Davood Rafiei. 2013. Predicting political preference of twitter users. *Social Network Analysis and Mining*, 4:1–15.

[132] University of Manchester, Institute of Teaching and Learning Man. 2023. Content Notes for Programmes, Course Units and Specific Activities and Resources. `https://www.staffnet.manchester.ac.uk/umitl/resources/inclusivity/content-notes-in-teaching/`. Last accessed: May 10, 2023.

[133] Adam Marcus and Aditya G. Parameswaran. 2015. Crowdsourced data management: Industry and academic perspectives. *Found. Trends Databases*, 6(1-2):1–161.

[134] Matej Martinc, Blaž Škrlj, and Senja Pollak. 2019. Who is hot and who is not? Profiling celebs on Twitter—Notebook for PAN at CLEF 2019. In *CLEF 2019 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

[135] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444.

[136] Eneldo Loza Mencía and Johannes Fürnkranz. 2008. Efficient pairwise multilabel classification for large-scale problems in the legal domain. In *ECML/PKDD (2)*, volume 5212 of *Lecture Notes in Computer Science*, pages 50–65. Springer.

[137] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR (Workshop Poster)*.

[138] George Mikros. 2013. Authorship Attribution and Gender Identification in Greek Blogs. In *Selected papers of the VIIIth International Conference on Quantitative Linguistics (QUALICO)*, pages 21–32.

[139] Mike D. Mintz, Steven Bills, R. Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Annual Meeting of the Association for Computational Linguistics*, pages 1003–1011.

[140] Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2018. Author profiling for abuse detection. In *Proceedings of the 27th international conference on computational linguistics*, pages 1088–1098.

[141] Rohan Mishra, Pradyumn Prakhar Sinha, Ramit Sawhney, Debanjan Mahata, Puneet Mathur, and Rajiv Ratn Shah. 2019. Snap-batnet: Cascading author profiling and social network graphs for suicide ideation detection on social media. In *Proceedings of the 2019 conference of the North American Chapter of the association for computational linguistics: student research workshop*, pages 147–156.

[142] Salman Mohammed, Nimesh Ghelani, and Jimmy Lin. 2017. Distant supervision for topic classification of tweets in curated streams. *CoRR*, abs/1704.06726.

[143] Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2020. ETHOS: an online hate speech detection dataset. *CoRR*, abs/2006.08328.

[144] Luis Gabriel Moreno-Sandoval, Edwin Puertas, Flor Miriam Plaza-del-Arco, Alexandra Pomares-Quimbaya, Jorge Andres Alvarado-Valencia, and L.Alfonso Ureña-López. 2019. Celebrity Profiling on Twitter using Sociolinguistic Features—Notebook for PAN at CLEF 2019. In *CLEF 2019 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

[145] Alexander A. Morgan, Lynette Hirschman, Marc E. Colosimo, Alexander S. Yeh, and Jeffrey B. Colombe. 2004. Gene name identification and normalization using a model organism database. *J. Biomed. Informatics*, 37(6):396–410.

[146] Fred Morstatter, Liang Wu, Tahora H. Nazer, Kathleen M. Carley, and Huan Liu. 2016. A new approach to bot detection: Striking the balance between precision and recall. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016, San Francisco, CA, USA, August 18-21, 2016*, pages 533–540. IEEE Computer Society.

[147] Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P. Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 462–472. Asian Federation of Natural Language Processing.

[148] Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. 2013. Learning with Noisy Labels. In *Neural Information Processing Systems*.

[149] Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250.

[150] Dong Nguyen, Noah A. Smith, and Carolyn P. Rosé. 2011. Author Age Prediction from Text Using Linear Regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 115–123. ACM.

[151] Andreas Niekler, Magdalena Wolska, Marvin Thiel, Matti Wiegmann, Benno Stein, and Manuel Burghard. 2023. Marco Polo's Travels Revisited: From Motion Event Detection to Optimal Path Computation in 3D Maps. In *Book of Abstracts from the Digital Humanities Conference 2023*, pages 357–359. Alliance of Digital Humanities Organizations.

[152] Curtis G. Northcutt, Tailin Wu, and Isaac L. Chuang. 2017. Learning with Confident Examples: Rank Pruning for Robust Classification with Noisy Labels. *ArXiv*, abs/1705.01936.

[153] University of Nottingham Nott. 2021. Content notes policy, 2021-22. `https://www.nottingham.ac.uk/educational-excellence/documents/content-notes-policy-2122.pdf`. Last accessed: May 10, 2023.

[154] Daniel J. O'Keefe. 2006. Persuasion. In *Encyclopedia of Rhetoric*. Oxford University Press.

[155] Jiaqi Pan, Rishabh Bhardwaj, Wei Lu, Hai Leong Chieu, Xinghao Pan, and Ni Yi Puay. 2019. Twitter homophily: Network based prediction of user's occupation. In *Proceedings of the 57th ACL*, pages 2633–2638. Association for Computational Linguistics.

[156] Hyunji Hayley Park, Yogarshi Vyas, and Kashif Shah. 2022. Efficient classification of long documents using transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 702–709. Association for Computational Linguistics.

[157] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. 2016. Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2233–2241.

[158] Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Audience-modulated variation in online social media. *American Speech*, 90.

[159] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[160] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[161] Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. 2011. Predicting Age and Gender in Online Social Networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, SMUC '11, pages 37–44, New York, NY, USA. ACM.

[162] Björn Pelzer. 2019. Celebrity Profiling with Transfer Learning—Notebook for PAN at CLEF 2019. In *CLEF 2019 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

[163] James W. Pennebaker, Matthias R. Mehl, and Kate G. Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54:547–577.

[164] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

[165] Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. 2023. Efficient benchmarking (of language models). *CoRR*, abs/2308.11696.

[166] Isaac Persing and Vincent Ng. 2017. Lightly-Supervised Modeling of Argument Persuasiveness. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 594–604, Taipei, Taiwan. Asian Federation of Natural Language Processing.

[167] Juraj Petrik and Daniela Chuda. 2019. Twitter feeds profiling with TF-IDF—Notebook for PAN at CLEF 2019. In *CLEF 2019 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

[168] Barbara Plank and Dirk Hovy. 2015. Personality traits on twitter - or - how to get 1, 500 personality tests in a week. In *WASSA@EMNLP*, pages 92–98. The Association for Computer Linguistics.

[169] Martin F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

[170] Martin Potthast, Tim Gollub, Kristof Komlossy, Sebastian Schuster, Matti Wiegmann, Erika Patricia Garces Fernandez, Matthias Hagen, and Benno Stein. 2018. Crowdsourcing a Large Corpus of Clickbait on Twitter. In *27th International Conference on Computational Linguistics (COLING 2018)*, pages 1498–1507. Association for Computational Linguistics.

[171] Martin Potthast, Tim Gollub, Matti Wiegmann, and Benno Stein. 2019. TIRA Integrated Research Architecture. In Nicola Ferro and Carol Peters, editors, *Information Retrieval Evaluation in a Changing World*, volume 41 of *The Information Retrieval Series*, pages 123–160. Springer, Berlin Heidelberg New York.

[172] Daniel Preotiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015. An analysis of the user occupational class through twitter content. In *ACL (1)*, pages 1754–1764. The Association for Computer Linguistics.

[173] Daniel Preotiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle H. Ungar. 2017. Beyond binary labels: Political ideology prediction of twitter users. In *ACL (1)*, pages 729–740. Association for Computational Linguistics.

[174] Daniel Preotiuc-Pietro and Lyle H. Ungar. 2018. User-level race and ethnicity predictors from twitter text. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1534–1545. Association for Computational Linguistics.

[175] Matthew Purver and S. Battersby. 2012. Experimenting with distant supervision for emotion classification. In *Conference of the European Chapter of the Association for Computational Linguistics*, pages 482–491.

[176] Quacquarelli Symonds Limited QS. 2023. QS World University Rankings 2023: Top global universities. `https://www.topuniversities.com/university-rankings/world-university-rankings/2023`. Last accessed: May 25, 2023.

[177] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, C. McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *ArXiv*, abs/2212.04356.

[178] Victor Radivchev, Alex Nikolov, and Alexandrina Lambova. 2019. Celebrity Profiling using TF-IDF, Logistic Regression, and SVM—Notebook for PAN at CLEF 2019. In *CLEF 2019 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

[179] Ricelli Ramos, Georges Neto, Barbara Barbosa Claudino Silva, Danielle Sampaio Monteiro, Ivandré Paraboni, and Rafael Dias. 2018. Building a corpus for personality-dependent natural language understanding and generation. In *LREC*. European Language Resources Association (ELRA).

[180] Francisco Rangel, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd Author Profiling Task at PAN 2015. In *Working Notes Papers of the CLEF 2015 Evaluation Labs*, volume 1391 of *CEUR Workshop Proceedings*.

[181] Francisco Rangel, Anastasia Giachanou, Bilal Ghanem, and Paolo Rosso. 2020. Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter. In *CLEF 2020 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

[182] Francisco Rangel and Paolo Rosso. 2019. Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling. In *CLEF 2019 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

[183] Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. 2014. Overview of the 2nd Author Profiling Task at PAN 2014. In *Working Notes Papers of the CLEF 2014 Evaluation Labs*, volume 1180 of *Lecture Notes in Computer Science*.

[184] Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. 2013. Overview of the Author Profiling Task at PAN 2013. In *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain*. CEUR-WS.org.

[185] Francisco Rangel, Paolo Rosso, Martin Potthast, and Benno Stein. 2017. Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In *Working Notes Papers of the CLEF 2017 Evaluation Labs*, volume 1866 of *CEUR Workshop Proceedings*.

[186] Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. 2016. Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations. In *Working Notes Papers of the CLEF 2016 Evaluation Labs*, volume 1609 of *CEUR Workshop Proceedings*.

[187] Francisco Rangel, Paolo Rosso, Manuel Montes y Gómez, Martin Potthast, and Benno Stein. 2018. Overview of the 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter. In *Working Notes Papers of the CLEF 2018 Evaluation Labs*, volume 2125 of *CEUR Workshop Proceedings*.

[188] Alexander J. Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. In *NIPS*, pages 3567–3575.

[189] University of Reading Read. 2023. Guide to policy and procedures for teaching and learning; Guidance on content warnings on course content ('trigger' warnings). `https://www.reading.ac.uk/cqsd/-/media/project/functions/cqsd/documents/qap/trigger-warnings.pdf`. Last accessed: May 10, 2023.

[190] Wendi Ren, Yinghao Li, Hanting Su, David Kartchner, Cassie S. Mitchell, and Chao Zhang. 2020. Denoising Multi-Source Weak Supervision for Neural Text Classification. *ArXiv*, abs/2010.04582.

[191] Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534.

[192] Marian-Andrei Rizoiu, Lexing Xie, Scott Sanner, Manuel Cebrián, Honglin Yu, and Pascal Van Hentenryck. 2017. Expecting to be HIP: hawkes intensity processes for social media popularity. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 735–744. ACM.

[193] Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. 2012. Supervised text-based geolocation using language models on an adaptive grid. In *Conference on Empirical Methods in Natural Language Processing*, pages 1500–1510.

[194] David Rolnick, Andreas Veit, Serge J. Belongie, and Nir Shavit. 2017. Deep Learning is Robust to Massive Label Noise. *ArXiv*, abs/1705.10694.

[195] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685. IEEE.

[196] Sara Rosenthal and Kathleen R. McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media

generations. In *ACL*, pages 763–772. The Association for Computer Linguistics.

[197] Russel Group. 2023. Russel Group: Our universities. `https://russellgroup.ac.uk/about/our-universities`. Last accessed: May 25, 2023.

[198] Derek Ruths and Jürgen Pfeffer. 2014. Social media for large studies of behavior. *Science*, 346(6213):1063–1064.

[199] A. Sadilek, Henry A. Kautz, and Jeffrey P. Bigham. 2012. Finding your friends and following them to where you are. In *Web Search and Data Mining*, pages 723–732.

[200] Umitcan Sahin, Izzet Emre Kucukkaya, and Cagri Toraman. 2023. ARC-NLP at PAN 2023: Hierarchical Long Text Classification for Trigger Detection . In *Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum*. CEUR-WS.org.

[201] Eric Arazo Sanchez, Diego Ortego, Paul Albert, Noel E. O?Connor, and Kevin McGuinness. 2019. Unsupervised label noise modeling and loss correction. *ArXiv*, abs/1904.11238.

[202] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. 2006. Effects of Age and Gender on Blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 199–205. AAAI.

[203] Anna Schmidt and Michael Wiegand. 2019. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the 5th International Workshop on Natural Language Processing for Social Media*, pages 1–10.

[204] H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. In *PLoS ONE*, page 8(9): e73791.

[205] Hosahalli Lakshmaiah Shashirekha, Asha Hegde, and Fazlourrahman Balouchzahi. 2023. Trigger Detection in Social Media Text . In *Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum*. CEUR-WS.org.

[206] Anshumali Shrivastava and Ping Li. 2014. In defense of minhash over simhash. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014*, volume 33 of *JMLR Workshop and Conference Proceedings*, pages 886–894. JMLR.org.

[207] Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. In *NIPS*, pages 1297–1304.

[208] Anders Søgaard. 2013. *Semi-Supervised Learning and Domain Adaptation in Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

[209] Thamar Solorio, Mahsa Shafaei, Christos Smailis, Brad J Bushman, Douglas A Gentile, Erica Scharrer, Laura Stockdale, and Ioannis Kakadiaris. 2021. White paper — objectionable online content: What is harmful, to whom, and why. *arXiv preprint arXiv:2104.03903*.

[210] Manuka Stratta, Julia Park, and Cooper deNicola. 2020. Automated content warnings for sensitive posts. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, CHI 2020, Honolulu, HI, USA, April 25-30, 2020*, pages 1–8. ACM.

[211] Yunsen Su, Yong Han, and Haoliang Qi. 2023. Siamese Networks in Trigger Detection task . In *Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum*. CEUR-WS.org.

[212] David Suendermann, Keelan Evanini, Jackson Liscombe, Phillip Hunter, Krishna Dayanidhi, and Roberto Pieraccini. 2009. From rule-based to statistical grammars: Continuous improvement of large-scale spoken dialog systems. In *ICASSP*, pages 4713–4716. IEEE.

[213] Nadine Tamburrini, Marco Cinnirella, Vincent Jansen, and John Bryden. 2015. Twitter users change word usage according to conversation-partner social identity. *Social Networks*, 40:84?89.

[214] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 613–624, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

[215] Duyu Tang, Furu Wei, Nan Yang, M. Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Annual Meeting of the Association for Computational Linguistics*, pages 1555–1565.

[216] Times Higher Education THE. 2023. World University Rankings 2023. `https://www.timeshighereducation.com/world-university-rankings/2023/world-ranking`. Last accessed: May 25, 2023.

[217] OTW The Organization for Transformative Works. 2023. Wrangling guidelines.

[218] Edward P. Tighe and Charibeth K. Cheng. 2018. Modeling personality traits of filipino twitter users. In *PEOPLES@NAACL-HTL*, pages 112–122. Association for Computational Linguistics.

[219] Stanford Teaching and Learning Hub TLHUB. 2022. Writing Content Notices for Sensitive Content. `https://www.reading.ac.uk/cqsd/-/media/project/functions/cqsd/documents/qap/trigger-warnings.pdf`. Last accessed: May 10, 2023.

[220] Oren Tsur and Ari Rappoport. 2012. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the Fifth International Conference on Web Search and Web Data Mining, WSDM 2012, Seattle, WA, USA, February 8-12, 2012*, pages 643–652. ACM.

[221] Twitter. 2018. FAQ: About verified accounts. `https://web.archive.org/web/20180701010748/https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts`, accessed 15.11.2018.

[222] Ben Verhoeven and Walter Daelemans. 2014. Clips stylometry investigation (CSI) corpus: A dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, pages 3081–3085. European Language Resources Association (ELRA).

[223] Ben Verhoeven, Walter Daelemans, and Barbara Plank. 2016. Twisty: A multilingual twitter stylometry corpus for gender and personality profiling. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.* European Language Resources Association (ELRA).

[224] Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *ArXiv*, abs/2306.07899.

[225] Svitlana Volkova and Yoram Bachrach. 2015. On predicting sociodemographic traits and emotions from communications in social networks and their implications to online self-disclosure. *Cyberpsy., Behavior, and Soc. Networking*, 18(12):726–736.

[226] Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, Dong Yu, D. Roth, and David A. McAllester. 2019. Evidence sentence extraction for machine reading comprehension. *ArXiv*, abs/1902.08852.

[227] Xiang Wang, Xiangnan He, Liqiang Nie, and Tat-Seng Chua. 2017. Item silk road: Recommending items from information domains to social users. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 185–194. ACM.

[228] Xiaolong Wang, Furu Wei, Xiaohua Liu, M. Zhou, and Ming Zhang. 2011. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *International Conference on Information and Knowledge Management*, pages 1031–1040.

[229] Yizhi Wang, Yuwan Dai, Hao Li, and Lili Song. 2021. Social media and attitude change: Information booming promote or resist persuasion? *Frontiers in Psychology*, 12.

[230] Yuan Wang, Yang Xiao, Chao Ma, and Zhen Xiao. 2016. Improving users' demographic prediction via the videos they talk about. In *EMNLP*, pages 1359–1368. The Association for Computational Linguistics.

[231] Zijian Wang, Scott A. Hale, David Ifeoluwa Adelani, Przemyslaw A. Grabowicz, Timo Hartmann, Fabian Flöck, and David Jurgens. 2019. Demographic inference and representative population estimates from multilingual social media data. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2056–2067. ACM.

[232] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84.

[233] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

[234] Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*. The Association for Computer Linguistics.

[235] Matti Wiegmann, Khalid Al-Khatib, Vishal Khanna, and Benno Stein. 2022. Analyzing Persuasion Strategies of Debaters on Social Media. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6897–6905. International Committee on Computational Linguistics.

[236] Matti Wiegmann, Jens Kersten, Friederike Klan, Martin Potthast, and Benno Stein. 2020. Analysis of Detection Models for Disaster-Related Tweets. In *17th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2020, May 2020*, pages 872–880. ISCRAM Digital Library.

[237] Matti Wiegmann, Jens Kersten, Hansi Senaratne, Martin Potthast, Friederike Klan, and Benno Stein. 2021. Opportunities and Risks of Disaster Data from Social Media: A Systematic Review of Incident Information. *Natural Hazards and Earth System Sciences*, 21(5):1431–1444.

[238] Matti Wiegmann, Jan Heinrich Reimer, Maximilian Ernst, Martin Potthast, Matthias Hagen, and Benno Stein. 2024. A Mastodon Corpus to Evaluate Federated Microblog Search. In *Proceedings of the First International Workshop on Open Web Search (WOWS 2024)*, volume 3689, pages 37–49. CEUR Workshop Proceedings.

[239] Matti Wiegmann, Benno Stein, and Martin Potthast. 2019. Celebrity Profiling. In *57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 2611–2618. Association for Computational Linguistics.

[240] Matti Wiegmann, Benno Stein, and Martin Potthast. 2019. Overview of the Celebrity Profiling Task at PAN 2019. In *Working Notes Papers of the CLEF 2019 Evaluation Labs*, volume 2380 of *CEUR Workshop Proceedings*.

[241] Matti Wiegmann, Benno Stein, and Martin Potthast. 2020. Overview of the Celebrity Profiling Task at PAN 2020. In *Working Notes Papers of the CLEF 2020 Evaluation Labs*, volume 2696 of *CEUR Workshop Proceedings*.

[242] Matti Wiegmann, Benno Stein, and Martin Potthast. 2024. De-Noising Document Classification Benchmarks via Prompt-based Rank Pruning: A Case Study. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 15th International Conference of the CLEF Association (CLEF 2024)*, volume 14958 of *Lecture Notes in Computer Science*, pages 172–178, Berlin Heidelberg New York. Springer.

[243] Matti Wiegmann, Michael Völske, Martin Potthast, and Benno Stein. 2022. Language Models as Context-sensitive Word Search Engines. In *Proceedings of the 1st Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, pages 39–45. Association for Computational Linguistics.

[244] Matti Wiegmann, Magdalena Wolska, Martin Potthast, and Benno Stein. 2023. Overview of the Trigger Detection Task at PAN 2023. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023)*, volume 3497 of *CEUR Workshop Proceedings*, pages 2523–2536.

[245] Matti Wiegmann, Magdalena Wolska, Christopher Schröder, Ole Borchardt, Benno Stein, and Martin Potthast. 2023. Trigger Warning Assignment as a Multi-Label Document Classification Problem. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12113–12134, Toronto, Canada. Association for Computational Linguistics.

[246] Wikipedia. 2018. notability guidelines for people. `https://en.wikipedia.org/wiki/Wikipedia:Notability_(people)`, accessed 15.11.2018.

[247] Magdalena Wolska, Matti Wiegmann, Christopher Schröder, Ole Borchardt, Benno Stein, and Martin Potthast. 2023. Trigger Warnings: Bootstrapping a Violence Detector for Fan Fiction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.

[248] Liang Wu and Huan Liu. 2018. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, pages 637–645. ACM.

[249] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1391–1399, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

[250] Guang Xiang, Bin Fan, Ling Wang, Jason I. Hong, and C. Rosé. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. *Proceedings of the 21st ACM international conference on Information and knowledge management*.

[251] Lin Xiao, Xin Huang, Boli Chen, and Liping Jing. 2019. Label-specific document representation for multi-label text classification. In *EMNLP/IJCNLP (1)*, pages 466–475. Association for Computational Linguistics.

[252] Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. 2020. Scalable and generalizable social bot detection through data selection. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 1096–1103. AAAI Press.

[253] Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. SGM: sequence generation model for multi-label classification. In *COLING*, pages 3915–3926. Association for Computational Linguistics.

[254] Zhichao Yang, Shufan Wang, Bhanu Pratap Singh Rawat, Avijit Mitra, and Hong Yu. 2022. Knowledge injected prompt based fine-tuning for multi-label few-shot ICD coding. *CoRR*, abs/2210.03304.

[255] A. H. Yazdavar, Hussein S. Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, K. Thirunarayan, Jyotishman Pathak, and A. Sheth. 2017. Semi-supervised approach to monitoring clinical depressive symptoms in social media. *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*.

[256] Jieyu Zhang, Cheng-Yu Hsieh, Yue Yu, Chao Zhang, and Alexander Ratner. 2022. A survey on programmatic weak supervision. *CoRR*, abs/2202.05433.

[257] Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational Flow in Oxford-style Debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 136–141. Association for Computational Linguistics.

[258] Zizhao Zhang, Han Zhang, Sercan Ö. Arik, Honglak Lee, and Tomas Pfister. 2019. Distilling Effective Supervision From Severe Label Noise. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9291–9300.

[259] Zhi-Hua Zhou. 2017. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53.

[260] D. Zhu, Michael A. Hedderich, Fangzhou Zhai, David Ifeoluwa Adelani, and Dietrich Klakow. 2022. Is bert robust to label noise? a study on learning with noisy labels in text classification. *ArXiv*, abs/2204.09371.