Enhancing Data Quality through Simple De-duplication: Navigating Responsible Computational Social Science Research

Yida Mu **Mali Jin** **Xingyi Song** **Nikolaos Aletras**
School of Computer Science, The University of Sheffield
{y.mu, m.jin, x.song, n.aletras}@sheffield.ac.uk

Abstract

Research in natural language processing (NLP) for Computational Social Science (CSS) heavily relies on data from social media platforms. This data plays a crucial role in the development of models for analysing socio-linguistic phenomena within online communities. In this work, we conduct an in-depth examination of 20 datasets extensively used in NLP for CSS to comprehensively examine data quality. Our analysis reveals that social media datasets exhibit varying levels of data duplication. Consequently, this gives rise to challenges like label inconsistencies and data leakage, compromising the reliability of models. Our findings also suggest that data duplication has an impact on the current claims of state-of-the-art performance, potentially leading to an overestimation of model effectiveness in real-world scenarios. Finally, we propose new protocols and best practices for improving dataset development from social media data and its usage.

1 Introduction

Research in natural language processing (NLP) for Computational Social Science (CSS) aims to analyze social behavior and sociolinguistic phenomena with computational methods on a large scale (Plank and Hovy, 2015; Guntuku et al., 2019). This relies upon using vast amounts of user-generated content in social media such as Twitter/X and Weibo (Edelmann et al., 2020).

Social media data exhibits distinctive characteristics such as rapid and continual topic evolution (Saha and Sindhwani, 2012; Yuan et al., 2013). Prior research has focused on introducing and developing novel resources including new tasks and datasets (Founta et al., 2018; Lazer et al., 2020; Hofman et al., 2021). Social media posts often contain a significant amount of near-duplicate or even identical content. For example, during social emergencies (e.g., COVID-19 pandemic), users tend to repeatedly post about the same topics (e.g., COVID-19 vaccination), resulting in a proliferation of near-duplicate content over a very short period of time (Ferrara, 2020; Zhang et al., 2022b). Furthermore, the presence of social bots (e.g., automatically generated posts from third-party applications) amplifies the rapid generation of near-duplicate content on these platforms, e.g., in political campaigns (Bessi and Ferrara, 2016; Stella et al., 2018).

Data quality is paramount for reliable and effective model performance. It includes various dimensions such as accuracy, completeness and consistency (Vidgen and Derczynski, 2020; Budach et al., 2022). Ensuring high data quality usually involves careful data management and processing such as data validation and cleaning (Denny and Spirling, 2018; Breck et al., 2019; Gröger, 2021; Li et al., 2021a). Previous works have examined and deduplication of samples in image datasets (Alam et al., 2017), generic real-word datasets (e.g., movie, restaurant) (Li et al., 2021a) and language modeling datasets (e.g., Wikipedia) (Lee et al., 2022). However, the quality of existing CSS datasets in NLP has been under-explored.

The goal of this paper is to conduct a large scale meta-analysis of current NLP for CSS datasets by considering the potential noise caused by duplicated text samples. To this end, we perform an in-depth re-examination of 20 social media datasets across various tasks such as offensive language detection and misinformation detection.[1] Our main contributions are as follows:

- Our systematic analysis shows that most of the examined social media datasets contain noise (e.g., duplicate and near-duplicate samples) despite the data cleaning process claimed by the developers (see examples in Table 1).
- We explore the impact of data duplication

[1]Data and code: https://github.com/YIDAMU/Clean_CSS

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 12477–12492
November 12-16, 2024

Tweet_ID	Samples	Label	Consequence
129839*	@USER @USER donald trump's lessons for republicans: consequences for lying...	Neutral	**Inconsistent Labeling**
129840*	@USER @USER donald trump's lessons for republicans: consequences for lying...	Against	
132605*	@USER FIGHT AGAINST TYRANNY!!! The coronavirus vaccine is a NEW technology RNA vaccine. It LITERALLY changes your DNA!!!! [Emojis]	Anti. Vaxx	**Label Leakage**
132607*	@USER FIGHT AGAINST TYRANNY!!! The coronavirus vaccine is a NEW technology RNA vaccine. It LITERALLY changes your DNA!!!! [Emojis]	Anti. Vaxx	
131120* (Dupli.) †	@USER trump	Neutral	**Label Leakage**
407156*	r.i.p to the driver who died with paul walker that no one cares about because he wasn't famous omg:(	True Rumor	**Label Leakage**
407164*	r.i.p to the driver that died with paul walker that no one cares about because he wasn't famous.	True Rumor	

Table 1: Examples of duplicate or near-duplicate samples in real-world social datasets (Ma et al., 2017; Cotfas et al., 2021; Kawintiranon and Singh, 2021) as well as their potential consequence. † denotes duplicate tweet ids in the dataset.

on model performance in various CSS tasks. We observe an overestimation of model performance in cases where duplicate or near-duplicate samples remain unfiltered. We find that the presence of duplicate samples results in label inconsistencies and data leakage, potentially causing unreliable model predictions.

- We propose a new data pre-processing protocol for a more responsible and effective use of social media resources and advocate for a 'minor revision' of the existing author checklist from major CSS communities.

Note that this study does not aim to criticize dataset creators. Instead, it acknowledges the crucial role these datasets play in advancing the field of CSS and seeks to provide constructive recommendations for improving responsible research and practices.

2 Related Work

2.1 Data Quality and Model Performance

Data quality is paramount in NLP tasks as the performance and reliability of models heavily depend on the quality of the training data. Several factors contribute to data quality such as accuracy, persistence, completeness, consistency and relevance (Zubiaga, 2018; Assenmacher et al., 2020; Koch et al., 2021; Budach et al., 2022). Budach et al. (2022) investigated the impact of different data quality dimensions on the performance of various machine learning algorithms covering classification, regression and clustering tasks. Similarly, Barry et al. (2023) explored the relationship between the quality of the training data and the fairness of model outputs on image classification across a range of algorithms. Furthermore, Li et al. (2021a) examined the impact of improving data quality on classification algorithm performance, resulting in the CleanML benchmark. Their work highlighted the importance of data cleaning methods such as addressing missing values and correcting mislabels in enhancing classifier predictions.

2.2 Data Preprocessing

Ensuring data quality in NLP often involves rigorous data preprocessing. Dealing with social media datasets often comes with unique challenges due to their unstructured nature and noisy text (Symeonidis et al., 2018).

The majority of CSS research follows a standard pre-processing pipeline for social media content (Nguyen et al., 2020; Antypas et al., 2023). Social media text often contains user mentions, URLs, hashtags, emojis, and other non-standard symbols. Pre-processing steps typically involve removing or replacing these tokens with generic placeholders. Additionally, social media language is informal, featuring slang, abbreviations, and misspellings, necessitating text normalization techniques such as lowercase conversion and spelling correction (Baldwin and Li, 2015; Naseem et al., 2021).

Previous studies have investigated the impact of various pre-processing strategies on model performance in different CSS downstream tasks such as sentiment analysis (Krouska et al., 2016; Jianqiang and Xiaolin, 2017; Mahilraj et al., 2020) and opinion mining (Dos Santos and Ladeira, 2014; Gull et al., 2016). For example, Symeonidis et al. (2018) examined up to 16 different pre-processing strategies for Twitter data and found that techniques such as lemmatization and removing numbers in

the text can enhance the predictive performance of transformer-based models in sentiment analysis.

2.3 Impact of Data Duplication

Existing language modeling datasets contain many duplicate and near-duplicate samples due to overlapping sources in the training corpus, which has emerged as a significant concern in recent research. Studies have demonstrated that large language models are vulnerable to privacy attacks due to the presence of duplicate sequences in commonly used training datasets (Kandpal et al., 2022). These sequences, when present multiple times in the training data, are regenerated at much higher frequencies by trained models, posing privacy risks. Furthermore, data duplication in language modeling leads to verbatim text duplication in model outputs (Lee et al., 2022). Efforts to mitigate this issue have focused on deduplicating training datasets, resulting in models emitting memorized text less frequently and requiring fewer training steps for comparable or improved performance (Lee et al., 2022).

In this study, we focus on examining and mitigating the potential issue of noise data (i.e., duplicate data) within existing CSS datasets, which has often been overlooked. To our knowledge, a thorough re-evaluation of of the quality of social media datasets has not been extensively conducted.

3 Tasks & Datasets

We evaluate a comprehensive collection of datasets that adequately cover a broad range of prevalent tasks in CSS. Based on recent publications in NLP and CSS venues (e.g., *ACL and ICWSM), we choose datasets from four main CSS tasks following previous work (Barbieri et al., 2020; Ziems et al., 2023; Antypas et al., 2023; Mu et al., 2024): (i) Offensive Language Detection, (ii) Misinformation Detection, (iii) Speech Act Detection & Sentiment Analysis, and (iv) Stance Detection.

3.1 Offensive Language Detection

Offensive language detection refers to the process of automatically identifying content that contains hate speech, harassment, profanity, or other forms of inappropriate language towards individuals, groups, or events (Chen et al., 2012; Davidson et al., 2017). It is particularly relevant for content moderation in online platforms and social media (Nobata et al., 2016).

For this task, we opt for the following datasets: `WASEEM` (Waseem and Hovy, 2016), `TBO` (Zampieri et al., 2023), `OLID` (Davidson et al., 2017), `FOUNTA` (Founta et al., 2018) and `HateEval'19` (Garibo i Orts, 2019).

3.2 Misinformation Detection

Misinformation detection in social media involves the use of algorithms to analyze and identify false or misleading information (e.g., fake news and false rumors) generated or diffused by end users (Shu et al., 2017; Zubiaga et al., 2018). These approaches typically rely on linguistic patterns (e.g., hand-craft features), source credibility (e.g., unreliable news sources), and contextual information (e.g., propagation network) to differentiate between true information and misinformation (Rashkin et al., 2017; Tian et al., 2022).

We conduct evaluation on five popular datasets covering different languages and social media platforms: `Twitter 15`, `Twitter 16` (Ma et al., 2017), `PHEME` (Zubiaga et al., 2016), `Weibo 16` (Ma et al., 2016), `Weibo 20` (Rao et al., 2021).

3.3 Speech Act Detection & Sentiment Analysis

Speech act detection and sentiment analysis deal with the detection and analysis of actions (e.g., requesting and complaining) and affecting content within texts. In recent years, there has been growing attention to automatically identifying speech acts and sentiment analysis in social media (Preoţiuc-Pietro et al., 2019; Farha et al., 2022; Zhang et al., 2022a). These two tasks are closely related as detecting one usually involves the other (Saha et al., 2021).

We choose the following speech acts and sentiment analysis datasets: `Complaint` (Preoţiuc-Pietro et al., 2019), `SemEval-2022 Task 6 Sarcasm` (Farha et al., 2022), `Bragging` (Jin et al., 2022), `Parody` (Maronikolakis et al., 2020) and `SemEval-2017 Task 4 Sentiment` (Rosenthal et al., 2019).

3.4 Stance Detection

Stance detection involves using computational approaches to automatically identify and classify a person's or a group's perspective or attitude towards a specific target, such as events (e.g., COVID-19 vaccination) or individuals (e.g., politicians) (Küçük and Can, 2020; Mu et al., 2023b).

Domain / Dataset	Source / Language	Mean Tokens	Self-claimed Deduplication	# of Post	# of Distinct / Ratio%	# of Distinct after Pre-proc. / Ratio %	# of Distinct after Removing Near-Dupli. / Ratio%
Offensive Language Analysis							
WASEEM	Twitter/En	15	✗	16,909	16,851 / 99.7%	16,568 / 98.0%	13,364 / 79.0%
TBO	Twitter/En	17	✓	4,000	3,998 / 99.9%	3,998 / 99.9%	3,946 / 98.7%
OLID	Twitter/En	22	✗	14,100	14,052 / 99.7%	14,031 / 99.5%	1,1509 / 81.6%
FOUNTA	Twitter/En	17	✗	99,996	91,940 / 91.9%	87,291 / 87.3%	88,263 / 81.5%
HatEval'19	Twitter/En	22	✗	19,600	19,342 / 98.7%	19,267 / 98.3%	17,851 / 91.0%
Misinformation Detection							
Twitter 15	Twitter/En	15	✗	1,490	1,428 / 95.8%	1,428 / 95.8%	1,345 / 90.2%
Twitter 16	Twitter/En	15	✗	818	761 / 93.0%	761 / 93.0%	740 / 90.5%
PHEME	Twitter/En	16	✗	5,802	5,789 / 99.8%	5,694 / 98.1%	5,236 / 90.2%
Weibo	Weibo/Zh	99	✗	4,664	4,516 / 96.8%	4,501 / 96.5%	3,322 / 71.2%
STANKER	Weibo/Zh	71	✗	6,068	6,040 / 99.5%	6,006 / 99.0%	4,703 77.5%
Speech Act & Sentiment Analysis							
Complaint	Twitter/En	15	✓	3,449	3,449 / 100.0%	3,408 / 98.8%	2,846 / 82.5%
Sarcasm	Twitter/En	18	✗	4,868	4,851 / 99.7%	4,849 / 99.6%	4,442 / 91.2%
Bragging	Twitter/En	22	✗	6,696	6,643 / 99.2%	6,636 / 99.1%	5,979 / 89.2%
Parody	Twitter/En	29	✗	46,622	46,587 / 99.9%	46,024 / 98.7%	43,591 / 93.5%
Sentiment	Twitter/En	18	✗	59,899	59,870 / 99.9%	59,836 / 99.9%	58,536 / 97.7%
Stance Detection							
CovidVaxx	Twitter/En	25	✗	2,792	2,787 / 99.8%	2,740 / 98.1%	2,567 / 91.9%
RumorEval	Twitter/En	33	✗	5,568	5,467 / 98.2%	5,467 / 98.2%	4,284 / 76.9%
US-Election	Twitter/En	25	✓	2,500	2,498 / 99.9%	2,498 / 99.9%	2,397 / 95.9%
P-Stance	Twitter/En	30	✓	21,574	21,571 / 99.9%	21,571 / 99.9%	21,551 / 99.8%
SemEval'16	Twitter/En	17	✓	4,063	4,063 / 100.0%	4,048 / 99.6%	3,926 / 96.6%

Table 2: Dataset Specifications. Cells in light grey indicate no or minor reduction (i.e., less than 0.1%) from duplicate samples. Note that some social media datasets have been pre-processed (e.g., by replacing @USER and URL with special tokens) before being publicly available such as Twitter 15 and Twitter 16. This results in the same values for *# of Distinct* and *# of Distinct Posts after Pre-proc.*.

We choose five datasets for the stance detection task including COVID-19 Vaccine Stance (Cotfas et al., 2021), SemEval-2019 Task 7 Rumor Stance (Gorrell et al., 2019), US-Election (Kawintiranon and Singh, 2021), P-Stance (Li et al., 2021b), and Semeval-2016 Task 6 (Mohammad et al., 2016).

3.5 Criteria for Dataset Selection

For each domain, we select five representative datasets based on criteria that include: (i) popularity, as indicated by citation counts; (ii) shared tasks from SemEval[2] and (iii) newly developed datasets (see Table 2 for tasks and related datasets).

For example, WASEEM (Waseem and Hovy, 2016) is one of the most popular and earliest dataset for studying online hate speech. Similarly, Twitter 15 & 16 (Ma et al., 2017) and PHEME (Zubiaga et al., 2016) have been widely used in recent work on computational rumor detection, as shown in (Mu et al., 2023a). Moreover, we consider SemEval datasets as they are generally of high interest to the NLP/CSS community. These shared tasks often release leaderboards for ranking participants, underscoring the importance of using a clean dataset. We also use some newly developed datasets that cover specific linguistic phenomena, such as Bragging (Jin et al., 2022), which is related to computational pragmatics.

[2]International Workshop on Semantic Evaluation: https://semeval.github.io/

4 Examining Dataset Quality

We focus on examining duplicate or near-duplicate samples in datasets to assess data quality.

4.1 Dataset Specifications

To assess the duplication and near-duplication quantity, we report the following specifications:

- **Self-claimed deduplication.** We report this feature by manually reviewing the source paper of the dataset. The purpose of this is to understand whether the original authors have performed data preprocessing and how data has been preprocessed. The label '✓' indicates that the developers have clearly mentioned their implementation of deduplicating in their datasets.

- **Number of posts in the datasets.** We present

the number of samples in the original datasets which were obtained either from authors or through the links provided in source papers.[3]

- **Number of distinct posts.** We display the number of *unique posts* in the original datasets. In this step, we only filter out samples that contain the same content.
- **Number of distinct posts after removing duplicates.** We present the number of distinct posts after preprocessing, i.e., replacing @USER and URL tokens with unified tokens.
- **Number of distinct posts after removing near-duplicates.** We further employ the Levenshtein distance (Levenshtein et al., 1966) to calculate the similarity between each sample and filter out near-duplicate samples after replacing @USER and URL tokens (threshold = 20). Note that this set of data excludes both duplicate and near-duplicate samples.

4.2 Results and Discussion

Table 2 presents the above features as well as basic information including source platforms, language and number of average tokens (word and character level for English and Chinese respectively). Generally, we observe that most existing social media datasets (18 out of 20 in total) contain duplicate samples. Replacing special tokens reveals additional duplicate samples in the majority of the datasets (17 out of 20 in total). Furthermore, we notice that datasets with high duplicate rates are usually developed through a keyword-based sampling method.

Additionally, we note that only a small fraction of papers claim that they have performed a deduplication process, e.g., `TBO` (Zampieri et al., 2023) and Complaint (Preoţiuc-Pietro et al., 2019) (see column 4). This suggests that many developers of social media datasets tend to neglect the data deduplication process. In addition to duplicated samples, we also observe a substantial number of near-duplicate samples in each dataset (see the last column).

5 Impact of Duplicate and Near-duplicate Samples

We conduct comparative experiments regarding three potential issues (data leakage, model rankings and inconsistent labels) on the selected social media datasets to verify the impact of those duplicate and near-duplicate posts on classification performance.

5.1 Data Leakage

Data leakage occurs because the model is evaluated on data it has already seen during training. When test data overlaps with training data (e.g., the dataset contains duplicates), the performance of the model is likely to be overestimated. To investigate the impact of data leakage on model performance, we compare the predictive results within a dataset before and after deduplication.

We use BERTweet (Nguyen et al., 2020) for all English datasets and Bert-base-Chinese (Devlin et al., 2019) for two Chinese datasets. For all datasets, we first replace user mentions and URLs with special tokens. Then we split each dataset into a training (80%) (*Train Set Original*) and a test set (20%). We construct *Train Set w/o Duplicates* by removing all posts in the training set that are identical to those in the test set. Similarly, we build *Train Set w/o Near-duplicate* using the same approach, removing near-duplicate posts. This deduplication process aims to prevent label leakage during model training. It is important to note that the test set remains unchanged for fair comparison. Detailed experimental setup is presented in Appendix A.

With the widespread use of large language models (LLMs), we aim to compare their predictive results with traditional pre-trained models that require training data (training on duplicates potentially leads to data leakage). To facilitate this, we employ the two most recent LLMs, GPT-4o[4] and LLaMA 3-8B-Instruct[5] for zero-shot classification. For LLMs, we use the same test set as for BERT-style models. Note that we only evaluate LLMs on the task of `Complaint`, `Bragging`, `Sarcasm`, `CovidVaxx` and `US-Election` since the sizes of other datasets are too large (see Table 2). Example prompts are provided in Appendix D.

5.2 Model Rankings

The prevalence of duplicates and near-duplicates in these datasets prompts us to investigate the consistency of model rankings before and after deduplication. Given the extensive workload of reproducing

[3] The size of some datasets varies from the original one due to recollection.

[4] https://platform.openai.com/docs/models/gpt-4o

[5] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

Dataset	Original F1	Original Acc.	w/o Duplicates F1	w/o Duplicates Acc.	w/o Near-Duplicates F1	w/o Near-Duplicates Acc.
Offensive Language Detection						
WASEEM	84.2±0.4	86.2±0.3	83.4±0.4	85.7±0.3	83.0±0.2	85.3±0.3
TBO	69.4±1.7	75.0±0.7	69.4±1.7	75.0±0.7	69.2±1.1	74.7±0.3
OLID	76.4±0.5	79.9±0.1	77.1±0.4	80.1±0.4	76.8±0.6	79.0±0.9
FOUNTA	85.9±0.0	86.0±0.0	85.7±0.2	85.9±0.1	85.9±0.1	86.0±0.1
HateEval'19	78.2±1.0	78.7±0.6	80.4±0.3	80.8±0.1	80.2±0.3	80.8±0.2
Misinformation Detection						
Twitter'15	60.4±2.2	61.0±1.7	58.4±0.5*	58.7±0.3*	57.1±3.2*	57.7±2.4*
Twitter'16	62.6±3.5	63.4±3.1	51.9±4.0*	55.7±2.5*	43.0±2.2*	50.2±1.9*
PHEME	84.7±0.7	86.2±0.4	84.1±0.4*	85.6±0.4*	83.8±0.2*	85.5±0.0*
Weibo	91.4±0.5	91.4±0.4	90.9±0.4*	90.9±0.4*	91.0±0.2*	91.0±0.2*
STANKER	92.2±0.2	92.2±0.2	91.7±0.2	91.7±0.2	92.6±0.2	92.6±0.2
Speech Act & Sentiment Analysis						
Complaint	88.9±1.3	89.8±1.3	89.5±1.3	90.3±1.2	89.2±0.3	90.0±0.2
Sarcasm	64.7±7.6	81.5±1.0	61.4±12.4	80.8±1.9	69.8±0.8	82.2±0.7
Bragging	77.8±0.7	91.4±0.5	77.2±0.5	91.1±0.3	77.8±0.7	91.3±0.7
Parody	-	-	-	-	-	-
Sentiment	74.7±0.2	75.0±0.3	74.5±0.2	75.0±0.4	74.5±0.3	74.9±0.3
Stance Detection						
CovidVaxx	80.7±0.4	80.7±0.5	79.2±0.9*	79.2±0.9*	79.0±0.8*	79.1±0.8*
RumorEval	48.6±1.2	72.7±0.5	48.6±0.4	72.8±0.5	46.4±0.9	72.4±1.1
US-Election	53.4±0.7	56.8±0.9	52.4±0.8	56.5±0.8	51.2±1.7	55.3±0.5
P-Stance	80.1±0.7	80.2±0.7	80.8±0.4	80.9±0.4	80.8±0.2	80.9±0.2
SemEval'16	65.9±0.3	69.5±0.5	65.7±0.6	69.1±0.4	63.1±1.1*	67.4±1.2*

Table 3: Model performance across datasets and duplicate rates. We were unable to conduct experiments on the Parody dataset due to the incomplete dataset we obtained. * denotes the statistic significance (t-test, $p < .05$) between original and w/o settings. We run all BERT-style models three times with different random seeds and then report the average F1 measure and accuracy.

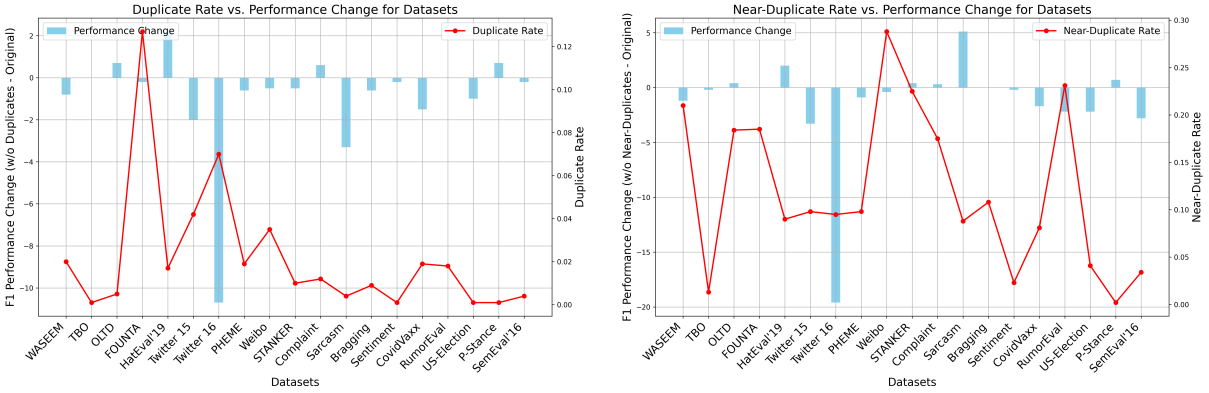

Figure 1: Duplicate (left) and Near-Duplicate (right) rates against performance changes in F1 scores. The positive values on the left y-axis (performance change) indicate an improvement while the negative values on the left y-axis indicate a decline. The zero line marks no change in performance.

all baseline and state-of-the-art approaches, we propose a new evaluation method. Firstly, we save 5 model checkpoints from the last 5 epochs during training, each representing a unique model with varying parameters. We then compare the ranking of these 5 models according to F1-macro scores before and after deduplication. We use *Train Set Original* and *Train Set w/o Duplicates* for training and the same test set for testing, similar to the previous experiment.

5.3 Inconsistent Labels

In the few-shot experiment, we aim to investigate the impact of label inconsistency in duplicates on the predictive performance of LLMs using two similar few-shot prompts:

- *Given Sample A (label X) and Sample A (label Y), predict if the following example is X or Y: Sample A*;
- *Given Sample A (label Y) and Sample A (label*

X), predict if the following example is X or Y: Sample A.

We then insert n randomly selected samples between two Sample A instances with different labels for both prompts, where n varies from 0 to 5. For comparison, we also use zero-shot evaluation with prompt *predict if the following example is X or Y: Sample A*.

5.4 Results and Discussion

Data Leakage. Table 3 presents the predictive results across all tasks with different training set configurations (i.e., Original, w/o Duplicates and w/o Near-Duplicates) and Figure 1 shows the duplicate (left) and near-duplicate (right) rates against F1 score changes. In general, we observe that the presence of duplicated samples in the training set results in an overestimation of the model's predictive performance (14 out of 19 datasets). For example, the model achieves 62.6 F1 in 'Original' and 51.9 F1 in 'w/o Duplicates' (t-test, $p < .05$) in `Twitter 16`. This is due to duplicate posts in the training and test set leading to label leakage during model training. Meanwhile, we notice that model performance on datasets (e.g., `Sentiment` and `P-Stance`) with minimal duplicates (i.e., Duplicated% < 0.01) remains steady. Additionally, we observe a similar result in 'w/o Duplicates' and 'w/o Near-duplicate' where the results of the latter are slightly worse.

Table 4 demonstrates that the zero-shot classification performance of recent LLMs may not always surpass that of fully fine-tuned BERT-style models. This suggests that supervised approaches remain indispensable in CSS research, thereby making the resolution of the data duplication issue inevitable.

Model Rankings. We present cross-model evaluation results based on macro-F1 scores for selected datasets in Table 5 and for all datasets in Appendix B. The findings reveal that the majority of model rankings exhibit inconsistency (17 out of 19 datasets) before and after deduplication. Taking `HatEval'19` for example, the top five model checkpoints are from Epoch 6, 7, 10, 7 and 9 respectively when containing duplicates; while the top five ones are from Epoch 9, 10, 8, 7 and 6 after deduplication. However, we notice the same rankings for dataset `Twitter 16` and `TBO`, which may result from the small size of the dataset (e.g., 818 tweets). This suggests that data duplication undermines the validity of claimed SoTA results. We contend that current SoTA approaches may require reassessment through essential data cleaning measures.

Inconsistent Labels. Table 6 presents the results of stance detection using few-shot prompts. We observe that the model can predict the stance correctly in zero-shot settings. Also, when presented with fewer instances containing posts with inconsistent labels, the model tends to make more accurate predictions. However, when provided with more than five instances, the model begins to align its predictions with the most recent label. This indicates that as the number of instances increases, predictions of the model may be less reliable, particularly when presented with conflicting information. We obtain similar findings from other social media tasks. This highlights the importance of data quality and consistency in training models for downstream CSS tasks.

5.5 Error Analysis

To further investigate the impact of such duplicates and near-duplicates in social media datasets, we manually perform an error analysis on the test sets of five misinformation detection datasets. We first mark the duplicates and near-duplicates in the test set based on *Train Set Original*. Figure 2 presents the percentage of these duplicates (upper) and these near-duplicates (bottom) in wrong predictions by models from five misinformation detection datasets (see Appendix C). We observe that most duplicated ($\sim$ 99%) and near-duplicated (more than 90%) samples across the training and test sets can be correctly predicted.

This emphasizes the potential overestimation of model performance on existing social media datasets as a result of label leakage. Nonetheless, the presence of duplicate posts can still pose challenges for the model, especially when there is label inconsistency, as demonstrated in Table 1. Our error analysis highlights the adverse impact of duplicated and near-duplicated samples on the reliability of model performance.

6 Effective Strategies for Development and Use of Social Media Datasets

This section outlines practical recommendations we propose for more effective development and use of datasets in responsible CSS research.

	Original		w/o Dupl.		w/o Near-dupl.		GPT-4o		LLaMA-3	
Dataset	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**
Complaint	88.9±1.3	89.8±1.3	89.5±1.3	90.3±1.2	89.2±0.3	90.0±0.2	81.2	81.4	77.4	77.7
Bragging	77.8±0.7	91.4±0.5	77.2±0.5	91.1±0.3	77.8±0.7	91.3±0.7	70.0	85.2	47.9	62.5
Sarcasm	64.7±7.6	81.5±1.0	61.4±12.4	80.8±1.9	69.8±0.8	82.2±0.7	53.1	53.9	39.4	59.1
CovidVaxx	80.7±0.4	80.7±0.5	79.2±0.9	79.2±0.9	79.0±0.8	79.1±0.8	77.6	76.9	63.6	63.1
US-Election	53.4±0.7	56.8±0.9	52.4±0.8	56.5±0.8	51.2±1.7	55.3±0.5	47.7	50.0	41.5	46.4

Table 4: Predictive results of LLMS (GPT-4o and LLaMA-3-8B) (right) vs. BERTweet (Original, w/o Dupl. and w/o Near-dupl.) (left).

Dataset	**NO.1**	**NO.2**	**NO.3**	**NO.4**	**NO.5**
HatEval'19-duplicate	80.25 (E6)	80.11 (E7)	80.01 (E10)	79.86 (E7)	79.41 (E9)
HatEval'19-noduplicate	80.18 (E9)	80.01 (E10)	79.78 (E8)	79.77 (E7)	79.34 (E6)
Twitter 16-duplicate	75.30 (E10)	74.65 (E9)	74.56 (E8)	73.97 (E7)	73.44 (E6)
Twitter 16-noduplicate	67.22 (E10)	67.04 (E9)	66.92 (E8)	66.40 (E7)	66.22 (E6)
Bragging-duplicate	80.00 (E7)	79.45 (E9)	79.33 (E8)	78.60 (E10)	78.47 (E6)
Bragging-noduplicate	78.22 (E6)	78.00 (E7)	77.86 (E9)	77.77 (E8)	77.39 (E10)
P-Stance-duplicate	81.17 (E10)	80.80 (E9)	80.72 (E8)	80.71 (E6)	80.39 (E7)
P-Stance-noduplicate	80.55 (E8)	80.38 (E10)	80.36 (E7)	80.35 (E9)	80.14 (E6)

Table 5: Rankings of five model checkpoints across selected datasets based on macro-F1 scores. E denotes epoch, e.g., E6 refers to 6th Epoch. Unchanged rankings are in light gray. Full experimental results are displayed in Appendix B.

Prompt	**Pred.**
Zero-shot	Against
Sample A (Neutral)	Against
Sample A (Against)	Against
Sample A (Neutral), Sample A (Against)	Against
Sample A (Against), Sample A (Neutral)	Neutral
Sample A (Neutral), Sample B, Sample A (Against)	Against
Sample A (Against), Sample B, Sample A (Neutral)	Against
Sample A (Neutral), Sample B, C, Sample A (Against)	Against
Sample A (Against), Sample B, C, Sample A (Neutral)	Against
Sample A (Neutral), Sample B, C, D, Sample A (Against)	Against
Sample A (Against), Sample B, C, D, Sample A (Neutral)	Neutral
Sample A (Neutral), Sample B, C, D, E, Sample A (Against)	Against
Sample A (Against), Sample B, C, D, E, Sample A (Neutral)	Neutral
Sample A (Neutral), Sample B, C, D, E, F, Sample A (Against)	Against
Sample A (Against), Sample B, C, D, E, F, Sample A (Neutral)	Neutral

Table 6: Predictive results of GPT-4o on stance detection using US-Election. Sample A is *@USER @USER donald trump's lessons for republicans: no consequences for lying. no consequences for hate. no consequences for immorality. no consequences for crimes. no consequences for collusion. no consequences for impeachment. no consequences for incompetence. sad. #votebluenomatterwho* and Sample B, C, D, E, F are randomly selected with original labels from the same dataset.

6.1 Developing New Datasets

Developing a new social media dataset typically involves several key steps in a well-defined pipeline including problem definition, data collection, annotation and labeling, preprocessing, optionally balancing and stratification.

However, we suggest performing an initial data cleaning before the annotation task. More specifically, we recommend first performing standard data preprocessing steps (e.g., replacing @USER and URL tokens)[6] and then removing duplicate samples. This process can help reduce human annotation costs (Nguyen et al., 2017) and avoid potentially inconsistent labeling. Additionally, manual rules (e.g., excluding posts with fewer than N tokens) can also be used for further data cleaning before annotating, as suggested by Zampieri et al. (2023) who developed a dataset with almost no duplicates (see Table 2).

Also, we suggest dataset developers provide different deduplication versions of datasets, for example, including and excluding near-duplicate posts, as both of them are likely to provide valuable information for the community.

6.2 Using Existing Datasets

An existing dataset may be used for model training as is, or enriched through fine-grained reannotation or incorporating multiple modalities. To make the most of existing datasets, attention to data quality is crucial, especially when dealing with duplicate samples. We suggest retaining one of the duplicate posts with consistent labels and excluding all posts with conflicting labels unless a reliable reannotation task is performed. In cases where duplicates are absent, consider employing out-of-box similarity checks using methods like Levenshtein

[6]For annotation tasks, developers may want to restore these tokens for providing the original posts to annotators.

Distance and Cosine Similarity to exclude near-duplicate samples.

Notably, for datasets like PHEME annotated with more than one type of label (e.g., Veracity label: *Rumor* or *Non-rumor* and Event-related label: *Charlie Hebdo Shooting* or *Ottawa Shooting*, consider task-specific data splitting strategies. The leave-one-out protocol (Lukasik et al., 2016), for instance, is advocated for more representative and effective evaluations.

6.3 Updating Data Checklists

Recently, most Computer Science venues have mandated a Data & Ethics checklist to promote *Responsible AI Research* emphasizing ethical and reproducibility considerations (Dodge et al., 2019; Rogers et al., 2021).

We argue that current versions of Data & Ethics checklists from two leading organizations in CSS, *ACL[7] and AAAI-ICWSM[8], may need a 'minor revision' to account for data deduplication. Therefore, we recommend including an additional question to remind researchers regarding the essential process of deduplicating data. Below are our revised versions for the two Data & Ethics checklists:

> **Under *ACL Checklist Section B**
>
> ***Title: Did you use or create scientific artifacts?***
>
> Did you perform dataset exploration, such as applying data pre-processing and **deduplication**, when creating new or utilizing existing data resources?
>
> **Under AAAI-ICWSM Checklist Section 5**
>
> ***Title: If you are using existing assets (e.g., code, data, models) or curating/releasing new assets.***
>
> Have you provided details of any meta-analysis conducted on the CSS datasets employed or developed in your research? This includes, but is not limited to, data pre-processing, **deduplication**, and other processes essential for responsible CSS research.

We believe this adjustment applies not only to CSS but also to broader research fields.

[7] https://aclrollingreview.org/responsibleNLPresearch/

[8] https://www.overleaf.com/latex/templates/aaai-icwsm-2024-paper-checklist/vxbztbhhrbch

7 Conclusion

Beyond focusing on innovative methodologies to improve model performance in downstream CSS tasks, we advocate a more comprehensive understanding of the datasets via meta analysis. In this work, we conduct an extensive evaluation of the data duplication issue in 20 selected social media datasets. We find that the majority of these datasets require additional preprocessing before the modeling, a critical step often overlooked in previous CSS research. Furthermore, we explore the potential impact of data duplication on model performance with a focus on label leakage, model rankings and label inconsistency. Finally, we offer targeted recommendations for more effective use of existing datasets and the creation of new datasets derived from social media sources.

Ethics Statement

The Research Ethics Committee of our institute has granted approval for our work. The datasets evaluated in this study were acquired from original authors by request or via links available in the source papers.

Limitations

Due to space limits, our work only focuses on 20 representative CSS datasets, which do not cover the entire scope of computational social science. However, our work has examined a similar or higher number of datasets compared to existing CSS benchmarks such as TweetEval (Barbieri et al., 2020) and SuperTweetEval (Antypas et al., 2023). In the future, we are committed to continually expanding our analysis by evaluating more datasets, with updates to be shared through our GitHub repository. We aim to create a processed CSS benchmark, similar to existing CSS benchmarks such as SuperTweetEval (Antypas et al., 2023). This will include a cleaned version of each dataset, along with re-evaluated baselines based on these deduplicated datasets.

Acknowledgments

This work has been supported by the UK's innovation agency (Innovate UK) grant 10039055 (approved under the Horizon Europe Programme as vera.ai, EU grant agreement 101070093). NA is supported by EPSRC grant EP/Y009800/1, part of the RAI UK Keystone projects.

References

Firoj Alam, Muhammad Imran, and Ferda Ofli. 2017. Image4act: Online Social Media Image Processing for Disaster Response. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, pages 601–604.

Dimosthenis Antypas, Asahi Ushio, Francesco Barbieri, Leonardo Neves, Kiamehr Rezaee, Luis Espinosa-Anke, Jiaxin Pei, and Jose Camacho-Collados. 2023. Supertweeteval: A challenging, unified and heterogeneous benchmark for social media nlp research.

Dennis Assenmacher, Indira Sen, Leon Fröhling, and Claudia Wagner. 2020. The end of the rehydration era the problem of sharing harmful twitter research data.

Tyler Baldwin and Yunyao Li. 2015. An In-depth Analysis of the Effect of Text Normalization in Social Media. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 420–429.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650.

Aki Barry, Lei Han, and Gianluca Demartini. 2023. On the Impact of Data Quality on Image Classification Fairness.

Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 us presidential election online discussion. *First monday*, 21(11-7).

Eric Breck, Neoklis Polyzotis, Sudip Roy, Steven Whang, and Martin Zinkevich. 2019. Data Validation for Machine Learning. In *MLSys*.

Lukas Budach, Moritz Feuerpfeil, Nina Ihde, Andrea Nathansen, Nele Noack, Hendrik Patzlaff, Felix Naumann, and Hazar Harmouch. 2022. The Effects of Data Quality on Machine Learning Performance.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80. IEEE.

Liviu-Adrian Cotfas, Camelia Delcea, Ioan Roxin, Corina Ioanăş, Dana Simona Gherai, and Federico Tajariol. 2021. The longest month: analyzing covid-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement. *Ieee Access*, 9:33203–33223.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Matthew J Denny and Arthur Spirling. 2018. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, 26(2):168–189.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194.

Fernando Leandro Dos Santos and Marcelo Ladeira. 2014. The role of text pre-processing in opinion mining on a social media language dataset. In *2014 Brazilian Conference on Intelligent Systems*, pages 50–54. IEEE.

Achim Edelmann, Tom Wolff, Danielle Montagne, and Christopher A Bail. 2020. Computational social science and sociology. *Annual Review of Sociology*, 46:61–81.

Ibrahim Abu Farha, Silviu Oprea, Steve Wilson, and Walid Magdy. 2022. Semeval-2022 task 6: isarcasmeval, intended sarcasm detection in english and arabic. In *The 16th International Workshop on Semantic Evaluation 2022*, pages 802–814. Association for Computational Linguistics.

Emilio Ferrara. 2020. What types of covid-19 conspiracies are populated by twitter bots?

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Òscar Garibo i Orts. 2019. Multilingual detection of hate speech against immigrants and women in Twitter at SemEval-2019 task 5: Frequency analysis interpolation for hate in speech detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 460–463, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Christoph Gröger. 2021. There is no AI without data. *Communications of the ACM*, 64(11):98–108.

Ratab Gull, Umar Shoaib, Saba Rasheed, Washma Abid, and Beenish Zahoor. 2016. Pre processing of twitter's data for opinion mining in political context. *Procedia Computer Science*, 96:1560–1570.

Sharath Chandra Guntuku, Daniel Preotiuc-Pietro, Johannes C Eichstaedt, and Lyle H Ungar. 2019. What twitter profile and posted images reveal about depression and anxiety. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 236–246.

Jake M Hofman, Duncan J Watts, Susan Athey, Filiz Garip, Thomas L Griffiths, Jon Kleinberg, Helen Margetts, Sendhil Mullainathan, Matthew J Salganik, Simine Vazire, et al. 2021. Integrating explanation and prediction in computational social science. *Nature*, 595(7866):181–188.

Zhao Jianqiang and Gui Xiaolin. 2017. Comparison research on text pre-processing methods on twitter sentiment analysis. *IEEE access*, 5:2870–2879.

Mali Jin, Daniel Preotiuc-Pietro, A. Seza Doğruöz, and Nikolaos Aletras. 2022. Automatic identification and classification of bragging in social media. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3945–3959, Dublin, Ireland. Association for Computational Linguistics.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pages 10697–10707. PMLR.

Kornraphop Kawintiranon and Lisa Singh. 2021. Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 4725–4735.

Bernard Koch, Emily Denton, Alex Hanna, and Jacob G Foster. 2021. Reduced, reused and recycled: The life of a dataset in machine learning research. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1.

Akrivi Krouska, Christos Troussas, and Maria Virvou. 2016. The effect of preprocessing techniques on twitter sentiment analysis. In *2016 7th international conference on information, intelligence, systems & applications (IISA)*, pages 1–5. IEEE.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

David MJ Lazer, Alex Pentland, Duncan J Watts, Sinan Aral, Susan Athey, Noshir Contractor, Deen Freelon, Sandra Gonzalez-Bailon, Gary King, Helen Margetts, et al. 2020. Computational social science: Obstacles and opportunities. *Science*, 369(6507):1060–1062.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Peng Li, Xi Rao, Jennifer Blase, Yue Zhang, Xu Chu, and Ce Zhang. 2021a. CleanML: A study for Evaluating the Impact of Data Cleaning on ML Classification Tasks. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 13–24. IEEE.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021b. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365.

Michal Lukasik, PK Srijith, Duy Vu, Kalina Bontcheva, Arkaitz Zubiaga, and Trevor Cohn. 2016. Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter. In *Proceedings of 54th Annual Meeting of the Association for Computational Linguistics*, pages 393–398. Association for Computational Linguistics.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 3818–3824.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 708–717.

Jenifer Mahilraj, Getahun Tigistu, and Sisay Tumsa. 2020. Text Preprocessing Method on Twitter Sentiment Analysis Using Machine Learning. *International Journal of Innovative Technology and Exploring Engineering*, 9(12):233–240.

Antonis Maronikolakis, Danae Sánchez Villegas, Daniel Preoţiuc-Pietro, and Nikolaos Aletras. 2020. Analyzing political parody in social media. In *Proceedings*

of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4373–4384.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.

Yida Mu, Kalina Bontcheva, and Nikolaos Aletras. 2023a. It's about time: Rethinking evaluation on rumor detection benchmarks using chronological splits. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 724–731.

Yida Mu, Mali Jin, Charlie Grimshaw, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023b. Vaxxhesitancy: A dataset for studying hesitancy towards covid-19 vaccination on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1052–1062.

Yida Mu, Ben P. Wu, William Thorne, Ambrose Robinson, Nikolaos Aletras, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2024. Navigating prompt complexity for zero-shot classification: A study of large language models in computational social science. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12074–12086, Torino, Italia. ELRA and ICCL.

Usman Naseem, Imran Razzak, and Peter W Eklund. 2021. A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter. *Multimedia Tools and Applications*, 80:35239–35266.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Dat Tien Nguyen, Firoj Alam, Ferda Ofli, and Muhammad Imran. 2017. Automatic Image Filtering on Social Networks Using Deep Learning and Perceptual Hashing During Crises.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Barbara Plank and Dirk Hovy. 2015. Personality traits on twitter—or—how to get 1,500 personality tests in a week. In *Proceedings of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 92–98.

Daniel Preoţiuc-Pietro, Mihaela Gaman, and Nikolaos Aletras. 2019. Automatically identifying complaints in social media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5019, Florence, Italy. Association for Computational Linguistics.

Dongning Rao, Xin Miao, Zhihua Jiang, and Ran Li. 2021. STANKER: Stacking network based on level-grained attention-masked BERT for rumor detection on social media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3347–3363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.

Anna Rogers, Timothy Baldwin, and Kobi Leins. 2021. 'just what do you think you're doing, dave?'a checklist for responsible data use in nlp. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4821–4833.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2019. Semeval-2017 task 4: Sentiment analysis in twitter.

Ankan Saha and Vikas Sindhwani. 2012. Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 693–702.

Tulika Saha, Apoorva Upadhyaya, Sriparna Saha, and Pushpak Bhattacharyya. 2021. Towards sentiment and emotion aided multi-modal speech act classification in Twitter. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5727–5737, Online. Association for Computational Linguistics.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Massimo Stella, Emilio Ferrara, and Manlio De Domenico. 2018. Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*, 115(49):12435–12440.

Symeon Symeonidis, Dimitrios Effrosynidis, and Avi Arampatzis. 2018. A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems with Applications*, 110:298–310.

Lin Tian, Xiuzhen Zhang, and Jey Han Lau. 2022. DUCK: Rumour detection on social media by modelling user and comment propagation networks. In *Proceedings of the 2022 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4939–4949, Seattle, United States. Association for Computational Linguistics.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat Thalmann. 2013. Who, where, when and what: discover spatio-temporal topics for twitter users. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 605–613.

Marcos Zampieri, Skye Morgan, Kai North, Tharindu Ranasinghe, Austin Simmmons, Paridhi Khandelwal, Sara Rosenthal, and Preslav Nakov. 2023. Target-based offensive language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 762–770.

Jianyang Zhang, Tao Liang, Mingyang Wan, Guowu Yang, and Fengmao Lv. 2022a. Curriculum knowledge distillation for emoji-supervised cross-lingual sentiment analysis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 864–875, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Menghan Zhang, Xue Qi, Ze Chen, and Jun Liu. 2022b. Social bots' involvement in the covid-19 vaccine discussions on twitter. *International Journal of Environmental Research and Public Health*, 19(3):1651.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science?

Arkaitz Zubiaga. 2018. A longitudinal assessment of the persistence of twitter datasets. *Journal of the Association for Information Science and Technology*, 69(8):974–984.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989.

Appendix

A Hyper-parameters and Experimental Setup

Following the standard pipeline (Devlin et al., 2019), we fine-tune BERT-style models by feeding the '[CLS]' token to a linear classifier with Softmax activation. We set the learning rate lr = 2e-5 and batch size bs = 64 for all datasets. All BERT variants are fine-tuned for up to 10 epochs using an early stopping strategy based on the validation loss (we use 10% of the data from the training set as validation sets for model selection). We run all BERT-style models three times with different random seeds and then report the standard deviations and average Precision, Recall, and F1-measure. For LLMs, we fix the random seed and temperature values to ensure reproducibility.

All supervised experiments are conducted on an Nvidia V100 GPU with 32 GB of memory, with a total running time of approximately 3 hours (on 20 datasets described in Table 2). All GPT prompting experiments cost around 3.5 USD in total, which can be fully covered by the free budget provided by OpenAI for new users.

B Complete results of model rankings

Table 7 shows the full results of rankings of five model checkpoints across all datasets.

C Error Analysis

Figure 2 presents the percentage of these duplicates (upper) and these near-duplicates (bottom) in wrong predictions by models from five misinformation detection datasets.

D Prompts for LLM Zero- and Few-shot Classification

Dataset	NO.1	NO.2	NO.3	NO.4	NO.5
Offensive Language Detection					
WASEEM-duplicate	83.62 (E6)	83.33 (E7)	83.17 (E10)	83.06 (E9)	83.00 (E8)
WASEEM-noduplicate	83.71 (E6)	82.89 (E10)	82.61 (E7)	82.53 (E9)	82.51 (E8)
TBO-duplicate	69.76 (E6)	69.72 (E10)	69.63 (E9)	69.61 (E8)	69.54 (E7)
TBO-noduplicate	69.76 (E6)	69.72 (E10)	69.63 (E9)	69.61 (E8)	69.54 (E7)
OLID-duplicate	75.30 (E7)	74.66 (E6)	74.52 (E9)	74.38 (E10)	74.01 (E8)
OLID-noduplicate	75.82 (E10)	75.62 (E6)	75.53 (E7)	75.21 (E9)	75.05 (E8)
FOUNTA-duplicate	84.37 (E6)	83.68 (E9)	83.61 (E8)	83.43 (E10)	83.25 (E7)
FOUNTA-noduplicate	84.20 (E6)	83.76 (E8)	83.71 (E7)	83.42 (E9)	83.29 (E10)
HatEval'19-duplicate	80.25 (E6)	80.11 (E7)	80.01 (E10)	79.86 (E7)	79.41 (E9)
HatEval'19-noduplicate	80.18 (E9)	80.01 (E10)	79.78 (E8)	79.77 (E7)	79.34 (E6)
Misinformation Detection					
Twitter 15-duplicate	69.61 (E8)	67.90 (E10)	67.76 (E9)	67.41 (E7)	66.47 (E6)
Twitter 15-noduplicate	67.58 (E7)	67.19 (E8)	67.00 (E10)	66.75 (E9)	65.45 (E6)
Twitter 16-duplicate	75.30 (E10)	74.65 (E9)	74.56 (E8)	73.97 (E7)	73.44 (E6)
Twitter 16-noduplicate	67.22 (E10)	67.04 (E9)	66.92 (E8)	66.40 (E7)	66.22 (E6)
PHEME-duplicate	86.93 (E8)	86.58 (E10)	86.54 (E7)	86.54 (E9)	85.92 (E6)
PHEME-noduplicate	85.95 (E8)	85.91 (E10)	85.83 (E9)	85.80 (E7)	85.61 (E6)
Weibo-duplicate	91.75 (E9)	91.43 (E10)	91.10 (E8)	90.78 (E7)	90.35 (E6)
Weibo-noduplicate	91.75 (E8)	91.64 (E9)	91.00 (E10)	90.68 (E6)	90.25 (E7)
STANKER-duplicate	93.82 (E7)	93.66 (E8)	93.33 (E9)	93.24 (E10)	93.16 (E6)
STANKER-noduplicate	92.92 (E10)	92.83 (E9)	92.75 (E8)	92.25 (E6)	92.17 (E7)
Speech Act & Sentiment Analysis					
Complaint-duplicate	90.73 (E6)	90.40 (E9)	90.34 (E10)	90.01 (E8)	89.60 (E7)
Complaint-noduplicate	90.85 (E7)	90.59 (E10)	90.52 (E6, E9)	90.52 (E6, E9)	90.44 (E8)
Sarcasm-duplicate	71.53 (E10)	71.45 (E9)	71.21 (E8)	70.76 (E6)	70.22 (E7)
Sarcasm-noduplicate	72.41 (E7)	71.79 (E10)	71.64 (E6)	71.49 (E8)	71.42 (E9)
Bragging-duplicate	80.00 (E7)	79.45 (E9)	79.33 (E8)	78.60 (E10)	78.47 (E6)
Bragging-noduplicate	78.22 (E6)	78.00 (E7)	77.86 (E9)	77.77 (E8)	77.39 (E10)
Parody-duplicate	-	-	-	-	
Parody-noduplicate	-	-	-	-	
Sentiment-duplicate	73.61 (E6)	73.11 (E7)	72.90 (E9)	72.86 (E10)	72.79 (E8)
Sentiment-noduplicate	73.39 (E6)	72.90 (E9)	72.81 (E10)	72.78 (E8)	72.56 (E7)
Stance Detection					
CovidVaxx-duplicate	80.93 (E8)	80.78 (E6)	80.43 (E9)	79.89 (E10)	78.99 (E7)
CovidVaxx-noduplicate	81.01 (E7)	80.32 (E9)	80.27 (E6)	80.27 (E8)	80.11 (E10)
RumorEval-duplicate	54.66 (E10)	54.44 (E9)	53.77 (E8)	53.76 (E6)	53.42 (E7)
RumorEval-noduplicate	54.27 (E6)	54.03 (E10)	53.96 (E9)	53.83 (E8)	53.23 (E7)
US-Election-duplicate	56.53 (E8)	56.04 (E7)	54.90 (E9)	54.63 (E6)	54.12 (E10)
US-Election-noduplicate	56.20 (E8)	55.36 (E6)	55.56 (E10)	55.46 (E9)	52.85 (E7)
P-Stance-duplicate	81.17 (E10)	80.80 (E9)	80.72 (E8)	80.71 (E6)	80.39 (E7)
P-Stance-noduplicate	80.55 (E8)	80.38 (E10)	80.36 (E7)	80.35 (E9)	80.14 (E6)

Table 7: Rankings of five model checkpoints across selected datasets based on macro-F1 scores. E denotes epoch, e.g., E6 refers to 6th Epoch. Unchanged rankings are in light gray.

Dataset	Prompt
Complaint	Read the given tweet, and categorize it into one of two categories: (1) Non-complaint (2) Complaint Only return the category number as your answer. Text: {list_of_text} Answer:
Bragging	Read the given tweet, and categorize it into one of two categories: (1) Not Bragging (2) Bragging Only return the category number as your answer. Text: {list_of_text} Answer:
Sarcasm	Read the given tweet, and categorize it into one of two categories: (1) Not Sarcasm (2) Sarcasm Only return the category number as your answer. Text: {list_of_text} Answer:
CovidVAXX	Read the given tweet, and categorize it according to the stance expressed about the COVID-19 vaccine: (1) Anti vaccine (2) Neutral (3) Pro vaccine Only return the category number as your answer. Text: {list_of_text} Answer:
US_election	Read the given tweet, and categorize it according to the stance expressed towards U.S. politicians: (1) None (2) Favor (3) Against Only return the category number as your answer. Text: {list_of_text} Answer:

Table 8: Example prompts used for LLM-based zero-shot classification.

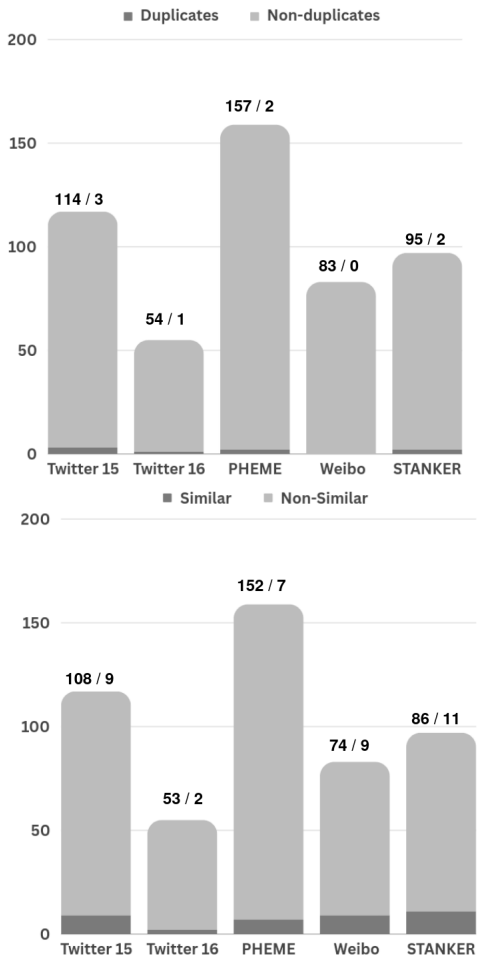

Figure 2: Ratio of duplicates (upper) and near-duplicates (bottom) in wrong predictions from five misinformation detection datasets.