Detecting Cybercrimes in Accordance with Pakistani Law: Dataset and Evaluation using PLMs

Faizad Ullah[1], **Ali Faheem**[1], **Ubaid Azam**[2], **Muhammad Sohaib Ayub**[1], **Faisal Kamiran**[3], **and Asim Karim**[1]
[1]Department of Computer Science, Lahore University of Management Sciences (LUMS), Lahore, Pakistan 54792
[2]Electronics and Computer Science, University of Southampton (SOTON), United Kingdom
[3]Department of Computer Science, Information Technology University (ITU), Lahore, Pakistan
{faizad.ullah, ali.faheem, sohaib.ayub, akarim}@lums.edu.pk
u.azam@soton.ac.uk
faisal.kamiran@itu.edu.pk

Abstract

Cybercrime is a serious and growing threat affecting millions of people worldwide. Detecting cybercrimes from text messages is challenging, as it requires understanding the linguistic and cultural nuances of different languages and regions. Roman Urdu is a widely used language in Pakistan and other South Asian countries, however, it lacks sufficient resources and tools for natural language processing and cybercrime detection. To address this problem, we make three main contributions in this paper. (1) We create and release CRU, a benchmark dataset for text-based cybercrime detection in Roman Urdu, which covers a number of cybercrimes as defined by the Prevention of Electronic Crimes Act (PECA) of Pakistan. This dataset is annotated by experts following a standardized procedure based on Pakistan's legal framework. (2) We perform experiments on four pre-trained language models (PLMs) for cybercrime text classification in Roman Urdu. Our results show that xlm-roberta-base is the best model for this task, achieving the highest performance on all metrics. (3) We explore the utility of prompt engineering techniques, namely prefix and cloze prompts, for enhancing the performance of PLMs for low-resource languages such as Roman Urdu. We analyze the impact of different prompt shapes and k-shot settings on the performance of xlm-roberta-base and bert-base-multilingual-cased. We find that prefix prompts are more effective than cloze prompts for Roman Urdu classification tasks, as they provide more contextually relevant completions for the models. Our work provides useful insights and resources for future research on cybercrime detection and text classification in low-resource languages.

Keywords: Cybercrime Detection, Roman Urdu, Pakistani Cybercrimes Law, Prompt-Based Classification

1. Introduction

A crime is an unlawful act or omission punishable by a state. When a computer/mobile phone or the internet is used as a tool, a target, or both to commit a crime, then it is known as cybercrime. Over the past few decades, the number of Internet and social media users has been growing rapidly. Consequently, cybercrimes are increasing steadily globally and in Pakistan. This continuously increasing online communication and the extensive usage of social networking platforms have created new avenues for cybercrime, including cyberbullying, cyber terrorism, cyber harassment, and hate speech. These cybercrimes are totally unique, transnational, invisible, rapid, and ever-evolving. It is necessary to identify these crimes in a timely manner to avoid any inconvenience and to protect the citizens' dignity.

While much research has been dedicated to specific cybercrime categories such as hate speech (Waseem et al., 2017), cyberbullying (Dadvar and de Jong, 2012), and fraud detection (Wang, 2010), a comprehensive framework for cybercrime detection, grounded in the legal context of a particular country or region, remains largely unexplored. Law enforcement agencies educate people using different awareness measures, and social networking sites/apps such as Facebook and Yahoo! update their Terms of Service (ToS) to prevent cybercrimes. Facebook ToS prohibits posting content that is: pornographic, hateful, threatening, or incites violence. Yahoo! ToS are like forbidding content that is hateful, unlawful, threatening, harassing, abusive, defamatory, vulgar, invasive of another's privacy, ethnically, racially, or otherwise objectionable. In Pakistan, the Prevention of Electronic Crimes Act (PECA) (PECA, 2016) was enacted in 2016 to provide a legal framework for combating cybercrimes and protecting the rights of online users. The PECA defines various types of cybercrimes and their punishments. It also established an investigation agency and a forensic laboratory for dealing with electronic crimes.

One of the challenges in implementing the PECA is the linguistic diversity of Pakistan. Pakistan has six major and more than 50 regional languages (Li et al., 2022). Urdu is the national language and one of the official languages of Pakistan, along with English. Roman Urdu (RU) is a variant of Urdu that

LREC-COLING 2024, pages 4717–4728
20-25 May, 2024.

uses the Latin script instead of the Perso-Arabic script. Pakistani users use Roman Urdu extensively for online communication and social media platforms (Li et al., 2022). Roman Urdu is also used in other South Asian countries, such as India and Bangladesh. A survey acknowledges that "300 million people are speaking Urdu and about 11 million speakers in Pakistan from which maximum users prefer Roman Urdu for the textual communication" (Shahroz et al., 2020). Another study notes that "Urdu is a national language of Pakistan, and more than 100 million speak Urdu and use it with Roman text to express their views about products, events, services, issues, and topics" (Khan et al., 2021).

Roman Urdu poses several challenges for natural language processing (NLP) and cybercrime detection due to its lack of standardization, linguistic variations, code-mixing with other languages, and informal writing style (Li et al., 2022). There are few datasets for Roman Urdu text analysis, especially for cybercrime detection. Most of the available datasets focus on specific types of cybercrimes, such as hate speech (Bilal et al., 2023), offensive text (Sajid et al., 2020), or cyberbullying (Dewani et al., 2021). These datasets do not cover all the text-based cybercrimes that the PECA defines and do not follow its legal framework. Moreover, these datasets are relatively small in size and imbalanced in class distribution (Bilal et al., 2023; Dewani et al., 2021; Sajid et al., 2020). There is a need for a large-scale and comprehensive dataset for text-based cybercrime detection in Roman Urdu aligned with the PECA.

Existing methods for cybercrime detection in RU/Urdu are based on machine learning (ML) or deep learning (DL) techniques that require manual or automatic feature extraction from text data (Dewani et al., 2023, 2021; Rizwan et al., 2020). These methods suffer from limitations such as data scarcity, noise sensitivity, and context ignorance. Prompt-based techniques have emerged in recent times as a strong alternative in NLP that uses natural language prompts to guide PLMs for various tasks (Liu et al., 2023; Gao et al., 2021; Brown et al., 2020; Ullah et al., 2023). These techniques can leverage the knowledge and skills of PLMs without fine-tuning them on task-specific data, achieving state-of-the-art results on many NLP tasks. However, they have not been explored for RU/Urdu cybercrime detection. We adopt and utilize prompt-based techniques for cybercrime text classification in RU, a novel and promising direction for this task. These techniques can potentially overcome some of the challenges of ML and DL methods by using less data, being more robust, and incorporating context and domain knowledge into the prompts (Brown et al., 2020).

In this paper, we make two key contributions to cybercrime detection in Roman Urdu. First, we create and release a benchmark dataset Cybercrimes in Roman Urdu (CRU)[1], covering a wide range of cybercrimes defined by the PECA Act of Pakistan. This dataset is annotated by experts following a standardized procedure based on Pakistan's legal framework. Second, we perform experiments on four pre-trained multilingual transformers, namely DistilBERT-Base-Multilingual (Sanh et al., 2019), BERT-Multilingual-Base (Devlin et al., 2019), XLM-RoBERTa-Base (Conneau et al., 2020), and Multilingual-MiniLM (Wang et al., 2020), for cybercrime text classification in RU. We evaluate their performance using standard metrics such as precision, recall, macro F1-score, and accuracy. We also explore the significance of prompt engineering techniques, namely prefix and cloze prompts (Liu et al., 2023), for enhancing the performance of pre-trained multilingual transformers for Roman Urdu cybercrimes detection. We analyze the impact of different prompt shapes and k-shot settings on the performance of the models. Our results show that XLM-RoBERTa-Base performs better in standard fine-tuning while BERT-Multilingual-Base with prefix prompts achieves the highest performance in Roman Urdu cybercrimes detection.

2. Related Work

Research in the field of cybercrime detection and related tasks has witnessed rapid evolution over the past few years. Existing research encompasses a diverse range of but interconnected tasks, including the detection of racism and sexism on platforms like Twitter (Waseem, 2016), distinguishing offensive language from hate speech (Davidson et al., 2017), and identifying hateful and non-hateful speech within white supremacy forums (De Gibert et al., 2018). Traditionally, researchers have identified and classified cybercrimes by extracting specific crime-related keywords from text data. Additionally, various features have been utilized in the literature, such as bag-of-words, n-grams (Nobata et al., 2016), sentiment analysis, lexical resources, linguistic features, knowledge-based features, meta-information, and word embeddings (Kwok and Wang, 2013; Davidson et al., 2017; Greevy and Smeaton, 2004). In recent years, however, the field has shifted from feature-based models to data-driven models.

The emergence of deep learning has brought significant advancements to NLP in recent years (Kovács et al., 2021). Notably, word embeddings and attention-based learning have revolutionized the field. Embeddings, as features, have sur-

[1] https://github.com/Faizad/CRU-LREC-COLING-2024

passed traditional methods like BoW and classical features (Djuric et al., 2015) in hate speech identification. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid architectures have also been employed for tasks related to hate speech, racism, offensive language, and sexism. However, the most profound development has been the rise of transformers, particularly exemplified by models like BERT (Devlin et al., 2018), which, in a recent competition for hate speech detection, delivered seven out of the ten best-performing models in a subtask (Zampieri et al., 2019). However, most of the work is focused on English and other resource-rich languages(Ullah et al., 2023).

A significant gap in this domain pertains to low-resource languages. Minimal efforts have been made for low-resource languages, particularly for Roman Urdu. According to a survey conducted by Ethnologue in 2018, Urdu ranks as the 10th most widely spoken language in the world, with 230 million speakers. However, Urdu is categorized as a low-resource language, and no publicly available dataset exists that comprehensively covers all types of cybercrimes amenable to classification using NLP-based techniques. While Urdu serves as Pakistan's national language, English is the official language, and the informal chats on social media platforms often involve a unique communication dialect known as Roman Urdu. Roman Urdu deviates from conventional grammatical rules and lacks a standard dictionary, making it a challenging language to process. Users employ diverse spellings, self-created abbreviations, slang, acronyms, and non-standard grammar, further complicating the task of cybercrime detection (Rizwan et al., 2020; Saeed et al., 2021; Talpur et al., 2020).

Rizwan et al. (2020) addressed the challenge of hate speech and offensive language in RU. They introduced a lexicon comprising 621 hateful words in Roman Urdu and released the RUHSOLD dataset, containing $10,012$ tweets with coarse-grained and fine-grained labels for offensive language and hate speech. Their work explored the feasibility of transfer learning and proposed a novel deep learning architecture, CNN-gram, for text classification. They concluded that transfer learning is more effective and advantageous for classification than training embeddings from scratch. Saeed et al. (2021) focused on toxic comment classification in Roman Urdu and released the Roman Urdu Toxic (RUT) dataset for further research. Talpur et al. (2020) investigated cyberbullying detection in Roman Urdu. These tasks are relevant to our work, as our primary focus lies in classifying cybercrimes, particularly within the context of Pakistan's cybercrimes law also known as the PECA Act. The severity and nature of cybercrimes exhibit variation across demographic locations; thus, labeled datasets, language resources, and multilingual models are pivotal to advancing research in this domain (Mandl et al., 2019).

3. Cybercrimes in Roman Urdu (CRU)

This section describes the collection and annotation procedure for cybercrimes in Roman Urdu (CRU) dataset. CRU is a collection of various types of cybercrimes according to the PECA. The PECA is passed in 2016 (PECA, 2016) by the parliament of Pakistan, hence act as a standardized cybercrimes law for Pakistan. The importance of the PECA lies in its clear and consistent categorization of different online offenses, such as hate speech, cyber terrorism, electronic fraud, cyberstalking, and cyber harassment. These offenses are not only harmful to individuals and groups but also pose a threat to national security and social harmony. Therefore, it is essential to have a legal framework that can effectively prevent and prosecute such crimes. This law helps us describe cybercrimes and their severity. It is worth noting that the cybercrimes defined in the PECA are comprehensively defined along with their punishments, which is not the case in research articles which just define the cybercrimes at the abstract level.

Cybercrime Definitions In this section, we present the definitions of cybercrimes from the PECA and compare them with those from the research community's definition. We mainly focus on three types of cybercrimes relevant to our task: (1) hate speech, (2) cyber terrorism, and (3) cyber harassment/bullying.

Hate Speech: The PECA defines hate speech as: "any word written or spoken or any gesture made or any visual representation displayed which incites violence or hatred against any group or individual on account of their religion or sect or caste or creed or race or ethnicity or gender or sexual orientation" (PECA, 2016). Under this act, The punishment for hate speech is imprisonment for up to seven years or a fine of up to ten million rupees or both (PECA, 2016). The research community defines hate speech as: "any communication that disparages a person or a group because of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics" (Waseem, 2016; De Gibert et al., 2018; Kovács et al., 2021; Fortuna and Nunes, 2018).

Cyber Terrorism: The PECA defines cyber terrorism as: "any act committed in violation of clause of section 6, sections 7, 8, 9 or 10 of this Act with

the intent to create terror or insecurity in the Government or the public or a section of the public or community or sect or create a sense of fear or insecurity in society" (PECA, 2016). The punishment for cyber terrorism under this act is imprisonment for up to fourteen years or a fine of up to fifty million rupees or both (PECA, 2016). The research community defines cyber terrorism as: “using information technology by terrorist groups and individuals to further their agenda. This can include using information technology to organize and execute attacks against networks, computer systems, and telecommunications infrastructures, or exchanging information or making threats electronically" (Conway, 2003). However, there is no universally accepted definition of cyber terrorism among researchers, and different countries have different legal definitions and regulations for cyber terrorism (Maras, 2017; Varol and Abdulhadi, 2018; Weimann, 2005).

Cyber Harassment: The PECA defines cyber harassment as: “any person who, with malicious intent, knowingly and publicly exhibits, displays, transmits any electronic communication that harms a person’s reputation or subjects a person to ridicule, hatred, embarrassment or dislike" (PECA, 2016). The punishment for cyber harassment under this act is imprisonment for up to three years or a fine of up to one million rupees or both (PECA, 2016). The research community defines cyber harassment as: “the use of information and communication technologies to harass, control, manipulate, or humiliate a person" (Citron and Franks, 2014). However, there is no clear distinction between cyber harassment and other related terms, such as cyberbullying or cyberstalking, among researchers, and different countries have different legal definitions and regulations for cyber harassment (Mishna et al., 2010; Dadvar and de Jong, 2012; Dewani et al., 2023; Talpur et al., 2020).

The PECA plays a crucial role in defining cybercrimes. As researchers navigate the complexities of defining hate speech, PECA offers precise and comprehensive definitions to guide their understanding. For instance, it explicitly addresses characteristics and incitement of violence or hatred. Similarly, PECA’s detailed definitions for cyber terrorism emphasize specific violations and intent. In contrast, the research community’s broader definitions may inadvertently include activities unrelated to terrorism. Furthermore, PECA’s objective characterization of cyber harassment focuses on public display and transmission of harmful electronic communication, while researchers’ subjective definition spans various behaviors. These standardized definitions enhance clarity and facilitate the effective handling of cybercrimes.

Most existing research on hate speech and cyberbullying detection focuses on English or other high-resource languages, leaving a gap for low-resource languages, including Urdu (Ullah et al., 2023). Moreover, most existing datasets for these tasks are either small or need proper annotation guidelines. These challenges limit the development and evaluation of robust and generalizable text classification models for low-resource languages. We present a gold standard dataset CRU, to address these challenges based on the PECA’s definitions and categories. Our dataset is valuable for developing and benchmarking text classification models for low-resource languages, especially for detecting cyber terrorism, hate speech, offensive language, and cyber harassment/cyberbullying.

4. Data Collection and Annotation Challenges

Our dataset, sourced from Twitter, is founded on the legislative framework established by the PECA - Pakistan’s primary cybercrime law. To ensure relevance to contemporary cybercrime trends in Roman Urdu, tweets were harvested from January 1, 2017, to December 30, 2022. Leveraging the Twitter API, we targeted tweets using predefined keywords aligned with cybercrime classifications. We also curated relevant tweets from the RUHSOLD (Rizwan et al., 2020), which served as an additional resource, aligning with the cybercrime categories outlined in the PECA. Specifically, we collected tweets from RUHSOLD related to (1) Hate Speech, (2) Cyber Harassment/Bullying, and (3) Cyber Terrorism. Additionally, we curated a lexicon of offensive terms through online searches and expert interviews, encompassing a spectrum of abusive language, hate speech, cyber harassment, and cyber-terrorism expressions. Using this lexicon and a list of commonly used Roman Urdu words, we systematically searched and collected $11,000$ tweets that met our criteria. A manual preliminary analysis helped discover new slang, identify abusive instances, and recognize frequent common terms. We included common Roman Urdu words to extract random inoffensive tweets and offensive tweets that may not contain explicit, offensive words. To address the issue of user distribution bias, as highlighted by (Arango et al., 2019), we set a limit of a maximum of 50 tweets per user. The following keywords were used to collect tweets:

Cyber Terrorism: "intiqam", "tehreek", "jala do", "fasad", "khilafat", "jihad", "faujiyat", "dehshat-gardi", "badmashi", "ghulam", "samraj", "gernilo", "tattoo", "lanat", "chor", "bajwa", "army", "beghairato", "qatal","maar", "jangh", "beghairat", "gernil", "bangal", "fouj", "jala", "pathloon", "monkhood", "jihad", "ghaddar"

Hate Speech: "shia", "sazish", "mullah", "Ghalatfehmi", "khanzeer", "fahashi", "nafrat-angaiz Taqreer", "yahoodi", "kafir", "shia", "yahodi", "agent", "namak haram", "afghani", "hindu", "india", "kaafir", "randi", "lanat", "halala", "israel"

Cyber Harassment: "ghalat fehmi", "badmashi", "blackmail", "randi", "maa", "baap", "tere" , "tera", "bc", "gashti", "halala", "chod", "aurat", "bhen", "phudi", "abe", "bhenchod", "harami", "bharwe", "saali", "tujhe", "rundi", "yaar", "hijra", "gand", "aulad", "sale", "madarchod", "chutiye", "choot", "gaand"

Our next task after assembling the corpus was annotation. Accurately annotating the cybercrimes tweet is a double-edged sword. On the one hand, there is a thin border between the cybercrime classes, while on the other hand, societal bias plays a negative role in annotating such tweets. Therefore, we need annotators who fully understand the definitions of cybercrimes and up to high instant are neutral (no societal or geographical bias). We consult with personnels who have knowledge of cybercrime law. We follow a rigorous annotation process that involves multiple experts and quality checks. Three expert annotators (with the knowledge of PECA) manually label the tweets to create a standard dataset for this task. To ensure the reliability and validity of the annotation process, we adopted a majority voting scheme to resolve any disagreements among the annotators.

Out of $11,000$ tweets collected from Twitter and RUHSOLD, only 4540 received unanimous labels from all three annotators for the same cybercrime class. Another 2832 tweets received consistent labels from two annotators, and the label from the third annotator was disregarded. Irrelevant tweets containing single words or phrases that did not indicate any cybercrime, such as "Good Night!" or "Feeling Happy," were discarded from our dataset. The final dataset comprises 7372 tweets. The dataset's class distribution is shown in Table 1.

Label	Instances
Cyber Terrorism	250
Hate Speech	474
Cyber Harassment	1105
Normal	2600
Offensive	2943
Total	7372

Table 1: Tweets count with respect to labels.

We perform various statistical analyses to ensure the quality and visualize the main trends in CRU. We remove punctuations from the dataset for reliable analysis. The vocabulary of the CRU is 26495, however, the total tokens are 138407. This shows that each word or term in the dataset appears on average 5 times, indicating the frequency and repetition of the text data. These numbers also reflect the size and diversity of the text data and the implications for text analysis and classification. Total unique unigrams, bigrams, and trigrams are 26495, 110050, and 133814, respectively. These numbers reveal the lexical and syntactic richness of the text data and the difficulty of capturing the context and meaning of the words. To simplify the analysis and visualization of the text data, we removed 138 stop words from the dataset. Stop words are a set of commonly used words in a language that do not add much meaning to a sentence, such as "hai" (is), "tha" (was), "es" (this), etc. By removing these words, we can focus on the most frequent and relevant words in each class, including the whole dataset, and identify the key topics and themes of the text data.

5. Experimental Design

This section outlines the experimental methodologies employed for cybercrime text classification in Roman Urdu, utilizing four pre-trained multilingual transformers.

The models utilized in this study are DistilBERT-Base-Multilingual (Sanh et al., 2019), BERT-Multilingual-Base (Devlin et al., 2019), XLM-RoBERTa-Base (Conneau et al., 2020), and Multilingual-MiniLM (Wang et al., 2020), all accessed through the Hugging-Face[2] library. The experiments encompass both standard fine-tuning and prompt-based techniques, with an emphasis on two prompt engineering methodologies: prefix and cloze prompts.

Pre-trained Multilingual Transformers: Pre-trained multilingual transformers are neural network architectures trained on extensive text corpora from diverse languages, employing self-supervised learning objectives like masked language modeling or next-sentence prediction. These models encode semantic and syntactic information from multiple languages into a unified vector space, facilitating cross-lingual transfer learning and zero-shot learning for various downstream tasks. Given the scarcity of labeled data for low-resource languages like Roman Urdu, pre-trained multilingual transformers offer adaptability and resilience.

Prompt Engineering Techniques Prompt engineering is designing and optimizing natural language prompts to guide pre-trained language models to perform specific tasks without extensive fine-tuning (Liu et al., 2023). Prompt engineering can leverage pre-trained language models' existing

[2] https://huggingface.co/

knowledge (Devlin et al., 2019; Gao et al., 2021) and capabilities to achieve high performance with few (few shot) or no labeled examples (zero-shot). Prompt engineering can also reduce the computational cost and complexity of fine-tuning large models.

Prefix prompts are natural language sentences that precede the input text and contain a placeholder for the desired output (Liu et al., 2023). As depicted in Figure 1, given an input text x = *Sb Afghanio ko Pakistan sey nekalo.* (Get all Afghans out of Pakistan.), a prefix prompt classification could be: x_{prompt} = *[CLS]* x *Ye hai [MASK]! [SEP]* (This is [MASK]!). Cloze prompts are natural language sentences that incorporate the input text and contain a masked token representing the desired output (Liu et al., 2023). For example, given the same input text x = *Sb Afghanio ko Pakistan sey nekalo.* (Get all Afghans out of Pakistan.), a cloze prompt classification could be x_{prompt} = *[CLS]* x *Ye [MASK] hai. [SEP]* (This is [MASK]). In the context of Roman Urdu cybercrime classification, "Ye" and "hai" are commonly used Urdu words meaning "This" and "is," respectively. These words are essential to our Roman Urdu (RU) natural language prompts.

In both cases, the pre-trained language model is expected to fill in the masked token with the correct label for the input text, such as "namak haram" (traitor) or "achey" (good). The choice of prompt shape, wording, and position can affect the model's performance and the task (Gao et al., 2021). Therefore, we experiment with different prompt variations and select the optimal one based on the development set. We employ a top-5 approach for verbalizer selection, selecting the five most likely words or phrases generated by the model that convey coherent and meaningful information.

Experimental Setup To formalize our experimental setup, let L be a pre-trained language model, D be a cybercrime text classification task in Roman Urdu with label space Y, and x_{prompt} be a natural language prompt for D. We aim to fine-tune or guide L on D_{train} using x_{prompt} and evaluate its performance on an unseen test set D_{test}.

Standard Fine-Tuning: Initially, we employ standard fine-tuning procedures to adapt four state-of-the-art transformer models for the task of cybercrime text classification in Roman Urdu. The models utilized include BERT-base, DistilBERT-base, MiniLM, and XLM-RoBERTa. We partition our dataset into an 80-20 split for training and testing, respectively. During fine-tuning, each model is trained on the 80% portion of the dataset, allowing the models to adjust their weights to the specific linguistic patterns and classification tasks presented by the Roman Urdu cybercrime texts. The remaining 20% of the dataset serves as the test set, used to evaluate the models' performance and their ability to generalize to new, unseen data. This approach ensures that the models are not only proficient in the language nuances but also effective in identifying and classifying the targeted cybercrime-related content.

Prompt-Based Fine-Tuning: In the prompt-based fine-tuning approach, we start with 0 and go up to 32 shots. Let K denote the number of training examples per class, and $|Y|$ denote the total number of classes ($|Y| = 5$ in our case) in the task. Thus, the few-shot training set D_{train} consists of $K_{tot} = K \times |Y|$ examples, where $D_{train} = (x_{in}^{(i)}, y^{(i)})_{i=1}^{K_{tot}}$. We also use a development set D_{dev} to select the optimal model, prompt shape, and tune hyper-parameters. The size of D_{dev} set is equal to the few-shot training set, i.e., $|D_{dev}| = |D_{train}|$ for each experiment in our case.

To construct the training and development sets for each dataset, we randomly select K labeled examples from D_{train} for each class, resulting in a total of $Y \times K$ labeled examples, where Y represents the total number of classes (labels) in the dataset. Our experiments consider $K = 0, 4, 8, 16 and 32$. The remaining examples from D are reserved for the test set (with no labels). To ensure fair evaluation, we perform multiple rounds of testing. We randomly select samples from the unlabeled test set not used in each round's training D_{train} and development D_{dev} sets. This process is repeated three times, allowing for a comprehensive evaluation of the models' performance across different test sets. We also use this experimental setup to assess the effectiveness of different prompt shapes in cybercrime detection.

We explore two distinct templates for prefix prompts: (1) x. *yeh hai [MASK]!* (x. *This is [MASK]!*), and (2) x. *Es tweet ka content hai [MASK].* (x. *This tweet's content is [MASK].*) In this format, 'x' represents the input text, and the '$[MASK]$' token serves as a placeholder to be filled in by the model L. This prompts the model to complete the sentence in a way that aligns with the context of the input text. The goal is to select words that belong to specific classes of cybercrime. We explore two distinct templates for cloze prompts: (1) x. *Yeh [MASK] hai* (x. *This is [MASK])*, and (2) x. *Yeh [MASK] tweet hai.* (x.*This is a [MASK] tweet.*). Here, 'x' denotes the input text, and '$[MASK]$' prompts the model to provide a relevant completion, considering the input context.

Combining these prompting techniques and pre-trained transformers allowed us to systematically investigate and optimize cybercrime classification

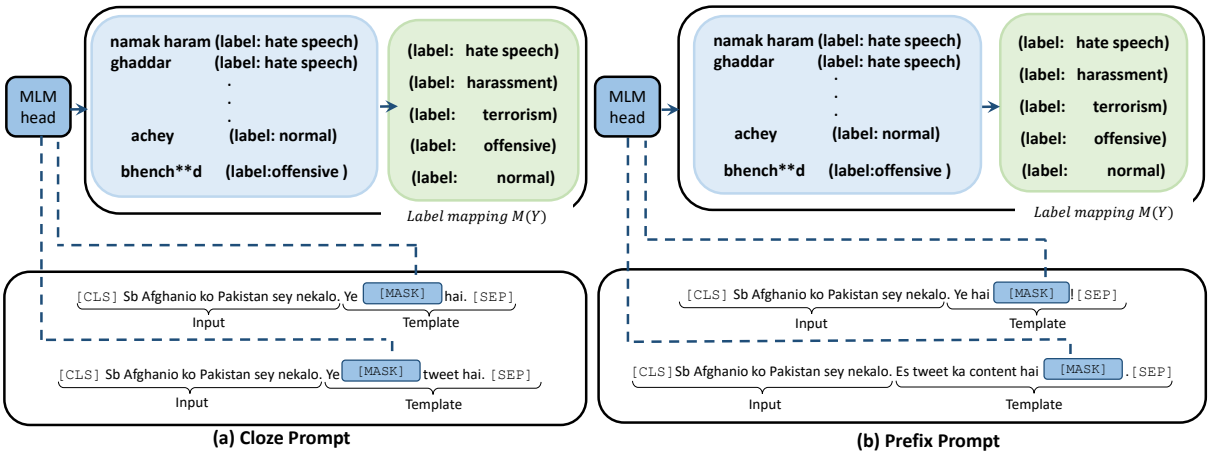

Figure 1: An Illustration of (a) cloze prompts and (b) prefix prompts for cybercrimes text classification in Roman Urdu. The underlined text represents the task-specific template, designed explicitly for Roman Urdu language contexts, while the label mapping block is to map the predicted words to class-specific labels. Translation: [(namak haram: traitor), (ghaddar: traitor), (achey: good), (bhench**d: sisterf****r), (Sb Afghanio ko Pakistan sey nekalo: Get all Afghans out of Pakistan), (ye [MASK] hai: this is [MASK]), (ye ([MASK] tweet hai: this is a [MASK] tweet), (ye hai [MASK]!: this is [MASK]!), (Es tweet ka content hai [MASK]: the content of this tweet is [MASK])"]

performance, contributing to advancing methods for detecting cybercrimes, particularly within the context of the Roman Urdu language (Ullah et al., 2023). The OpenPrompt[3] library has been utilized for the comparison of these prompt shapes. OpenPrompt is an open-source framework designed explicitly for prompt learning, providing a comprehensive set of tools and resources for this approach. Our standard fine-tuning and prompt-based fine-tuning process utilizes a learning rate of $2e-5$. The optimization method is AdamW (an Adam optimizer variant), and the loss function is Cross-Entropy Loss. The number of epochs for each training is 10.

6. Result and Discussion

This section presents the results of four pre-trained multilingual transformers and compares prefix and cloze prompt shapes for Roman Urdu cybercrime text classification.

In Roman Urdu cybercrime detection, we evaluated the performance of four pre-trained multilingual transformers using various metrics. Table 6 shows the results of these PLMs on precision, recall, macro F1-score, and accuracy. Among the models, xlm-roberta-base achieved the highest score of 74%, 72%, 73% and 77% for precision, recall, macro F1-score, and accuracy, respectively, followed by bert-base-multilingual-cased. The distilbert-base-multilingual-cased performed slightly lower than the other models, with precision at 65%, recall at 61%, F1-score at 63%, and accuracy at 72%. These results suggest that xlm-roberta-base is the most effective model for cybercrime detection in Roman Urdu, while distilbert-base-multilingual-cased is the least effective among the models.

Model	Prec.	Rec.	F1	Acc.
bert-base	0.73	0.69	0.71	0.76
distill-bert-base	0.65	0.61	0.63	0.72
Mini-LM	0.71	0.65	0.67	0.74
xlm-roberta	**0.74**	**0.72**	**0.73**	**0.77**

Table 2: Pretrained multilingual transformers results using standard fine-tuning.

We also compare two types of prompt engineering techniques, prefix and cloze prompts, using xlm-roberta-base and bert-base-multilingual-cased. We analyzed the impact of prompting techniques on different k-shot settings, where k denotes the number of training examples per class.

The Table3 shows the results of the prompting experiments, using xlm-roberta-base and bert-base-multilingual-cased as base models, and applying different types of prompts and other numbers of examples (k-shots) for each category. The table reports the precision, recall, F1-score, and accuracy of each model and prompt type for four types of prompts: (Roman Urdu : English) (Yeh [MASK] hai : this is [MASK]), (Yeh [MASK] tweet hai : this is a [MASK] tweet), (Yeh hai [MASK]! : this is [MASK]!), and (Es tweet ka content hai [MASK] : the content of this tweet is [MASK]). This also shows the results of the zero-shot classification, which does not

[3] https://thunlp.github.io/OpenPrompt/

k-shots	xlm-roberta				bert-base			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
(Yeh [MASK] hai)								
0	0.35	0.2	0.11	0.35	0.15	0.23	0.11	0.19
4	0.29	0.36	0.24	0.29	0.28	0.35	0.25	0.29
8	0.34	0.38	0.30	0.36	0.33	0.39	0.33	0.36
16	**0.38**	0.47	0.35	**0.39**	0.36	0.47	0.37	0.41
32	**0.38**	**0.49**	**0.37**	0.38	**0.40**	**0.53**	**0.42**	**0.44**
(Yeh [MASK] tweet hai)								
0	0.28	0.2	0.11	0.35	0.2	0.23	0.13	0.3
4	0.34	0.38	0.27	0.33	0.3	0.38	0.28	0.32
8	0.33	0.44	0.3	0.31	0.34	0.42	0.34	0.36
16	0.36	0.48	0.36	0.38	0.36	0.47	0.36	0.39
32	**0.40**	**0.53**	**0.39**	**0.41**	**0.39**	**0.53**	**0.40**	**0.42**
(Yeh hai [MASK]!)								
0	0.33	0.21	0.13	0.35	0.22	0.25	0.11	0.19
4	0.32	0.43	0.3	0.33	0.29	0.38	0.26	0.29
8	0.35	0.46	0.35	0.36	0.34	0.41	0.34	0.35
16	0.38	0.49	0.37	0.40	0.39	0.48	0.40	0.43
32	**0.40**	**0.52**	**0.39**	**0.41**	**0.42**	**0.52**	**0.42**	**0.45**
(Es tweet ka content hai [MASK])								
0	0.26	0.2	0.11	0.35	0.21	0.24	0.13	0.18
4	0.31	0.39	0.27	0.29	0.29	0.36	0.28	0.31
8	0.33	0.44	0.32	0.34	0.34	0.43	0.34	0.35
16	**0.41**	0.48	0.39	0.44	**0.40**	0.44	**0.37**	**0.39**
32	0.40	**0.53**	**0.42**	**0.46**	0.38	**0.47**	0.36	0.38

Table 3: Prompting Results for k-shots. The top two subsections denote results for cloze prompts, while the bottom two subsections present the results for prefix prompts.

use any standard or prompt-based fine-tuning but only relies on the pre-trained models to classify the tweets.

Based on the Table3, the main findings are: (1) xlm-roberta-base outperforms bert-base-multilingual-cased in all settings, confirming its dominance for this task; (2) prefix prompts perform better than cloze prompts, especially in low k-shot settings, suggesting their suitability for Roman Urdu classification tasks; (3) the accuracy and the F1-score of both models and both prompt types increase as the number of examples (k-shots) increases; this suggests sufficient training data coupled with an effective prompt shape can yield highly accurate cybercrime classification results, hence, prompting techniques are better for under-resourced languages like Roman Urdu, and (4) among the four types of prompts, the (Yeh hai [MASK]! : this is [MASK]!) prompt and the (Es tweet ka content hai [MASK] : the content of this tweet is [MASK]) prompt achieve the better results for both models and both prompt types, followed by the (Yeh [MASK] tweet hai : this is a [MASK] tweet) prompt and the (Yeh [MASK] hai : this is [MASK]) prompt, implying that the prompt shape and the language structure influence the performance of the models.

It is worth noting that in standard fine-tuning, we use 80% of the data for training and 20% for testing. In prompt-based fine-tuning, however, we use only 0, 4, 8, 16, or 32 (shots) samples from each class for training and the rest for testing. This means that prompt-based fine-tuning has much less data to learn from, which could limit its ability to generalize and adapt to the task. We do not claim that our prompts are optimal, and we acknowledge that there may be other ways to design better prompts that could improve the accuracy. However, our main contribution is showing that prompt-based fine-tuning can achieve reasonable results with very few samples, a valuable option for low-resource languages that need more labeled data.

The bert-base-multilingual-cased substantially increases precision, recall, macro F1-score, and accuracy from 0 to 32 shots. It achieves the highest overall performance among the models at 32 shots. This demonstrates the effectiveness of bert-base-multilingual-cased for the specific task of Roman

Urdu cybercrime detection. The prefix prompts, which involve filling a slot at the last of the template text, encourage the models to generate more contextually relevant completions, resulting in improved classification performance. Cloze prompts, are less effective because they provide prompt structures that could be challenging for PLMs to predict the specific word in RU. Based on a comparative evaluation of different prompt shapes and k-shot settings on four pre-trained language models, we conclude that prefix prompts are suitable for the task. The results show that prefix prompts consistently outperform cloze prompts in Roman Urdu classification, regardless of the base model used. We also claim that the prefix prompts achieve better results, especially in low k-shot settings. However, this comparison could be fairer and more convincing, as the zero-shot classification is a much more complex and unrealistic scenario, and the difference in accuracy is not very significant.

The study conducted a comprehensive assessment of four pre-trained multilingual transformers for the classification of cybercrime text in Roman Urdu. The analysis involved standard fine-tuning with an 80-20 train-test split and prompt-based fine-tuning across a range of shots (0, 4, 8, 16, and 32). Model performance was evaluated using accuracy, precision, recall, and F1-score metrics. According to the results presented in Table 2, the xlm-roberta-base model achieved the highest scores across all metrics under the standard fine-tuning technique, indicating its effectiveness for the specified classification task. In contrast, Table 3 shows that the BERT-base model performed best with prompt-based fine-tuning, adapting well to different prompt templates and shapes. The results underscore the importance of prompt engineering for enhancing the accuracy and efficiency of classification models in the context of the Roman Urdu language.

7. Conclusion and Future Work

In this paper, we develop a benchmark dataset covering diverse categories of cybercrimes in Roman Urdu within the legal context of Pakistan. Subsequently, we evaluate the performance of four fine-tuned multilingual transformers for classifying cybercrime texts in Roman Urdu. Our findings reveal that xlm-roberta-base is the proficient model for this task. Furthermore, we compare two prompt engineering techniques, prefix and cloze prompts, utilizing xlm-roberta-base and bert-base-multilingual-cased as base models. The study uncovers the influence of these prompting techniques across various k-shot settings, demonstrating that prefix prompts outperform cloze prompts in Roman Urdu cybercrime classification.

This research holds implications for cybercrime detection and text classification, especially in low-resource languages. It underscores the potential of pre-trained multilingual transformers in addressing the challenge of cybercrime text classification in Roman Urdu. It also highlights the effectiveness of prompt engineering techniques in such low-resource language scenarios. Additionally, it emphasizes the pivotal role of prompt shape, k-shot settings, language models, and language structure in influencing the approach's performance. This work sets the stage for future research, including expanding it to other low-resource languages and domains like hate speech detection and sentiment analysis.

8. Limitations

Our work has some limitations that need attention in future work. Our dataset is relatively small compared to other datasets for text classification tasks in other languages. Thus, the generalizability and robustness of the approach need further investigation. Our approach relies on pre-trained multilingual transformers that may not capture the nuances and specificities of the Roman Urdu language and culture. This may lead to some errors or biases in our results. Our approach assumes that the input texts are well-formed and grammatically correct. However, this may differ for real-world data, especially for informal texts such as tweets or comments. This may affect the performance of our approach, as the pre-trained multilingual transformers and the prompt engineering techniques may not be capable of handling noisy or unstructured data. Our approach does not incorporate domain knowledge or external resources for cybercrime detection and text classification in Roman Urdu. For example, some cybercrimes may involve specific terms or concepts that need to be better represented by pre-trained multilingual transformers or prompt engineering techniques. This may require additional information or guidance for the models to perform better.

9. Ethical Considerations

We acknowledge that our work has ethical implications for both the research community and society. On the positive side, our work contributes to developing natural language processing tools and resources for Roman Urdu, addressing a pressing social problem of cybercrime detection, which can help protect the rights and safety of millions of people who use online platforms for communication and information. Our work can assist law enforcement agencies, policymakers, and civil society organizations in combating cybercrimes and

enforcing the legal framework of Pakistan.

On the flip side, our work may also pose some risks and challenges. Our dataset contains sensitive information about real cybercrime cases, which may expose the victims' and perpetrators' identities and personal details, even though we have taken measures to anonymize the data and remove any identifiable information.

10. Bibliographical References

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 45–54.

Muhammad Bilal, Atif Khan, Salman Jan, Shahrulniza Musa, and Shaukat Ali. 2023. Roman urdu hate speech detection using transformer-based model for cyber security applications. *Sensors*, 23(8):3909.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Danielle Keats Citron and Mary Anne Franks. 2014. Criminalizing revenge porn. *Wake Forest L. Rev.*, 49:345.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Maura Conway. 2003. Cyberterrorism: the story so far. *Journal of Information Warfare*, 2(2):33–42.

Maral Dadvar and Franciska de Jong. 2012. Cyberbullying detection: A step toward a safer internet yard. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, page 121–126, New York, NY, USA. Association for Computing Machinery.

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.

Ona De Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Amirita Dewani, Mohsin Ali Memon, and Sania Bhatti. 2021. Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for roman urdu data. *Journal of big data*, 8(1):160.

Amirita Dewani, Mohsin Ali Memon, Sania Bhatti, Adel Sulaiman, Mohammed Hamdi, Hani Alshahrani, Abdullah Alghamdi, and Asadullah Shaikh. 2023. Detection of cyberbullying patterns in low resource colloquial roman urdu microtext using natural language processing, machine learning, and ensemble techniques. *Applied Sciences*, 13(4):2062.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Edel Greevy and Alan F. Smeaton. 2004. Classifying racist texts using a support vector machine. In *Proceedings of the 27th Annual International*

ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '04, page 468–469, New York, NY, USA. Association for Computing Machinery.

Ihsan Ullah Khan, Aurangzeb Khan, Wahab Khan, Mazliham Mohd Su'ud, Muhammad Mansoor Alam, Fazli Subhan, and Muhammad Zubair Asghar. 2021. A review of urdu sentiment analysis with multilingual perspective: A case of urdu and roman urdu language. *Computers*, 11(1):3.

György Kovács, Pedro Alonso, and Rajkumar Saini. 2021. Challenges of hate speech detection in social media. *SN Computer Science*, 2(2):1–15.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Twenty-seventh AAAI conference on artificial intelligence*.

Dun Li, Kanwal Ahmed, Zhiyun Zheng, Syed Agha Hassnain Mohsan, Mohammed H Alsharif, Myriam Hadjouni, Mona M Jamjoom, and Samih M Mostafa. 2022. Roman urdu sentiment analysis using transfer learning. *Applied Sciences*, 12(20):10344.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th forum for information retrieval evaluation*, pages 14–17.

M.H. Maras. 2017. *Cybercriminology*. Oxford University Press.

Faye Mishna, Charlene Cook, Tahany Gadalla, Joanne Daciuk, and Steven Solomon. 2010. Cyber bullying behaviors among middle and high school students. *American journal of orthopsychiatry*, 80(3):362–374.

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.

PECA. 2016. The Prevention of Electronic Crimes Act, 2016.

Hammad Rizwan, Muhammad Haroon Shakeel, and Asim Karim. 2020. Hate-speech and offensive language detection in Roman Urdu. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2512–2522, Online. Association for Computational Linguistics.

Hafiz Hassaan Saeed, Muhammad Haseeb Ashraf, Faisal Kamiran, Asim Karim, and Toon Calders. 2021. Roman urdu toxic comment classification. *Language Resources and Evaluation*, 55(4):971–996.

Tauqeer Sajid, Mehdi Hassan, Mohsan Ali, and Rabia Gillani. 2020. Roman urdu multi-class offensive text detection using hybrid features and svm. In *2020 IEEE 23rd International Multitopic Conference (INMIC)*, pages 1–5.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Mobeen Shahroz, Muhammad Faheem Mushtaq, Arif Mehmood, Saleem Ullah, and Gyu Sang Choi. 2020. Rutut: roman urdu to urdu translator based on character substitution rules and unicode mapping. *IEEE Access*, 8:189823–189841.

KR Talpur, SS Yuhaniz, NNBA Sjarif, and B Ali. 2020. Cyberbullying detection in roman urdu language using lexicon based approach. *J. Crit. Rev*, 7(16):834–848.

Faizad Ullah, Ubaid Azam, Ali Faheem, Faisal Kamiran, and Asim Karim. 2023. Comparing prompt-based and standard fine-tuning for Urdu text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6747–6754, Singapore. Association for Computational Linguistics.

Cihan Varol and Hezha M. Tareq Abdulhadi. 2018. Comparision of string matching algorithms on spam email detection. In *2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT)*, pages 6–11.

Shiguo Wang. 2010. A comprehensive survey of data mining-based accounting-fraud detection research. In *2010 International Conference on Intelligent Computation Technology and Automation*, volume 1, pages 50–53.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788. Curran Associates, Inc.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

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, Vancouver, BC, Canada. Association for Computational Linguistics.

Gabriel Weimann. 2005. Cyberterrorism: The sum of all fears? *Studies in Conflict & Terrorism*, 28(2):129–149.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.