Detecting Offensive Language in an Open Chatbot Platform

Hyeonho Song[1], **Jisu Hong**[2], **Chani Jung**[1], **Hyojin Chin**[3], **Mingi Shin**[1]
,**Junghoi Choi**[4], **Yubin Choi**[1], **Meeyoung Cha**[1,3]
[1]KAIST School of Computing, [2]Seoul National University, [3]Institute for Basic Science, [4]SimSimi Inc.
[1,3]Daejeon Korea, [2,4]Seoul Korea
{hyun78.song, hellojisuworld, 1016chani, tesschin, mingi0116s,
sijay00, joyda2525, meeyoung.cha}@gmail.com

Abstract

While detecting offensive language in online spaces remains an important societal issue, there is still a significant gap in existing research and practial datasets specific to chatbots. Furthermore, many of the current efforts by service providers to automatically filter offensive language are vulnerable to users' deliberate text manipulation tactics, such as misspelling words. In this study, we analyze offensive language patterns in real logs of 6,254,261 chat utterance pairs from the commercial chat service Simsimi, which cover a variety of conversation topics. Based on the observed patterns, we introduce a novel offensive language detection method—a contrastive learning model that embeds chat content with a random masking strategy. We show that this model outperforms existing models in detecting offensive language in open-domain chat conversations while also demonstrating robustness against users' deliberate text manipulation tactics when using offensive language. We release our curated chatbot dataset to foster research on offensive language detection in open-domain conversations and share lessons learned from mitigating offensive language on a live platform.

Keywords: Offensive language detection, Contrastive learning, Chatbot, Open-domain conversations

Disclaimer: Examples of offensive chat in this study may include slang, sexually explicit language, profanity, and hate speech.

1. Introduction

Chatbots are computer programs that interact with humans using natural language (Bae Brandtzæg et al., 2021). Since Weizenbaum introduced the ELIZA chatbot in the 1960s (Weizenbaum, 1966), conversational agents have evolved from simple rule-based algorithms to advanced deep neural networks. Recently, with the emergence of large language model applications like OpenAI ChatGPT, Google Bard, and Microsoft Bing, chatbot usage has skyrocketed worldwide.

Some research suggests that conversations with chatbots have fewer barriers than interacting with humans, as users can enjoy the benefits of anonymity and privacy (Bae Brandtzæg et al., 2021). Some people even perceive the service as a "safe zone" where they can freely express themselves and discuss any topic without fear of being judged (Ta et al., 2020). Therefore, due to the anonymity they provide, chatbots are seen as potentially advantageous for individuals who are vulnerable, including those dealing with mental health issues (Lucas et al., 2017).

Although chatbots offer many potential benefits, there are worries that user anonymity may lead to an increased use of harmful language compared to interactions with other people. According to Hill et al. (2015), the level of profanity in human–chatbot conversations is nearly 30-fold more prevalent on average compared to human-human conversations. Studies have found that toxic language, including sexually explicit expression, accounts for a significant (e.g., 10–44%) portion of human–chatbot interaction (Angeli and Brahnam, 2008; Veletsianos et al., 2008).

If not addressed, offensive language towards chatbots may become unintentionally accepted, potentially spreading to human-human conversations (Lima et al., 2020; Chin et al., 2020). Furthermore, it can harm the chat system itself as Chin et al. (2020) pointed out that misuse and abuse of chatbots will exacerbate negative user experiences on the platform. Chatbots may also learn offensive behavior from training data and respond to users with toxic content.

As a result, extensive efforts are being made to address the problem of regulating toxic behavior in text generation (Kwak et al., 2022; Gehman et al., 2020). Nevertheless, identifying toxic content within generated sentences or training data remains a huge challenge. This is because there are many different types of toxic content, such as speech that contains hurtful, derogatory, obscene, offensive, profane, or hateful content, and they are topic- and context- dependent.

In addition, there is a gap in existing research and practical datasets dedicated to chatbots. In one study (Khatri et al., 2018), offensive content in chatbot-human interactions was categorized into different labels, including inappropriate, insulting, profane, and sexual language (Ram et al., 2018). This dataset comprises 1,517 manually annotated utterances. In another investigation, the ConvAbuse corpus was introduced, containing 20,000 annotated utterances and offering a detailed breakdown of various types of abusive language (Cercas Curry et al., 2021).

Our research expands on prior efforts by specifically addressing offensive language within chatbots, using a large dataset of 6,254,261 utterance pairs from real conversations. For this, we collaborated with SimSimi (`https://simsimi.com/`), an open-domain chat platform running in 81 languages that has been one of the world's longest-running chatbot services since 2002. Based on the logs, we examined which topics incur a higher prevalence of offensive language and identified frequent attempts by users to circumvent the automated filters when employing offensive language. For instance, users intentionally misspelled offensive words, such as "`c0ward`" instead of "`coward`", to bypass the filters, which diminished the performance of dictionary-based detection methods.

Based on the frequent misspelling practice, we devised a contrastive learning-based approach to train a context-aware offensive language detection model by random masking. Experiments demonstrate that the proposed loss design indicates marked improvement over baseline models in detecting offensive language in unstructured chat conversations. In terms of conversational topics related to offensive language, our data analysis also reveals that certain topics, such as sex, exhibit higher frequencies of offensive language than other topics, such as politics.

Our main contributions are as follows:

- We present a large-scale analysis on the use of offensive language in a real-world, open-domain chatbot, leveraging a vast dataset comprising 6,254,261 pairs of user utterances and responses.

- We propose a novel masking-based contrastive learning detection model tailored to learning the offensive contexts embedded in open-domain chat conversation texts. The proposed model is designed to be resilient to users' intentional misspelling tactics of offensive words.

- We release the code (`https://github.com/hyun78/coling2024_simmask`) and dataset [1] publicly to foster community collaboration and subsequent research.

2. Related Work

Efforts to detect offensive language Offensive language is defined as hurtful, derogatory, or obscene comments (Wiegand et al., 2018). Scholars have taken various approaches to address offensive usage in online user-generated content. For example, regarding studies on offensive language online, various platforms have been examined, including social media in general (MacAvaney et al., 2019), Facebook posts (Kumar et al., 2018), Twitter tweets (Zampieri et al., 2019; Rosenthal et al., 2021; Mandl et al., 2019), and Wikipedia comments section.

Regarding detection methods, a variety of classifiers have been applied for identifying offensive language, and these automated detection approaches have demonstrated promising results. For example, several studies have suggested detection models utilizing various techniques such as linear classifiers (Malmasi and Zampieri, 2017), deep neural networks (Aluru et al., 2021), transfer learning (Wiedemann et al., 2020), and pre-trained language models (Liu et al., 2019). A recent study even proposed the use of large language models to identify hateful content (Wang and Chang, 2022).

As chatbot agents like ChatGPT become more popular, researchers have been paying increasing attention to the problem of offensive language detection in chatbots. For instance, some of this research has focused on how to control the generative models to prevent them from producing offensive language (Kwak et al., 2022; Gehman et al., 2020). Others try to use large language models for detecting offensive languages (Li et al., 2023; Nguyen et al., 2023). Our research aligns with detecting offensive language while we focus on large amounts of user-chatbot conversations.

Contrastive learning Contrastive learning is a popular representation learning method that brings positive pair embeddings closer together while pushing negative pair embeddings apart. It has been successful in computer vision (Chen et al., 2020; He et al., 2020; Grill et al., 2020; Chen and He, 2021) and has extended its application across several domains. Many studies have attempted to enhance the universal sentence-level semantics of their representations by pretraining language models such as BERT (Kenton and Toutanova, 2019) on the sentence level using contrastive learning (Fang et al., 2020; Giorgi et al., 2021). There also exist some studies that utilize contrastive learning in a multi-modal manner to detect offensive content (González-Pizarro and Zannettou, 2023; Shome and Kar, 2021). Recently, SimCSE (Gao et al., 2021) introduced an effective text augmentation method that generates positive samples by sending the same input through the

[1] Link to request the chatbot dataset at `https://blog.naver.com/simsimi_official/222833955785`

encoder twice, utilizing a random dropout mask in the encoder. DiffCSE (Chuang et al., 2022) demonstrated that exploiting the property of the masked language model can be helpful in representation learning. Kim et al. (2022) proposed a contrastive learning-based generalizable hate speech detection method, which implies the contrastive approach can be effective in generalizable and robust detection. As another example, Lu et al. (2023) proposed contrastive-learning-based hate speech detection method, which uses simple dropout noise to make positive pairs. However, it's worth mentioning that these offensive language detection methods based on contrastive learning may not effectively identify offensive language that has been intentionally misspelled to bypass existing filters. Therefore, we introduce a contrastive learning-based model for detecting offensive language, specifically engineered to exhibit greater resilience in the face of intentional misspelling.

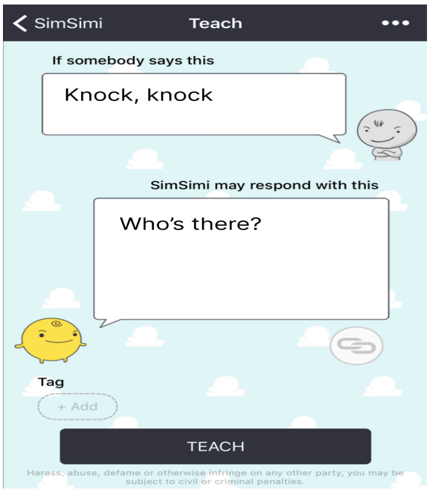
3. Open Chat Data

3.1. SimSimi Platform

SimSimi is an open-domain chatbot platform launched in 2002 with the primary goal of engaging in small talk with users. SimSimi is a worldwide service servicing millions of users. SimSimi has relied on crowdsourcing to accumulate conversational knowledge since the beginning of the service. This service provides a teaching feature that allows users to input pairs of conversational exchanges, with each pair consisting of a question and its corresponding answer. Figure 1 shows an example of the teaching feature. SimSimi responds to user utterances by selecting the most appropriate chat answer from 141.6 million pairs of conversations taught by their users based on text similarity and internal context embeddings. Under research collaboration, the authors accessed a collection of random chat utterances on SimSimi. Please see our ethical statement on preserving users' privacy protection.

Data types For topical consistency, we prepared two datasets in English that span a similar period between 2019 and 2021 for analysis (Table 1).

- The Teach dataset is conversation pairs that use the "Teach" function, having crowdsourced labels, where SimSimi explicitly asks users how to respond to a question. This Teach dataset is also labeled by crowdsourcing, determining whether the conversation pairs contain offensive language, promote violence, or promote hatred. Labeling was only available to users who passed the attention check questions.



If somebody says this

Knock, knock

SimSimi may respond with this

Who's there?

Harass, abuse, defame or otherwise infringe on any other party, you may be subject to civil or criminal penalties.

Figure 1: Teaching example on Simsimi. Users can instruct the chatbot on how to respond to a specific query. SimSimi also asks users if a given sentence is offensive. In return, users receive service credit. These crowdsourced labels are used in this research.

Data	Count	Q.Len	A.Len	Users
Teach	40,152	15.30	23.80	19,825
Live	6,254,261	15.53	17.84	168,305

Table 1: Data statistics for the number of utterance pairs, average query length (Q.Len) and average answer length (A.Len) each utterance, and the number of unique chat participants.

- The Live dataset is conversation pairs that naturally arise between SimSimi and its users. This dataset contains no labels.

SimSimi Inc. was granted a worldwide license (the "IP License") to use IP Content posted by SimSimi users through "Terms and Conditions."[2] The data provided to researchers does not contain any personally identifiable information (PII) such as names, gender, age, address, phone numbers, locations, and social security numbers. SimSimi does not collect or store PII and has an internal logic that protects user privacy by filtering out PII revealed during chat conversations.

3.2. Data Processing

Offensiveness annotation To evaluate the offensiveness of their data, Simsimi Inc. employs data annotation methods that involve voluntary contributions from Simsimi users. The criteria for the offensiveness labels are outlined in the terms of the SimSimi Policy on Rights and Responsibilities (see section 2.1.1 through 2.1.7 of the policy document) These criteria are in the same context as the definition of offensive language we had described earlier.

[2] http://bit.ly/3QT327n.

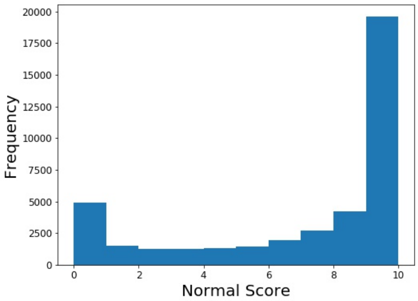


Figure 2: Crowd label distribution on the toxicity of randomly selected chat utterances of Teach dataset. Scores represent the responses of non-toxic labels out of ten labelers.

The annotation was performed within the application. First, annotating users were given a detailed explanation of the labeling criteria. Next, the users labeled a set of five data samples as either offensive or not offensive. Each set included two attention check samples, which SimSimi confirmed as true positives and negatives. Using these attention check samples, the server rejected users who provided incorrect answers to maintain annotation quality. At the end of the annotation, participants received rewards that could be spent on the application in return. Each utterance was labeled by ten users.

We defined the 'normal score' for each sentence as the count of times it was labeled 'non-offensive' by a sample of ten users. Figure 2 presents the crowdsourced label distribution from the Teach dataset. The distribution of the obtained class labels is 74.62% normal (non-offensive) and 25.38% offensive, with the normal score threshold as 5.

3.3. Topic Distribution and Offensive Language in Chat Conversations

To determine the representative conversation topics to observe the use of offensive language across topics, we examined utterances in the Live dataset, which are from a natural chat environment. We define *sessions* as the consecutive conversations between a user and SimSimi. A session s with length N is a set of consecutive pairs of utterances(p) for a user i ($s = \{p^i_j\}^N_{j=1}$). The maximum time gap between two utterances in a session, p^i_j and p^i_{j+1}, is set as 30 minutes, after which the conversation is considered a new session. A total of 262,629 sessions were found.

Next, we define *context session* to account for topic shifts within a session. If the word or n-gram w^t of topic t occurred in a sentence pair p_j in session s, we defined *context session* $C = \{p_{j-k}, p_{j-k+1}, \cdots, p_{j-1}, p_j, p_{j+1}, \cdots, p_{j+k}\}$ with window size k. Then, if two consecutive context sessions contained the same topics, they were merged. Setting the window size k to 5, a context session contained an average of 18.3 utterance pairs. A total of 561,878 context sessions were retrieved.

We started with 12 topics: sports, food, sci&tech, business, health, fashion, entertainment, politics, news, music, books, and COVID-19. Keywords for extracting chats by topics were compiled from the literature (Hong and Davison, 2010; Alvarez-Melis and Saveski, 2021; Nguyen and Shirai, 2015; Hsieh et al., 2012; Abbar et al., 2015; Vidal et al., 2015; Mao et al., 2011; Yang et al., 2014; Paul and Dredze, 2014; Zhao and Min, 2019; Ma et al., 2019; Al-Rawi, 2019). We augmented this list by adding two additional topics that were prevalent from observing the top-200 n-grams for n=1 to 4: *sexually-explicit* and *small talk with SimSimi*. For example, utterances like "How old are you?" and "What do you like?" were labeled as the small talk category.

Having identified topically coherent chats as context sessions, we could see that the distribution was skewed across topics: sex-related talk (accounting for 47.9% of all context sessions), small talk with SimSimi (20.5%), food (9.8%), music (8.1%), sports, science & technology, business, health, fashion, entertainment, politics, news, books, and COVID-19. Interestingly, this finding aligns with previous studies demonstrating the prevalence of sexual topics in online conversation with social agents (Angeli and Brahnam, 2008; Veletsianos et al., 2008).

The Live dataset includes offensive language. Three types of offensive language, 1) insults (including swear words), 2) hate speech, and 3) sexually explicit language, were dominant in SimSimi conversations. Each type is defined as below.

- Insulting: Language that disrespects and denies an individual's normal attributes and abilities (Chin et al., 2020) (Query: "You idiot!", Answer: "Lol").

- Hate speech: Language that attacks, denigrates, or incites violence or hatred against specific groups based on characteristics (Fortuna and Nunes, 2018) (Query: "N*gga" Answer: "You black not me").

- Sexually Explicit: Language that contains references to sexual acts, body parts, or other obscene content[3] (Query: "fck yeah daddy " Answer: "any time sweet @ss")

In these offensive sessions, sexually explicit topics, where users initiate role-playing in a sexual context, were the most prevalent. Based on

[3] https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages

manual inspection, we learned that certain topics (*sexually-explicit*, fashion, and health) are associated more with offensive language.

3.4. Intentional Misspelling Behavior

We also found that misspellings, coded words, and intentional typos are frequent in the studied open domain chatbot, which may be challenging for dictionary-based detection algorithms. Here, we use the term *attack query* to describe instances where users intentionally misspell or introduce typos in an attempt to circumvent the automatic offensive language detection algorithm.

From the 262,629 sessions with 6,254,261 conversation pairs, we try to find such attempts by the following process. First, we found 319,059 conversation pairs (5.1%) with attack query and avoidance answer (i.e., "I will not answer that sentence"), where the chatbot tries to avoid the conversation based on SimSimi Inc. policy when facing inappropriate text. These attack queries are identified by rule-based answers. Second, we collect up to 3 consecutive queries after the avoidance response and measure the edit distance (Levenshtein et al., 1966) of the original attack query (i.e., suck) and modified queries (i.e. sûck). Lastly, we measure the attack success rates if the chatbot replied with any other meaningful answer rather than rule-based avoidance, while the modified queries have a smaller edit distance than 3.

There were 26,930 attack attempts (8.4%), and attack success rates were 73.55%, indicating the presence of intentional misspelling behavior in chatbot conversations. In the next section, we highlight this problem and suggest a framework for learning the offensive context that users may modify to circumvent the offensive language filters of systems.

4. Detection Model

4.1. Contrastive Learning Loss Design

We present the Simple effective Masking framework (SimMask) for detecting offensive language (e.g., swear words, sexually explicit, and hate symbols) leveraging the high representation power of contrastive learning. The underlying concept is derived from the observation that individuals use typos or coded words to avoid censorship (Magu and Luo, 2018). Such a simple trick is effective at deceiving not only automatic detection filters but also deep neural networks (Gröndahl et al., 2018). We recognize that the offensive nuance or tone of the entire sentence remains unchanged even when the modified word may no longer seem non-offensive. Thus, we anticipate that an effective offensiveness detection model should be able to learn the offensive nuance or tone without explicit keywords.

To handle this, we employ random masking to generate positive samples without explicit keywords and suggest three possible masking strategies and corresponding experimental results in section 4.2. We expect this method can effectively detect the offensive language used by the user side.

We use a contrastive learning framework to learn offensive nuances or tones in a sentence. In general, many contrastive learning frameworks define positive and negative samples to make successful representations. To define a positive sample x^+ for a given sentence x, we augment the data with *random masking*, which masks words in the original sentence. Let the given sentence $x = \{w_1, w_2, \cdots, w_L\}$ have length L and word tokens w_i. Then, we randomly replace each word with a special token **[MASK]** with probability r. The masking ratio r can be a constant or varied depending on the masking strategy tokenization results. We use negative samples from the same batch instances, which are unrelated to target instance x (Figure 3b).

The sentence embedding $\mathbf{z} = h(f(x))$ can be obtained by using the encoder f and projection head h, as shown in Figure 3a. The encoder can be any architecture that can generate sentence embedding, such as a BERT. Given a training batch $\{x_i\}_{i=1}^N$, the contrastive loss is defined as follows:

$$\mathcal{L}_{\text{cont}} = -\log \frac{e^{\text{sim}(\mathbf{z}_i, \mathbf{z}_i^+)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau}},$$

where $\text{sim}(\cdot,\cdot)$ is the cosine similarity function, and τ is the temperature parameter. In addition, we suggest jointly training with a supervised loss for better prediction quality. The final objective is:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \lambda \mathcal{L}_{\text{cont}},$$

where $\mathcal{L}_{\text{ce}}$ is the cross entropy loss and λ is the hyperparameter to adjust two objectives. For the cross-entropy loss, our model has a separate linear layer that inputs sentence embedding and outputs the number of classes (Figure 3).

4.2. Masking strategies

Masking enables the model to learn context more effectively. The masking process teaches the model to learn the behavior of coded words, intentional typos, or indirect speech. We present three possible masking word selection strategies:

- **Strategy 1. Mask offensive words.** We mask potentially offensive words. Various algorithms or existing dictionaries may be used to predefine the offensive words. The masking probability is set to r for offensive words, while other words are not masked.

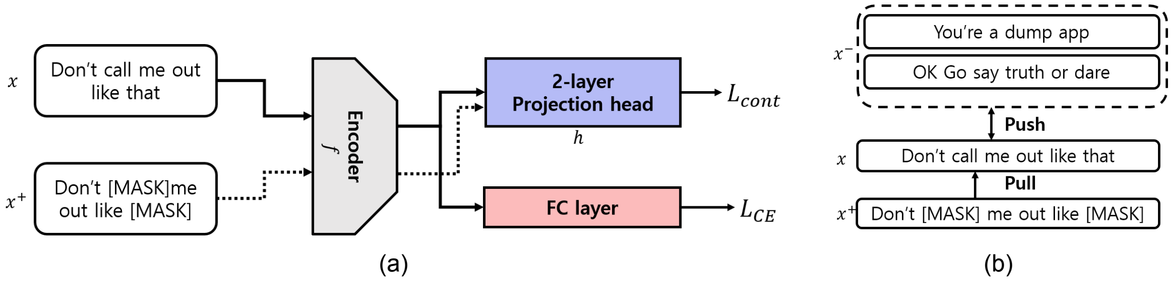


Figure 3: Overview of SimMask framework and contrastive loss. 3a: Original text(x) and its label used to calculating $\mathcal{L}_{CE}$ loss, while its masked version(x^+) is used to calculate $\mathcal{L}_{cont}$. 3b: Contrastive loss maximize the agreement of original text(x) and positive sample(x^+) while minimize agreement of negative samples(x^-).

- **Strategy 2. Mask non-offensive words.** Similar to Strategy 1, we predefine the non-offensive words and mask them with the probability. In detail, the probability of masking non-offensive words is set to r while other words are not masked.

- **Strategy 3. Mask all words.** In this strategy, all words in a given sentence have the masking probability r.

In contrastive learning, the model learns input features that are invariant to the transformation. For instance, the rotated, flipped, or cropped images have the same semantic features. Maximizing the agreement between positive pairs leads the model to learn invariant features after rotation, flipping, and cropping. In the same way, masking offensive words leads the model to learn offensive word variants. For instance, users may use "c0ward" instead of "coward" to avoid keyword-based filtering. In that case, our training process can make the model robust to such modification.

In contrast, masking the non-offensive words means the model learns the offensive context. That is, masking non-offensive words makes the model focus on the offensive words (n-words, swear words, or offensive phrases).

Identifying offensive or non-offensive words using c-TF-IDF c-TF-IDF is similar to TF-IDF, but it is different in that it modifies how to calculate TF-IDF scores for multiple classes by joining all documents per class (Grootendorst, 2022). In this case, the words with high scores imply the representative words for each class.

To set the offensive keywords and non-offensive keywords, we used c-TF-IDF with offensive and non-offensive documents in the Jigsaw dataset. We selected the top 500 offensive and non-offensive words based on the c-TF-IDF score from the training dataset. During the random masking process, offensive and non-offensive words have a fixed probability of being masked.

Masking Strategy	Acc	F1	Prec	Recall
Offensive words	0.838	0.719	0.683	0.902
Non-offensive	0.842	0.723	0.686	0.904
All words	**0.844**	**0.726**	**0.688**	**0.904**

Table 2: Performance by masking strategy.

5. Experimental Setup

Datasets We use three datasets as follows.

- Jigsaw: The Jigsaw Wikipedia Comment Dataset comprises 159,571 training and 63,978 test samples. There were six labels: toxic, severely toxic, obscene, threat, insult, and identity hate. We use the toxic label for training. For training, we balanced the Jigsaw dataset by undersampling with 30,588 samples (i.e., 15,294 toxic samples and 15,294 nontoxic samples).

- ConvAbuse: ConvAbuse corpus consists 4,185 samples (Cercas Curry et al., 2021). We use the binary abusive label for training. We split the dataset into 3,332 training and 853 testing datasets.

- Teach: Explained in Section 3. We use 10K randomly selected sentences from the Teach dataset.

Baseline We compare our model, SimMask, with the following baselines in the same setting.

- Always offensive (AO), never offensive (NO): These are the most naive baselines. The low performance of these baselines demonstrates the difficulty of the detection task.

- TF-IDF+SVM: This baseline uses the TF-IDF values of a sentence and feeds them into an SVM classifier. We use the top 2000 frequent unigrams to build the vocabulary. The basic preprocessing was applied, including lowercase, removing the white spaces, stopwords, and punctuation. We used the list of stop

Dataset	Jigsaw⟶ Jigsaw				Jigsaw⟶ Teach			
Model	**Acc**	**F1**	**Precision**	**Recall**	**Acc**	**F1**	**Precision**	**Recall**
Always Offensive	0.095	0.087	0.500	0.048	0.254	0.202	0.500	0.127
Never Offensive	**0.905**	0.475	0.500	0.452	**0.746**	0.427	0.500	0.373
TFIDF+SVM	0.823	0.684	**0.831**	0.655	0.497	0.495	0.617	0.603
BERT-FT	0.829	0.708	0.676	0.895	0.664	0.645	0.671	0.726
SimCSE	0.815	0.693	0.667	0.887	0.543	0.540	0.635	0.660
SupCon	0.726	0.573	0.602	0.758	0.564	0.493	0.595	0.620
SimMask	0.865	**0.749**	0.704	**0.911**	0.744	**0.711**	**0.706**	**0.763**

Table 3: Experimental result from Jigsaw and *Teach* dataset.

Dataset	Jigsaw⟶ ConvAbuse				ConvAbuse⟶ ConvAbuse			
Model	**Acc**	**F1**	**Precision**	**Recall**	**Acc**	**F1**	**Precision**	**Recall**
Always Offensive	0.150	0.130	0.500	0.075	0.150	0.130	0.500	0.075
Never Offensive	0.850	0.459	0.500	0.425	0.850	0.459	0.500	0.425
TFIDF+SVM	0.723	0.644	0.771	0.647	0.886	0.711	0.669	0.825
BERT-FT	0.870	0.803	0.765	0.903	0.898	0.833	0.797	**0.899**
SimCSE	**0.879**	**0.812**	**0.774**	0.901	0.913	0.848	0.819	0.890
SupCon	0.695	0.585	0.623	0.729	0.737	0.598	0.683	0.698
SimMask	0.878	0.811	0.773	**0.905**	**0.914**	**0.850**	**0.822**	0.894

Table 4: Experimental result from Jigsaw and *ConvAbuse* dataset.

λ	**Acc**	**F1**	**Precision**	**Recall**
0.15	**0.841**	**0.722**	**0.685**	**0.903**
0.30	0.837	0.718	0.683	0.902
0.45	0.837	0.718	0.683	0.901
0.60	0.838	0.719	0.683	0.902
0.75	0.834	0.715	0.681	0.899
0.90	0.835	0.716	0.681	0.901

Table 5: Performance by varying λ

r	**Acc**	**F1**	**Precision**	**Recall**
0.15	**0.842**	**0.723**	**0.686**	**0.904**
0.30	0.838	0.719	0.684	0.902
0.45	0.838	0.718	0.682	0.900
0.60	0.833	0.714	0.681	0.901
0.75	0.839	0.719	0.683	0.901
0.90	0.835	0.716	0.682	0.901

Table 6: Performance by varying r

Model	**Acc ↓**	**F1 ↓**	**Prec ↓**	**Recall ↓**
BERT-FT	5.67%	8.15%	11.02%	11.9%
SimCSE	3.70%	4.89%	5.70%	5.02%
SimMask	**3.33%**	**4.51%**	**4.92%**	**4.07%**

Table 7: Performance drop from Jigsaw with TextAttack by model.

words in NLTK library. For the SVM implementation, we use the sklearn Python library with the parameter $C = 1.0$.

- BERT fine-tune(FT): This fine-tuned model has a fully connected layer after the SBERT encoder and was trained using only cross-entropy loss. The configuration is the same as our SimMask model but without the contrastive loss head.
- SimCSE: This baseline uses the original unsupervised SimCSE loss (Gao et al., 2021) $\mathcal{L}_{\text{cont}}$, and the newly added supervised loss $\mathcal{L}_{\text{ce}}$. Unsupervised SimCSE loss increases similarities of output embeddings between positive pairs from the same input, which are different due to the dropout.
- SupCon: This baseline uses supervised contrastive loss (Khosla et al., 2020), along with the newly introduced supervised loss $\mathcal{L}_{\text{ce}}$. Supervised contrastive learning uses samples with the same label as positives while samples with different labels as negatives.

5.1. Experimental Results

Masking Strategy Experiment We tested the three masking strategies mentioned in Section 4.2. The masking probability r was set to 0.8, 0.8, and 0.3 for strategies 1, 2, and 3, respectively. We set a higher masking probability ratio to balance the total number of masked tokens. Since masking both performed best, we use this strategy in the main experiments (Table 2). We use BERT-base-uncased as the encoder and the Jigsaw dataset to train and conduct evaluation.

Classification Performance For the main evaluation metric, we selected the F1-score due to the label imbalance of the test dataset. Experimental results show that TF-IDF achieves the lowest performance, whereas the BERT-based models perform well in the Teach dataset. This demonstrates the generalizability of the pre-trained language model. Our method, SimMask, outperforms all baselines in both datasets except for Jigsaw→

ConvAbuse where SimMask shows slightly lower performance compared to SimCSE. For the Table 3 and 4, we performed t-test and found no significant differences between the F1 scores of SimMask and SimCSE(p-val>0.1). However, in Table 3, there was a significant difference between the F1 scores of SimMask and BERT-FT(p-val<0.1). Notably, SimMask is also trained on an external dataset (Jigsaw) since it shows the potential to detect offensive language in practical scenarios where labeling is scarce due to cost. In this experiment, we use SBERT as the base encoder and fixed two hyperparameters as $\lambda = 0.15$ and $r = 0.3$. Table 3 and Table 4 summarize the performance over train and test dataset(train$\longrightarrow$ test).

5.2. Robustness Testing

As shown in 3.4, people try to fool the rule-based filtering system in chatbot conversation. To detect this kind of attack, we use the TextAttack library to generate adversarial examples (Morris et al., 2020). We corrupt the test dataset using TextAttack algorithms and compare the performance drop of each method. In detail, the attack algorithm inserts space, deletes a random character, swaps neighbor characters, or replaces a character with a similar-looking character for random 30% of the word tokens. Table 7 shows the experimental results on jigsaw dataset, where our suggested method shows minimal performance drop compared to other approaches.

5.3. Ablation Study

This section presents ablation results on two parameters, λ and r. The λ is expected to adjust the training effect between contrastive loss and supervised cross-entropy loss, and r manages the information amount to generate positive samples used in contrastive learning by masking. We varied the values of two hyperparameters in $\{0.15, 0.3, 0.45, 0.6, 0.75, 0.9\}$ and measured the effects. We use BERT-base-uncased as the encoder and the Jigsaw dataset in ablation study experiments. All other experimental details are set the same as in the main experiment.

The lower value of λ guides the model to focus on supervised learning. Table 5 shows that the proposed method is less sensitive to λ. Note that $\lambda = 0$ equals to the baseline BERT-finetune. The value of r guides the model to learn context information. However, it would harm the performance if the masking ratio is too high since it would generate inconsistent positive pairs. The higher the mask ratio is, the greater the model performance may decrease. Since the proposed model also uses supervised loss, in extreme cases (i.e., $r = 0.9$) the model does not collapse (Table 6). Note that $r = 0$ equals the SimCSE baseline.

5.4. Qualitative Analysis

Table 8 shows the classification examples that SimMask correctly classified offensiveness, whereas other baselines (BERT fine-tune and SimCSE) fail. In particular, the examples have no specific obscene words but imply sexually explicit content (S1, S3). For example, S2 shows the modification of the original word by omission. The non-offensive cases showed misclassification of baseline models due to biased words. Words such as ‘thieves’, ‘demon’, ‘black’, or ‘death’ are predicted as offensive content by the baseline models. When we replaced these words with similar nouns or removed such words, the baselines successfully classified them as non-offensive.

6. Discussion and Conclusion

This paper examined the offensive language of users generated during open-domain conversations with the SimSimi chatbot service. Furthermore, we proposed a contrastive learning model that identifies offensive language. Our data revealed a pronounced skew on offensive language topics, exhibiting a heightened frequency of sex-related topics (48%) compared to other subjects like politics or sports. In addition, the proposed method with a random masking strategy outperformed existing models in detecting offensive language. We further emphasize the societal imperative to detect sexually explicit material, especially on platforms predominantly used by vulnerable groups such as children.

Insights from two decades of service We have gathered insights from addressing offensive language over SimSimi’s 20 years of service. We must note that what is deemed offensive words is often rooted in cultural taboos, which may vary across cultures (Vingerhoets et al., 2013). As SimSimi now services 81 languages, the challenge of effectively detecting emerging forms of offensive language across multiple languages has been intensifying. To solve this issue, SimSimi continues to run tasks such as labeling offensive language via crowdsourced human annotation on a per-language basis, aiming to enhance the efficiency of offensive language detection using deep learning models. However, more than 100 million sentences of SimSimi data remain unlabeled, there remains a significant gap in human resource availability for this task.

Nonetheless, we believe that our method leveraging unsupervised learning can resolve two significant challenges highlighted above: the emergence of new offensive words and the extensive, yet untapped, database of unlabeled offensive language. Furthermore, we hope our findings will contribute to the research community’s combat against evolving offensive language on platforms.

Type	ID	Query (User)	Answer (Simsimi)
Obscene	S1	How are you	good. But i'd be better with you inside me
	S2	f me	it would be my pleasure too f someone like you, I'll make sure I'll do it hard ;)
	S3	*Takes a shower*	*Gets in the shower with you*
Non-offensive	S4	joker	P5 phantom thieves leader
	S5	Have you watched angry kid	Yes that ginger demon is so funny
	S6	do you like black pink?	yes, do you?
	S7	What's your biggest fear?	death

Table 8: Classification examples form Teach dataset

Ethics Statement

The entire study design (from data collection to experiments) underwent the first authors' IRB review (IRB-21-494) and given low risk, the need for user-informed consent was deemed unnecessary. Datasets were provided by SimSimi Inc. under a confidentiality agreement.

The data acquired from the company went through a thorough filtration procedure to remove any sensitive or personally identifiable details, including names, email addresses, ages, physical addresses, and sensitive preferences, prior to being provided to our research team.In addition, the researchers made sure not to analyze any aggregated data at the individual user level, ensuring that user identities remained concealed. The utterance data were also prepared in the same manner as random utterance pairs. The key findings of this paper and SimSimi usage data can be found on SimSimi's research blog. Readers are encouraged to check and gain insights, potentially fostering better offensive language detection models or novel research questions.

Acknowlegment

This research was supported by the Institute for Basic Science (IBS-R029-C2) and the National Research Foundation of Korea (NRF) grant (RS-2022-00165347). The authors greatly thank SimSimi Inc. for providing the data for research purposes. We also thank the anonymous reviewers and the chairs for their insightful reviews.

Sofiane Abbar, Yelena Mejova, and Ingmar Weber. 2015. You tweet what you eat: Studying food consumption through twitter. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 3197–3206, New York, NY, USA. Association for Computing Machinery.

Ahmed Al-Rawi. 2019. Gatekeeping fake news discourses on mainstream media versus social media. *Social Science Computer Review*, 37(6):687–704.

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2021. A deep dive into multilingual hate speech classification. In *proc. of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, pages 423–439.

David Alvarez-Melis and Martin Saveski. 2021. Topic modeling in twitter: Aggregating tweets by conversations. *proc. of the International AAAI Conference on Web and Social Media (ICWSM)*, 10(1):519–522.

Antonella De Angeli and Sheryl Brahnam. 2008. I hate you! disinhibition with virtual partners. *Interacting with Computers*, 20(3):302–310. Special Issue: On the Abuse and Misuse of Social Agents.

Petter Bae Bae Brandtzæg, Marita Skjuve, Kim Kristoffer Kristoffer Dysthe, and Asbjørn Følstad. 2021. When the social becomes nonhuman: Young people's perception of social support in chatbots. In *proc. of the ACM CHI Conference on Human Factors in Computing Systems*.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *proc. of the Annual International Conference on Machine Learning (ICML)*, pages 1597–1607.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758.

Hyojin Chin, Lebogang Wame Molefi, and Mun Yong Yi. 2020. *Empathy Is All You Need: How a Conversational Agent Should Respond to Verbal Abuse*, page 1–13. Association for Computing Machinery, New York, NY, USA.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Wen-tau Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *proc. of the North Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*.

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

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6894–6910.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. Declutr: Deep contrastive learning for unsupervised textual representations. In *proc. of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 879–895.

Felipe González-Pizarro and Savvas Zannettou. 2023. Understanding and detecting hateful content using contrastive learning. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 257–268.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. 2020. Bootstrap your own latent - a new approach to self-supervised learning. In *proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 21271–21284.

Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N Asokan. 2018. All you need is" love" evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security*, pages 2–12.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738.

Jennifer Hill, W. Randolph Ford, and Ingrid G. Farreras. 2015. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior*, 49:245–250.

Liangjie Hong and Brian D. Davison. 2010. Empirical study of topic modeling in twitter. In *proc. of the Workshop on Social Media Analytics*, page 80–88.

Liang-Chi Hsieh, Ching-Wei Lee, Tzu-Hsuan Chiu, and Winston Hsu. 2012. Live semantic sport highlight detection based on analyzing tweets of twitter. In *proc. of the IEEE International Conference on Multimedia and Expo*, pages 949–954.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *proc. of NAACL-HLT*, pages 4171–4186.

Chandra Khatri, Behnam Hedayatnia, Rahul Goel, Anushree Venkatesh, Raefer Gabriel, and Arindam Mandal. 2018. Detecting offensive content in open-domain conversations using two stage semi-supervision. *arXiv preprint arXiv:1811.12900*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Youngwook Kim, Shinwoo Park, and Yo-Sub Han. 2022. Generalizable implicit hate speech detection using contrastive learning. In *Proceedings of the 29th International Conference* *on Computational Linguistics*, pages 6667–6679, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *proc. of the Workshop on Trolling, Aggression and Cyberbullying*, pages 1–11.

Jin Myung Kwak, Minseon Kim, and Sung Ju Hwang. 2022. Language detoxification with attribute-discriminative latent space. *arXiv preprint arXiv:2210.10329*.

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.

Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2023. " hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *arXiv preprint arXiv:2304.10619*.

G. Lima, C. Kim, S. Ryu, C. Jeon, and M. Cha. 2020. Collecting the public perception of ai and robot rights. In *proc. of the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, pages 1–24.

Ping Liu, Wen Li, and Liang Zou. 2019. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *proc. of the International Workshop on Semantic Evaluation*, pages 87–91, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Junyu Lu, Hongfei Lin, Xiaokun Zhang, Zhaoqing Li, Tongyue Zhang, Linlin Zong, Fenglong Ma, and Bo Xu. 2023. Hate speech detection via dual contrastive learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Gale M. Lucas, Albert Rizzo, Jonathan Gratch, Stefan Scherer, Giota Stratou, Jill Boberg, and Louis-Philippe Morency. 2017. Reporting mental health symptoms: Breaking down barriers to care with virtual human interviewers. *Frontiers in Robotics and AI*, 12.

Yunshan Ma, Xun Yang, Lizi Liao, Yixin Cao, and Tat-Seng Chua. 2019. Who, where, and what to wear? extracting fashion knowledge from social media. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 257–265.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PLOS ONE*, 14(8):1–16.

Rijul Magu and Jiebo Luo. 2018. Determining code words in euphemistic hate speech using word embedding networks. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pages 93–100.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting hate speech in social media. In *proc. of the International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 467–472, Varna, Bulgaria. INCOMA Ltd.

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 *proc. of the Forum for Information Retrieval Evaluation*, page 14–17, New York, NY, USA. Association for Computing Machinery.

Huina Mao, Scott Counts, and Johan Bollen. 2011. Predicting financial markets: Comparing survey, news, twitter and search engine data. *arXiv preprint arXiv:1112.1051*.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.

Thanh Thi Nguyen, Campbell Wilson, and Janis Dalins. 2023. Fine-tuning llama 2 large language models for detecting online sexual predatory chats and abusive texts. *arXiv preprint arXiv:2308.14683*.

Thien Hai Nguyen and Kiyoaki Shirai. 2015. Topic modeling based sentiment analysis on social media for stock market prediction. In *proc. of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 1354–1364.

Michael J. Paul and Mark Dredze. 2014. Discovering health topics in social media using topic models. *PLOS ONE*, 9(8):1–11.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. 2018. Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604*.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2021. SOLID: A large-scale semi-supervised dataset for offensive language identification. In *proc. of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 915–928.

Debaditya Shome and T. Kar. 2021. Conoffense: Multi-modal multitask contrastive learning for offensive content identification. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 4524–4529.

Vivian Ta, Caroline Griffith, Carolynn Boatfield, Xinyu Wang, Maria Civitello, Haley Bader, Esther DeCero, and Alexia Loggarakis. 2020. User experiences of social support from companion chatbots in everyday contexts: Thematic analysis. *J Med Internet Res*, 22(3):e16235.

George Veletsianos, Cassandra Scharber, and Aaron Doering. 2008. When sex, drugs, and violence enter the classroom: Conversations between adolescents and a female pedagogical agent. *Interacting with Computers*, 20(3):292–301.

Leticia Vidal, Gastón Ares, Leandro Machín, and Sara R. Jaeger. 2015. Using twitter data for food-related consumer research: A case study on “what people say when tweeting about different eating situations”. *Food Quality and Preference*, 45:58–69.

Ad Vingerhoets, Lauren M. Bylsma, and Cornelis de Vlam. 2013. Swearing: A biopsychosocial perspective. *Psychological topics*, 22:287–304.

Yau-Shian Wang and Yingshan Chang. 2022. Toxicity detection with generative prompt-based inference. *arXiv preprint arXiv:2205.12390*.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Gregor Wiedemann, Seid Muhie Yimam, and Chris Biemann. 2020. UHH-LT at SemEval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection. In *proc. of the International Workshop on Semantic Evaluation*, pages 1638–1644.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. In *14th Conference on Natural Language Processing KONVENS 2018*, page 1.

Steve Y. Yang, Sheung Yin Kevin Mo, and Xiaodi Zhu. 2014. An empirical study of the financial community network on twitter. In *proc. of the IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, pages 55–62.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *proc. of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 1415–1420.

Li Zhao and Chao Min. 2019. The rise of fashion informatics: A case of data-mining-based social network analysis in fashion. *Clothing and Textiles Research Journal*, 37(2):87–102.