TST-IOC: A TEXT STYLE TRANSFER-BASED APPROACH TO AUTOMATIC INTERVENTION OF ONLINE OFFENSIVE CONTENT ON SOCIAL MEDIA TO IMPROVE ONLINE SAFETY

by

Zhihui Liu

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing & Information System

Charlotte

2023

Approved by:

_____
Dr. Dongsong Zhang

_____
Dr. Lina Zhou

_____
Dr. Minwoo Lee

_____
Dr. Shaoyu Li

ABSTRACT

ZHIHUI LIU. **TST-IOC**: A Text Style Transfer-based Approach to Automatic Intervention of Online Offensive Content on Social Media to Improve Online Safety.
(Under the direction of DR. DONGSONG ZHANG)

Social media platforms such as Facebook, TikTok and Instagram have witnessed increasing use of offensive language by online users, which can be harmful to other users. Recently the continuance of the pandemic has propelled the propagation of offensive content associated with Covid-19 on social media. Some researchers begin to develop effective methods for detecting online offensive language from social media content automatically, yet automatic intervention of offensive language after it is detected remains largely understudied. To address the gaps, this dissertation develops an effective text style transfer-based approach, TST-IOC, for automatic offensive intervention tasks. The promising outcome suggests that our proposed method shows significant potential and could be a preferred choice among users for offensive intervention tasks. This dissertation provides some contributions. First, it contributes significantly to the field of offensive language research by introducing a new pipeline for generating parallel offensive/non-offensive datasets and a novel text style transfer-based approach, which has been rarely explored in existing intervention studies. This approach shows a step forward in the development of an automatic offensive intervention system, addressing the limitations of current filtering systems deployed by social media platforms. Second, existing research has mainly focused on using performance metrics for evaluating offensive intervention methods quantitatively. However, this study goes beyond by proposing a comprehensive automatic evaluation paradigm. By exploring both quantitative and qualitative aspects of

automatic intervention assessment, it fills a crucial gap in the current offensive language research landscape. Finally, it recognizes the scarcity of studies comparing human evaluation with automatic evaluation in automatic intervention systems. To bridge this gap, we conduct a user study, which allows for an investigation of user acceptance of the proposed automatic intervention approach in real-world scenarios. The insights gained from this user study not only guide the design of more comprehensive automatic intervention systems but also hold the potential to shape the development of human-centric automatic intervention systems in the future.

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to my advisor, Dr. Dongsong Zhang, for providing me with the invaluable opportunity to pursue my Ph.D. studies and engage in exciting research projects in the field of Computer Science and Information Systems. His generous support has been instrumental in enabling me to complete this dissertation successfully. I am immensely grateful for the wealth of knowledge and critical thinking skills that I have acquired under his guidance. Dr. Zhang's passion for research has truly been inspiring me, and I have learned to be a responsible and dedicated researcher through his mentorship. Above all, I am deeply appreciative of his exceptional patience throughout my research and projects, which has taught me the importance of being rational and conscientious in my academic pursuits.

I extend my sincere thanks to my dissertation committee members: Dr. Lina Zhou, Dr. Minwoo Lee, and Dr. Shaoyu Li. Their kindness and dedication in serving on my dissertation committee have been truly appreciated. The valuable feedback and guidance they provided throughout this journey have been enlightening and motivating, significantly enriching my research insights and overall experience.

To my beloved family, I am forever grateful for their unwavering emotional support. My parents' constant encouragement never fails to remind me never to give up. I can never make this day happen without them.

I would also like to extend my gratitude to my lab mates. Abdulrahman Aldkheel, Hamad Alsaleh, Jaewan Lim, Kanlun Wang, Marran Aldossari, and Zhe Fu. Their constant support and understanding have been a source of inspiration and motivation during my academic journey.

To everyone who has played a part in my academic and personal journey, I am indebted

to you all for your support, encouragement, and friendship. Your contributions have shaped me into the researcher and person I am today, and I will always cherish these meaningful connections and experiences.

DEDICATION

With the deepest love and gratitude, I dedicate this dissertation to my mother and father, my unwavering supporters, pillars of strength, and constant motivators throughout my academic journey. Your unconditional love and encouragement have been the driving force behind my pursuit of knowledge and academic excellence. Even in moments of self-doubt, your belief in me gave me the courage to overcome challenges and push my limits.

Thank you for always being there to listen, offer guidance, and share in both my triumphs and setbacks. Your unwavering support and faith in my abilities have played an integral role in reaching this significant milestone.

I am forever grateful for the love, values, and invaluable lessons you have bestowed upon me. As I embark on this next chapter of my academic journey, I carry with me the profound impact of your love and unwavering support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

TST	Text Style Transfer
IOC	Intervention of Offensive Content
NLP	Natural Language Processing
SVM	Support Vector Machines
RF	Random Forest
LR	Logistic Regression
NB	Naïve Bayes Classifier
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
GANs	Generative Adversarial Networks
TL	Transfer Learning
CTF	Conventional Knowledge Transfer Learning
MIMCT	Multi-Channel Transfer Learning based model
CNN-LSTM	Convolutional Neural Network- Long Short-Term Memory
BN	Batch Normalization Layer
FC	Fully Connected Layer
NST	Neural Style Transfer
LLM	Large Language Model
BERT	Bidirectional Encoder Representations from Transformers
PBMT	Phrase-Based Machine Translation

BiLSTM	Bidirectional Long Short-Term Memory
BiGRU	Bidirectional Gated Recurrent Unit
NER	Named Entity Recognition
OLID	Offensive Language Identification Dataset
BLEU	Bilingual Evaluation Understudy
TA	Transformed Accuracy
O_SCORE	Offensive Score
CS	Cosine Similarity
GA	Grammatical Accuracy
PPL	Perplexity
FR	Fully Removal
PR	Partial Removal
CC	Content Control
AMT	Amazon Mechanical Turk
HIT	Human intelligence task

CHAPTER 1: INTRODUCTION

1.1 Research Problem

Social media has played a central role in millions of people's daily life. Via social media, people can not only access news and share information but also share their experiences and discuss interesting topics without a face-to-face meeting. However, the wide adoption of social media also brings several negative effects to our society. One of the growing problems is offensive behavior on social media, which is becoming a pressing issue. Some users' behaviors can not only rub others the wrong way but also offend people. In this study, we define offensive behavior as the egregious things people do on social media that can hurt others' feelings, such as hateful speech. Given their anonymity feature, social media platforms have been abused by some people for harassing other individuals and become increasingly known for offensive behavior such as personal attacks with offensive language [1, 2]. Offensive language is typically described as remarks that are hurtful, derogatory, or obscene in nature and are directed from one individual towards another [3].

Based on recent data, global social media usage has soared to over 4.2 billion users, with individuals spending an average of 145 minutes daily engaging on these platforms [4]. A comprehensive survey encompassing 2,500 adults and teenagers [5] sheds light on the prevalence of social media among young individuals aged 13-17, revealing that a staggering 95% possess a social media account. The dominant platforms in this demographic include YouTube (79%), followed by Instagram (73%), Snapchat (66%), and Facebook (45%).

However, the benefits that teenagers glean from social media are counterbalanced by the potential exposure to substantial quantities of offensive online content. A study conducted by the Chartered Institute of Marketing in 2019 indicates that a notable 46% of teenagers aged 13-17, who are active on social media, have encountered posts that they deem unsuitable for online spaces. Despite the widespread occurrence of children stumbling upon potentially harmful content on social media, the measures taken to safeguard them remain inadequately implemented.

It is crucial to underscore that adolescents are particularly susceptible to the adverse impacts of prejudiced and harmful content, which can manifest in forms like cyberbullying, driven by the propagation of offensive messages across social media platforms. A staggering 60% of students who are exposed to offensive online content reveal that it substantially disrupts their academic pursuits and personal lives [6]. The ramifications of such online toxicity extend to more severe consequences, including a heightened risk of mental health disorders and even instances of suicide. Alarming evidence indicates that students subjected to offensive behaviors, such as cyberbullying, are nearly twice as likely to contemplate or attempt suicide compared to their counterparts who do not experience such distressing encounters [6].

In 2020, the pandemic triggered extensive lockdowns, confining billions of individuals to their homes. As a direct consequence, people increasingly turned to prominent social media platforms for connectivity and information amidst the crisis. This surge in social media usage amplified the potential of these platforms to wield significant influence over people's attitudes and behaviors, particularly in relation to issues of racism [6, 7].

During the initial phases of the pandemic, social media unfortunately emerged as a conduit for discriminatory practices targeting Asian Americans [9]. Certain media outlets propagated prejudiced narratives through headlines such as "Chinese virus pandemonium" or "China kids stay home" [9]. Notably, by early April 2020, Instagram alone featured approximately 72,000 posts with the hashtag #WuhanVirus and 10,000 posts with the hashtag #KungFlu [10]. Regrettably, these types of posts have engendered adverse effects on those engaging with online content. Within this context, social media has come to be recognized as a prominent catalyst for fostering discrimination [11]. The widespread dissemination of such offensive material holds the potential to significantly shape individuals' beliefs and attitudes, ultimately culminating in unpredictable and far-reaching consequences.

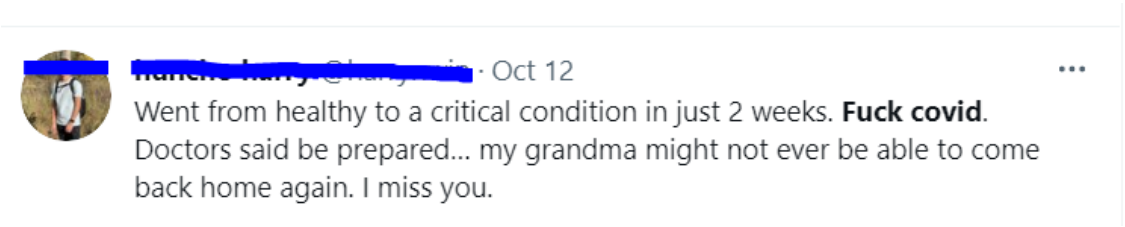


Figure 1. Offensive content about Covid-19 on twitter

Figure 2. A tweet with hashtag #Wuhan Virus

As the problem of offensive content on social media grows, there has been increasing research on identifying potential offensive languages from social media. This

line of research has a potentially significant impact not only on online users, communities, and media platforms, but also on the society as a whole.

In general, there are two major tasks in mitigating offensive language on social media. The first one is automatic detection of offensive content from social media posts, and the second one is automatic intervention. In offensive language detection, many researchers have applied text classification approaches [1, 11–15] in the past decades. Based on the review of existing research, generally, most of the studies utilize machine learning techniques, starting with collecting and annotating social media messages, then training machine learning methods to classify an online message as either offensive or not. Despite that some studies have made efforts on automatic detection and achieved good results [12–15], there have been few studies exploring the solutions to the automatic detection and intervention of offensive content on social media, which leads to a new task in addressing the online offensive language problem.

The current strategy of mitigating offensive language used by social media platforms like Twitter and Facebook is enforcing community standards, which lays out the type of contents not allowed on the platforms. Usually, social media platforms follow two general approaches to deal with offensive content. The first approach is to deploy an automatic filtering system, which can filter words appearing in a post with offensive keywords in a blacklist. However, filtering results may not be enough to address the offensive language problem because directly filtering offensive words from posts breaks the readability of the original message. Further, users can often easily guess what the offender wants to express or even infer the filtered offensive words [17], which makes this approach unable to protect users from offensive content because an offensive intention has

been delivered to victims successfully. The second approach is to review offensive content by human administrators of social media platforms. The administrator manually reviews and deletes the posted contents with any offensive language [1]. This manual filtering system can achieve the best performance, but it is very time and labor consuming. This limitation determines that it cannot be widely applied in practice.

An additional drawback of filtering methods is that while users engage with online content, they might not necessarily desire the complete removal of a message. Instead, they might prefer the message to be reworded in a non-offensive manner, allowing other users to view the post without taking offense [18]. Conversely, for individuals who unknowingly share offensive content, a platform that not only notifies them of the offensiveness of their content and the possibility of it being blocked, but also provides an alternative version with offensive language substituted, could motivate users to reduce the use of offensive language in their social media posts [18]. Thus, if an automatic process can transform offensive content into a non-offensive version while keeping the original content meaning intact, it may effectively mitigate the propagation of offensive language and protect users from exposure to potential hurt or risk in the virtual space.

1.2 Research Questions

This study is aimed at filling the gaps and limitations of prior studies. The first gap is the relative lack of automatic intervention studies on the problem of offensive language on social media. Although several prior studies exist with a focus on the problem of offensive language detection [12–15], there are few related studies that dive into automatic intervention strategies and answer the question: after offensive content is detected, what

can be done with the detected offensive language to minimize its negative impact? This is an under-explored research avenue yet a crucial task in offensive language research area. Further, there are no prior studies conducting human evaluation to assess the performance of offensive content intervention systems [18]. There is a lack of existing user studies that explore the difference in evaluation performance between human and automatic intervention systems when assessing the generated results [18]. This draws forth the three key research questions in this study:

RQ1: How to develop an effective approach to intervention by rephrasing the detected offensive content to mitigate its impact on users?

RQ2: How to assess the performance of an automatic offensive content intervention method comprehensively and effectively?

RQ3: Is there a significant difference between human judgment and objective evaluation when assessing the performance of an automatic intervention method?

In this study, we address the above research questions by investigating the development of an end-to-end deep learning method and the assessment of the proposed method via a systematic evaluation paradigm and a user study. In the first development stage, we proposed, implemented, and evaluated TST-IOC, a text style transfer-based approach that is aimed at minimizing the changes to the content of original offensive sentences while removing offensiveness as much as possible. TST-IOC treats automatic intervention as a sequence-to-sequence Natural Language Processing (NLP) task. In the second assessment stage, we first propose a systematic evaluation paradigm. The evaluation and performance measures of automatic intervention are severely under

explored. This systematic evaluation paradigm extends the current automatic intervention research for assessing the performance of our proposed method from various perspectives. Second, we conduct a user study to compare the discrepancy between the evaluation performance of our proposed method and human judgment. The importance of this comparison is that modern deep learning models highly depend on data collection and annotation, thus, this assessment can also assist future researchers to build fair and equitable deep learning models and reduce the impact of human biases in automatic offensive intervention tasks.

This study provides some research and practical contributions to the field of mitigation of offensive language on social media. Firstly, the study extends current offensive language research by implementing a novel text style transfer-based approach, which has been rarely investigated in existing offensive intervention studies. This solution takes one step further towards the solution of an automated social media intervention system to substitute the ineffective and time-consuming filtering systems currently deployed by social media platforms and improve the safety of both online users and communities.

Secondly, although researchers have conducted several automatic metrics to evaluate offensive intervention methods from quantitative side, there are no prior studies to investigate a systematic evaluation paradigm and human evaluation to assess the performance of automatic intervention systems [18]. This study extends the current offensive language research by proposing a framework of an automatic evaluation paradigm and exploring the assessment of automatic intervention from both quantitative and qualitative aspects.

Thirdly, as there are few studies comparing automatic evaluation and human evaluation in automatic intervention systems, we also conduct a user study to investigate the difference between human judgment and automatic evaluation methods This allows researchers to investigate the user acceptance of the proposed automatic intervention approach in practice. Furthermore, this user study not only provides guidance for designing a comprehensive automatic intervention system but also might have profound implications for designing a human-centric automatic intervention system in the future.

Social media has provided a free space to users to share their opinions and thoughts with others. Crowds using offensive language pertaining to hate speech, harassment, and cyberbullying make it difficult to maintain the intricate balance between freedom of expression and the defense of human dignity. Offensive online behavior isn't just an issue of civility and healthy discourse. It poses a threat to individual online users, social media platforms, and society. It is imperative to prevent the propagation of such offensive content on social media. The primary goal of this study is to mitigate the spreading of offensive language on social media by leveraging advanced deep learning techniques. This research not only protects online users to avoid potential harm from offensive content but also benefits online communities by providing a substitute to them instead of current inefficient offensive content filtering strategies.

1.3 Research Outline

This study is organized as follows. In Chapter 2, we present a review of the basic concepts of offensive language. Then we present some existing studies and datasets for offensive language research. We also present the importance of intervention in offensive

language. Chapter 3 introduces our automatic intervention method to address problems of offensive language and describes the collection process of the datasets.

In Chapter 4, we introduce two parts of the evaluation methods for automatic intervention of offensive language. The first part is objective evaluation, we describe the baselines, evaluation method, and the framework of a systematic evaluation paradigm for an automatic intervention system and performance metrics that are used to assess the performance of our proposed method. The second part is subjective evaluation, we describe a human evaluation-based user study that is adopted in this study to assess the performance of our proposed approach. In Chapter 5, we report the results of this study from the experiments. We conclude with the major findings and research contributions. Then we also discuss the limitations and future research in Chapter 6.

CHAPTER 2: LITERATURE REVIEW

This chapter first introduces the concepts of offensive language, then reviews existing research on detection of offensive language, including traditional machine learning methods based on feature engineering and deep learning methods. The rest of the section reviews existing automatic filtering methods that have been deployed by social media platforms and highlights the urgency of research on effective automatic intervention of offensive language in existing literature.

2.1 Concepts of Offensive Language

A universal definition for "offensive language" does not exist. According to Jay et al. [19], offensive language encompasses vulgar, pornographic, and hateful expressions. Vulgar language refers to speech or writing that is crude, obscene, or offensive in nature. It often includes profanity, explicit sexual content, or derogatory terms that are considered socially unacceptable and disrespectful. Pornographic language refers to explicit or graphic speech or writing that is sexually explicit and intended to arouse sexual desire. It often includes explicit descriptions of sexual acts, body parts, or content that is explicit in nature. Pornographic language is typically considered explicit and not suitable for general audiences, as it can be offensive or inappropriate in many contexts.

Offensive language constitutes a broad category encompassing profanity and insults of various kinds. Fortuna et al. [28] have compiled a comprehensive overview of these concepts' definitions in Table 1.

Table 1. Different definitions of offensive language in previous research

Concept	Definition
Hate speech	Displaying aggression or hostility without providing a stated rationale for such behavior [20].
Cyberbullying	A deliberate and aggressive action executed by an individual or a group, utilizing electronic means of communication, consistently and over a period of time, targeting a victim who faces difficulty in effectively protecting themselves [21].
Discrimination	The procedure by which a distinction is recognized and subsequently exploited as the foundation for unjust treatment [22].
Flaming	Flaming refers to the use of hostile, profane, and intimidating remarks that have the potential to disrupt engagement within a community [25].
Abusive language	The phrase "abusive language" was employed to denote hurtful speech, encompassing hate speech, derogatory expressions, and profanity as well [15].
Profanity	obscene word or phrase [23].
Toxic language or comment	Toxic remarks consist of impolite, disrespectful, or irrational messages that have a high likelihood of prompting an individual to exit a discussion [24].
Extremism	An ideology linked to extremists or hate organizations, advocating for violence and frequently seeking to divide populations while asserting dominance, portraying outgroups as either instigators or less privileged groups [26].
Radicalization	Radicalization is the process where individuals or groups adopt extreme ideologies, deviating from mainstream values, which can span various areas like politics, religion, or social beliefs [27].

2.2 Automatic Detection of Offensive Language in Social Media Content

People are widely using social media to show their personal views and opinions and share feelings about their social life with others. By taking advantage of posting content freely, more and more people begin to misuse such online platforms and post offensive content. The government issued the lockdown to keep people safe during the early stage of the pandemic last year. The lockdown not only had a negative impact on the social and psychological well-being of people but also affected the economy. Many people voiced

their opinions online and there are many posts filled with offensive content. Those offensive contents easily instigate violent behavior offline. One of the examples is racism against Asian Americans related to the pandemic, which happened not long ago [29]. Identifying and controlling such offensive content is important to keep innocent users away from such attacks. It is impossible to rapidly detect and remove offensive content manually due to the excessive volume of content posted on social media platforms in real time.

Automatic detection of offensive language can be treated as a text classification task that classifies a social media post as offensive or not. In the existing offensive language detection research, there are three main categories of approaches: traditional machine learning approaches, deep learning approaches, and transfer learning approaches.

2.2.1 Traditional Machine Learning Approaches

Machine learning has been proven to be useful for solving text classification tasks. Many studies have implemented machine learning techniques for offensive language detection [1, 11–15, 17].

Support Vector Machines (SVM) is one of the most widely used methods for classification in offensive language research field [1, 12, 19, 30–32]. For example, Warner et al. [13] collected hate speech data from Yahoo, and the American Jewish Congress, and they used an SVM classifier to detect hate speech. Chen et al. [1] proposed the Lexical Syntactic Feature (LSF) architecture, which is one of the first systems to use a combination of lexical and syntactic features to detect offensive language in YouTube comments in order to protect adolescents. In [30], authors experimented with a multiclass, multilabel classification model on hateful comments corpus collected from YouTube and Facebook. They found that linear SVM achieved the best performance using TF-IDF features.

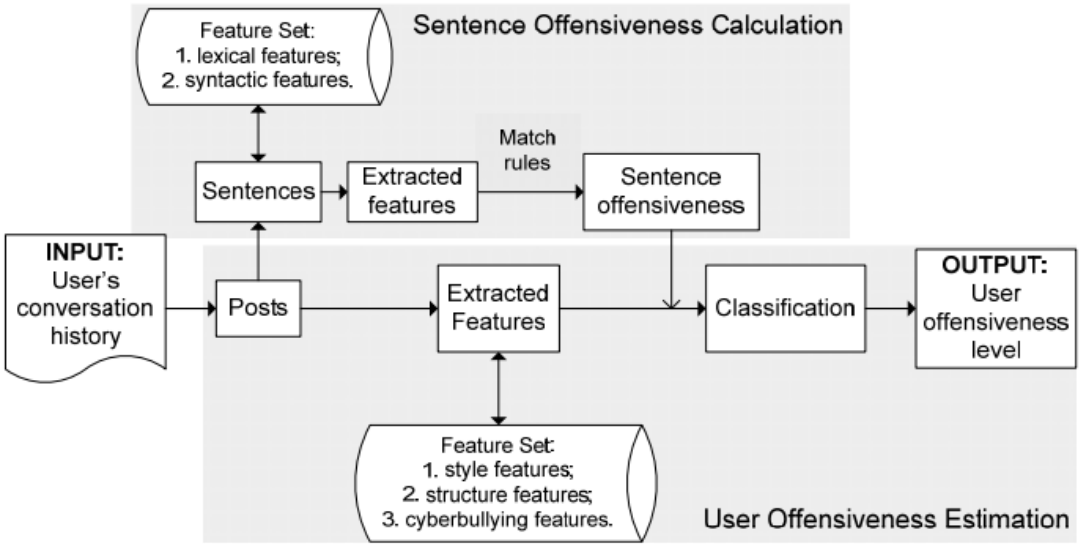


Figure 3. The structure of LSF-based offensive language detection [1]

Random Forest (RF) is an ensemble of several decision trees. It has been employed in many studies for offensive language detection [32, 33]. Agarwal et al. [33] proposed a cascaded ensemble learning classifier for identifying the posts containing racist or radicalized intent on Tumblr microblogging website. The results revealed that Random Forest outperformed Naïve Bayes and Decision Tree algorithms. Burnap et al. [34] used a random forest decision tree to create a rule-based approach for distinguishing hateful speech.

Logistic Regression (LR) is another machine learning algorithm that has been used in many offensive language detection studies [11, 14, 15, 30, 31, 35]. For instance, Davidson et al. [31] used a logistic regression with L2 regularization for the classification model. Xiang et al. [12] proposed a novel semi-supervised approach for detecting profanity-related offensive content on Twitter. Nonata et al. [15] experimented with different syntactic features and types of embedding features to detect the hate speech in

online user comments with a regression model. They found the model to be more powerful when combining these new features with the standard NLP features. Djuric et al. [16] trained a logistic regression classifier with a paragraph2vec embeddings to classify language in user comments as abusive or clean. The results showed paragraph2vec performed better than BOW model and required less memory and training time.

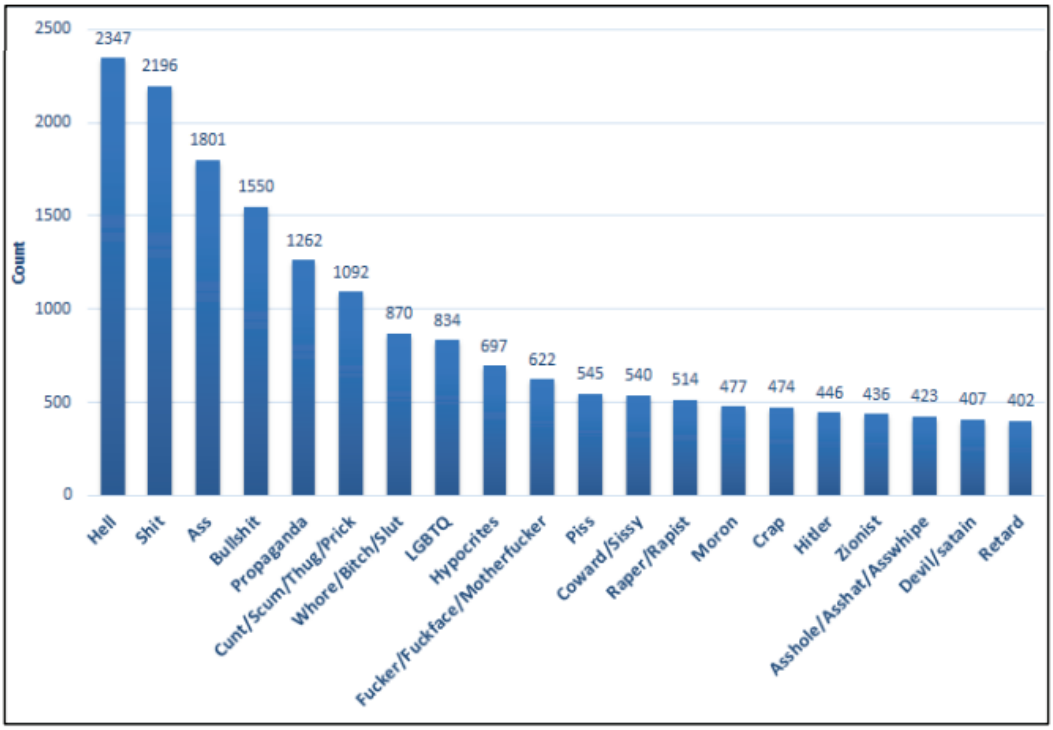


Figure 4. Distribution of Offensive Language appeared in a Facebook and YouTube comments dataset [30]

Naïve Bayes Classifier (NB) is one of the most efficient and effective machine learning algorithms. Some studies used the Naïve Bayes classifier in offensive language detection tasks [13, 33, 36]. Kwok et al. [14] applied a Naïve Bayes classifier, employing inexpensively acquired labeled data from diverse Twitter accounts to learn a binary classifier for detection of ‘racist’ and ‘nonracist’ tweets. In [36], authors incorporated an

n-gram representation and sentiment features into Naïve Bayes classification model. The results showed that 5-gram based models achieved significant performance in precision for the racism related web content.

2.2.2 Deep Learning Approaches

Deep learning is one of the most popular and widely explored techniques in recent years. Deep learning methods have achieved extremely promising results in many research fields, such as text mining and natural language processing [37–39]. Compared with traditional machine learning approaches that are based on manually crafted features, which not only is time-consuming but also may cause biased feature selection, deep learning approaches can learn representations of social media content. The problem shifted from modeling relevant input features to modeling the network. There have been many studies on offensive language detection with deep learning methods [40–46].

Convolutional Neural Network (CNN) is a class of deep learning models that have achieved well performance for processing and analyzing visual data [47]. CNN has also been applied to offensive language detection in a lot of recent studies [41, 43, 44]. To detect sexist and racist language, Park et al. [43] proposed a HybridCNN model built with a dataset of 20K tweets. Singh et al. [44] experimented with the CNN model to detect aggressive posts.

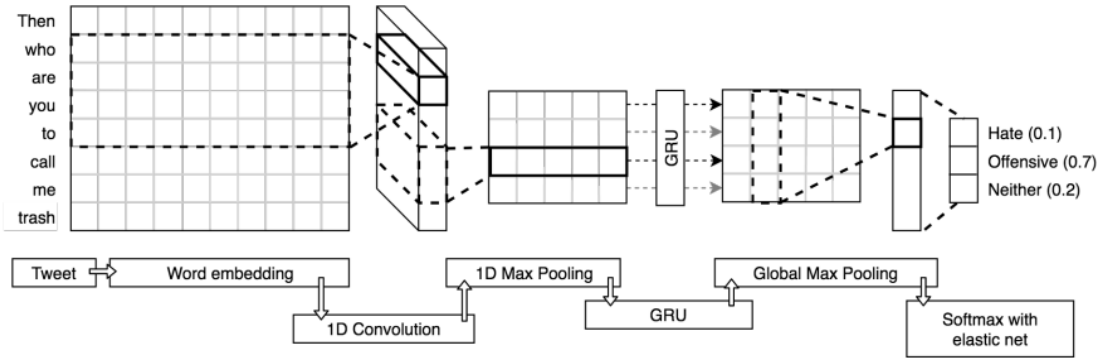


Figure 5. The architecture of CNN-GRU model [45]

Recurrent Neural Network (RNN) is a type of neural network architecture designed to handle sequential data and time-series data, such as text or speech [48]. In particular, Long Short-Term Memory (LSTM) is a recurrent neural network architecture that can learn order dependence in sequence prediction problems [40, 42, 46]. [40] proposed an LSTM model to discover discrimination from tweets. Mishra et al. [42] developed an RNN model for classifying tweets as racism, sexism, or none by incorporating author profiling features to improve the model performance. Hu et al. [46] proposed an LSTM neural network by using a pronunciation-based representation of hate speech and offensive language from Twitter, and they found the pronunciation-based presentation could significantly reduce noise and enhance the performance.

(a) *None* class (b) *Sexism* class

Figure 6. Visualization of authors from different classes [42]

Although various studies have shown that both traditional machine learning and deep learning approaches can achieve good performance in a number of applications, one limitation of those approaches is that the effectiveness of models hinges on the congruence between training and testing data, where both are sourced from an identical feature space and exhibit an identical distribution. [49]. It is essential to avoid retraining a model from scratch, making transfer learning between different domains with potentially different label spaces or conditional distributions highly desirable.

Table 2. Categories of existing approaches to offensive language detection

Category	Features used	Algorithms	Year
Feature engineering approaches	TF-IDF, POS, sentiment, hashtags, mentions, retweets, URLs, number of characters, words, syllables	Logistic Regression, SVM	2017 [31]
	POS, sentiment analysis, word2vec, CBOW, N-grams, text features	SVM, LSTM	2017 [32]
	Lexicon-based features, abusing language dictionary of words and phrases	SVM	2018 [3]
	N-gram, semantic and syntactic features	Logistic Regression, Decision Tree, Linear SVM	2018 [30]
	N-gram, TF-IDF, user features	Logistic Regression	2018 [35]
	Topic modelling, sentiment analysis, tone analysis, semantic analysis, contextual metadata	One- class Classifiers, Random Forest, Naïve Bayes, Decision Trees	2016 [33]
	Rule-based approach, sentiment analysis, typed dependencies	Rule-based classifier	2015 [50]
	TF-IDF, N-grams, topic similarity, sentiment analysis	Naïve Bayes	2014 [36]
Deep learning approaches	Word2vec	LSTM	2016 [40]
	GloVe embedding	CNN, LSTM, FastText (like BoWV model)	2017 [41]
	Author profile	LSTM	2018 [42]
	Word2vec, one-hot encoding	CNN	2017 [43]
	Count of offensive words, number of tokens, size of post, presence of URLs, presence of hashtags, presence of phone numbers.	CNN, LSTM	2018 [44]
	Word embedding	CNN-GRU model	2018 [45]
	GloVe embedding	Bi-directional LSTM	2020 [46]

2.2.3 Transfer learning approaches

A person can learn to drive a truck quickly after learning to drive a passenger car first and learning mathematics makes it easier for students to study physics. Transfer of learning can be defined as the process when learning in one context or with a set of resources influences the performance of the same task in another context or with other related resources [51]. It can extend the definition of transfer of learning [52] to the context of machine learning as the extent to which the learning of offensive language detection contributes to the subsequent learning in a downstream task. There are different perspectives that predict and explain human performance in knowledge transfer [53]:

- Analogical transfer [54] is a three-step process: 1) Retrieve a prior knowledge structure that is relevant to the new situation or problem; 2) Create a mapping between the prior knowledge structure and the new situation or problem; 3) Use the mapping to generate or adapt a new knowledge structure that is relevant to the new context. The transferred knowledge is typically assumed to be a declarative representation, meaning that it is a factual representation of the world.
- Knowledge compilation [55] is the process of translating declarative knowledge into a form that can be used to solve problems. This can be done by translating the knowledge into a set of procedures or actions, or by translating it into a more compact representation that can be used more efficiently. Knowledge compilation has a tradeoff between applicability and efficiency. On the one hand, it can be applied to a wider range of problems than other approaches to knowledge representation, such as logic programming. On the other hand, it can be more complicated to apply, and it may not be as efficient as other approaches for certain problems.

- The error correction theory [56] states that declarative knowledge helps learners identify and correct their own errors. Declarative knowledge is factual knowledge about the world, such as facts, concepts, and rules. It can be used to constrain possible solutions to a problem. When incomplete or faulty procedural knowledge generates undesirable outputs, it will be revised by the rules accordingly.

Typically, there are three major research problems in transfer learning, including what to transfer, how to transfer, and when to transfer [49]. In the context of neural networks, a common transfer learning technique is to fine-tune the last layer of a pre-trained model. This entails adjusting the weights and parameters of only the final layer while keeping the rest of the model frozen [55, 56]. Mathur et al. [57] applied learning models with multiple features using transfer learning, which was pre-trained with English tweets. Wiedemann et al. [58] investigated potential strategies for transfer learning. They used a pre-trained BiLSTM-CNN model with a “One Million Post” corpus, which provided non-offensive annotated labels for over 11,000 user comments to learn the actual offensive language task. Unlike the above two studies that leveraged pre-trained knowledge from a background corpus, in the study by Rizoiu et al. [59], they devised an automated text analytics technique that can simultaneously learn a unified representation of hate using disparate, smaller datasets.

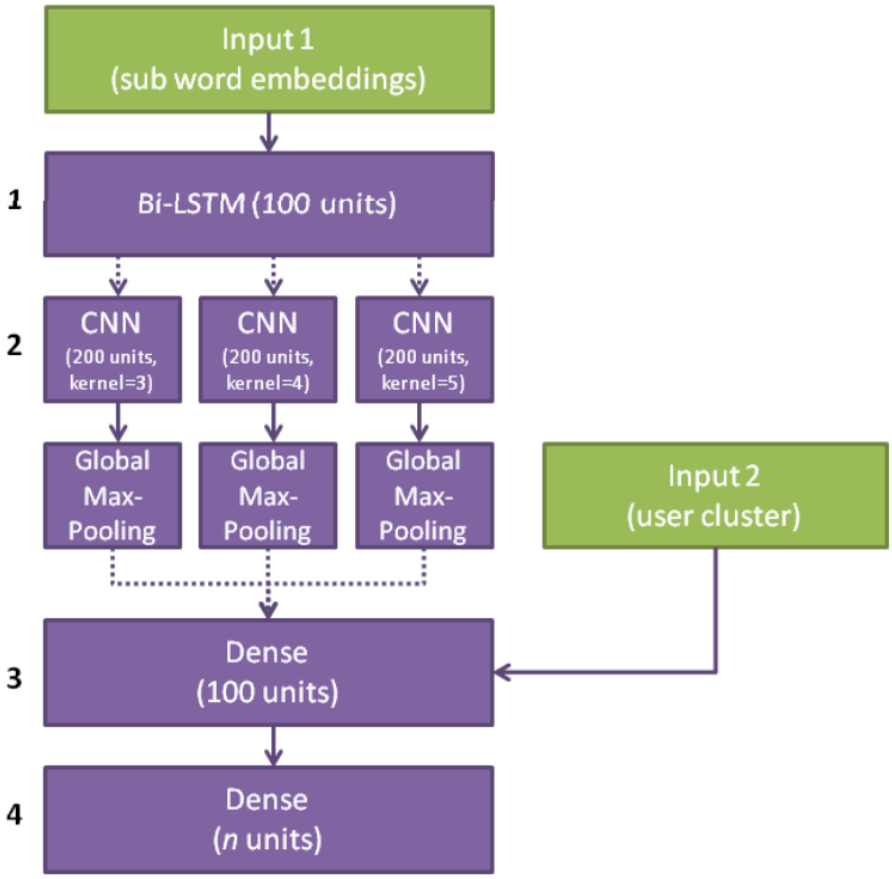


Figure 7. The architecture of BiLSTM-CNN model [58]

2.3 Intervention of Offensive Language

Social media provides a virtual environment to users. It, however, has also unintentionally encouraged the use of offensive language. Although people have been aware of the problem of offensive language on social media platforms and increasing efforts have been made on detecting offensive language [1, 11–14]. But only detecting offensive content is not enough to protect people from the harm of bullying/harassment. More importantly, how the detected offensive content can be intervened after it is detected remains intensely studied.

In the existing studies, there are two main categories of approaches that are widely used by social media platforms to intervene in offensive content [1, 16]: automatic filtering and manual filtering.

2.3.1 Automatic Filtering

The process of removing social media posts with offensive content is called offensive language filtering [17]. A sentence consists of a sequence of words. The process of identifying and removing offensive words is called offensive language filtering. The existing automatic filtering approaches can be categorized into two methods: keyword censoring and content control [17].

The keyword censoring method involves comparing words in a post with offensive terms in a blacklist. Offensive words detected are either removed, partially masked (e.g., "a***"), or entirely replaced (e.g., "***"). This method has been extensively used on platforms like YouTube and World of Warcraft websites [17]. Despite its popularity, this approach might not always be the best solution for tackling offensive language. Simply erasing words could disrupt content clarity. Additionally, users often deduce the removed offensive words, undermining the efficacy of filtering. In such cases, offensive content still reaches its intended recipients, indicating the limitations of the filtering process.

The content control method is commonly employed by users to prevent exposure to inappropriate online content. This method utilizes rules like identifying URL addresses, offensive words, and topic categorization for filtering [17]. For instance, a prevalent rule involves counting offensive words in a sentence; if they surpass a certain limit, the filter eliminates the sentence. However, this approach is susceptible to evasion by offenders familiar with the system's rules. Additionally, automatic filters often block entire posts,

even if only a small portion is offensive. This practice can hinder users' engagement in social media and impede online community expansion.

2.3.2 Manual Filtering

Manual filtering method is another content filter adopted by social media platforms. This method usually requires a real human in the backend to review the published posts.

The administrator review method requires human administrators of social media platforms to review the submitted content manually [1]. Typically, users' posts are reviewed by community administrators before being posted on a social media platform. If the posts contain any inappropriate content, the administrators will remove the posts. This approach does not seem a feasible solution due to the large volume of social media posts and the significant human resources efforts required for manual inspection. It is impractical to be widely applied. In addition, in a real-world case, when users post a message on a social media platform, they expect the message to be displayed as soon as possible and seen by others, but a manual review by administrators often delays the delivery of the posts.

2.3.3 Existing Studies In Automatic Offensive Intervention Problem

Automatic intervention of offensive language is the task of removing offensive language from a sentence without changing the meaning of the sentence. This is a text generation task, and it can be solved using neural models.

Recurrent neural networks (RNNs) [82], Bidirectional LSTM [87], and Bidirectional GRU [88] have been used for text generation, but they require large datasets to train. Pretrained language models, RoBERTa [89], XLNet [90] have made significant progress in downstream tasks [80], and they can be used for automatic intervention of

offensive language. However, these models are usually trained on general corpora and lack domain-specific knowledge. To address this limitation, some domain-specific models have been trained for different domains, such as finance (FinBERT) [91] and scientific texts (SciBERT) [92]. In offensive language research, Sarkar et al. [83] proposed a transformer-based model, fBERT, for offensive language identification. However, offensive language domain-specific models have not been explored for automatic intervention tasks. Some previous studies focused on exploring the development of offensive intervention methods with non-parallel datasets. Krishna et al. [113] conducted a comprehensive investigation of the domain of paraphrase generation. They proposed a versatile paraphrasing tool that could effectively remove stylistic cues from the source text. Dale et al. [114] proposed a fine-tuned T5 paraphraser with style-informed language models for automatic offensive intervention tasks. They achieved 95% in style accuracy; 66% for content preservation and 0.8 for language fluency of the paraphrased ... An encoder-decoder model with a pretrained style classifier has been applied in previous offensive intervention tasks [76, 86]. That model achieved a style accuracy of 86%; 30.2 for BLEU and 0.19 for fluency.

A simple solution to automatic intervention can be substituting an offensive word with another word without distorting/changing the meaning of the original sentence while eliminating the offensiveness. In this study, we define offensiveness as the single offensive word that appears in the source sentence. Thus, the goal is to find a way to substitute a target offensive word. Although the pre-trained transformer models have been used in offensive language classification [82, 83], there are rare applications of pre-trained transformer-based models for automatic intervention. In this dissertation, we propose a

novel text style transfer-based approach that adapts the transformer model for this challenging task.

2.4 Summary

In this section, we reviewed the existing research on detection of offensive language that includes both traditional machine learning methods using feature engineering but also deep learning methods. The performance of those approaches heavily depends on the amount of training data. The lack of publicly available datasets of annotated social media posts involving offensive language is a crucial challenge. It is very expensive to build high-quality labeled datasets. Some studies work on transfer learning to reduce the need and effort to recollect large-scale training data and retrain a model from scratch when detecting offensive language. We also reviewed automatic and manual filtering methods that are used by social media platforms to intervene in offensive language. Then we demonstrate their limitations in mitigating offensive content. Finally, we reviewed the existing studies in offensive intervention problem.

CHAPTER 3: METHODOLOGY

In this chapter, we describe the proposed novel automatic intervention approach that addresses the limitations of current offensive language research. To fill the gap of automatic intervention research, as outlined in the previous section, we first present an encoder-decoder automatic offensive intervention approach when a parallel dataset is not available. Then we introduce our proposed text style transfer-based approach for offensive intervention tasks.

3.1 Intervention of Offensiveness: when parallel datasets are not available

Automatic intervention can be considered as a text style transfer problem. Text style transfer is the process of changing the style of a sentence while maintaining the original content. The goal is to create a sentence that has the same meaning as the original sentence, but that is written in a different style. In automatic intervention, an offensive input sentence can be modified to a non-offensive output sentence by substituting the target offensive word with a non-offensive one. This sentiment modification can be treated as one kind of text style transfer [18]. Utilizing parallel corpora to train a deep learning model requires the datasets, which contain parallel sentence pairs: each offensive sentence is aligned with a matched non-offensive sentence. However, due to the lack of parallel offensive language datasets and the time-consuming process of data collection/annotation, finding an existing parallel dataset is usually impractical. Therefore, to address this problem, exploring a non-parallel text transfer method is a promising solution and studied by many previous studies in various style domains [72, 86]. In this study, we introduce a text style transfer-based automatic offensive intervention approach. The basic idea behind this approach is that we

utilize the non-parallel style transfer approach to generate target results. In the text style transfer task, a style is typically defined as positive or negative; informal or formal and so on. For example, a negative sentence: “my goodness this dish is disgusting.” can be transferred to a positive sentence: “my goodness this dish is delicious.” by substituting the target word in the source sentence. For another example of offensive/non-offensive language style transfer, an offensive sentence: “fuxk is crazy here.” could be transferred to a non-offensive sentence: “it is crazy here.” By substituting the target offensive keyword. Therefore, we hypothesize that a text style transfer approach will achieve the transformation of offensive sentences successfully.

In the automatic offensive intervention task, we define an input sentence as two styles: offensive and non-offensive. We assume $X = \{(T_i, S_i)\}_{i=1}^{N}$ is a set of offensive sentences (source domain). We want to rephrase an offensive sentence T_i with offensive style S_i to a non-offensive sentence $\widehat{T_i}$ with non-offensive style $\widehat{S_i}$. To achieve this purpose, we can encode the input offensive sentence T_i to its content representation c_i by an encoder model. Then we can use a decoder model to reconstruct the non-offensive sentence $\widehat{T_i}$ with the content representation c_i and non-offensive style $\widehat{S_i}$. The process can be divided into the following steps:

1) We define an encoder-decoder model as (E, D). The encoder E encodes the input offensive sentence T_i to its content representation c_i. To estimate the conditional distribution of $\widehat{T_i}$ when giving the content representation and non-offensive style. We can use the decoder D to get it as:

$$p\left(\widehat{T_i} \mid c_i, \widehat{S_i}\right) = \prod_{t=1}^{X} p\left(\hat{T}_i^t \mid \hat{T}_{i<t}^t, c_i, \widehat{S_i}\right) \qquad (1)$$

In equation (1), $\widehat{T}_i^t$ represents the t^{th} word in non-offensive sentence $\widehat{T}_\iota$. $\widehat{T}_{i<t}^t$ represents the prefix of non-offensive sentence $\widehat{T}_\iota$.

2) Since a parallel dataset is not available, we cannot estimate (1) directly. We can formulate an encoding reconstruction loss function as:

$$L_e^X = -\mathbb{E}_{T_i \sim X} \log p(\, T_i |\, c_i,\, S_i) \qquad (2)$$

3) We assume D aims to faithfully recover the original stylistic property of input text, denoted as offensive sentence T_i, when provided with the corresponding offensive style S_i. The non-offensive sentence $\widehat{T}_\iota$ is sampled from the distribution $\widehat{T}_\iota \sim p\big(\, \widehat{T}_\iota |\, c_i, \widehat{S}_\iota\big)$. However, optimizing equation (2) can result in the model failing to rephrase the offensive style as intended. To address this issue, we can employ a style classifier as a style regularization technique [76]. This style classifier C^T, which ensures that the model can effectively output non-offensive sentence $\widehat{T}_\iota$ with its correct style label $\widehat{S}_\iota$, C^T is pretrained on the offensive/non-offensive language dataset.:

$$L_{style}^X = -\mathbb{E}_{\widehat{T}_\iota \sim p(\, \widehat{T}_\iota |\, c_i, \widehat{S}_\iota)} \log p_{C^T}\big(\, \widehat{S}_\iota |\, \widehat{T}_\iota\big) \qquad (3)$$

4) The final loss function for optimization is as follows:

$$L^X = L_e^X + L_{style}^X \qquad (4)$$

3.2 Intervention of Offensiveness: TST-IOC

3.2.1 Problem Formulation

The problem of automatic intervention can be represented as the following scenario: given an offensive sentence $T_i = (t_1, t_2, \dots, t_k, \dots, t_m)$, where $t_1, t_2, \dots, t_k, \dots, t_m$ are

individual words, an offensive word t_k in this sentence needs to be substituted with a non-offensive word t_k'. Then the offensive sentence T_i will be transformed to a new sentence $Y_i = (t_1, t_2, \dots t_{k-1}, t_k', t_{k+1} \dots, t_m)$, which is a non-offensive version of the original sentence. Thus, the task is equivalent to finding an alternative non-offensive sentence *Y* that maximizes the conditional probability of non-offensive sentence *Y* given an offensive source sentence *X*, $arg\ max\ p(Y \mid T)$. In this task, a parameterized model will be trained to maximize the conditional probability of a generated non-offensive sentence given its paired offensive sentence using a parallel training dataset. Once the conditional distribution is learned by the model, given a source offensive sentence, a corresponding transformed non-offensive sentence will be automatically generated by searching for an alternative sentence that maximizes the conditional probability.

Our intuitive design of transforming an offensive language sentence is that if there exists a dataset with paired sentences $\{T_1, Y_1; T_2, Y_2; \dots; T_n, Y_n\}$, we can easily formulate automatic intervention as a sequence-to-sequence NLP task. The sequence-to-sequence task can be solved by implementing an encoder-decoder structure language model. In recent years, transformer-based models have made a dramatic impact on text generation tasks such as machine translation [80] and dialogue systems [94]. However, the effectiveness of a transformer model in the automatic intervention task remains to be explored. In this study, we investigate the effectiveness of the proposed language model for automatic intervention tasks to fill this gap.

3.2.2 Neural Style Transfer

Gatys et al. [60] pioneered the application of convolutional neural networks (CNNs) to transfer the artistic style of renowned paintings onto ordinary photographs. They

demonstrated that a CNN can extract content details from photographs and artistic styles from iconic artworks. Their technique involves iteratively refining an image to align with the CNN's feature distributions, combining content and style information. This seminal work established the groundwork for neural style transfer (NST), a method utilizing CNNs to apply diverse styles to content images.

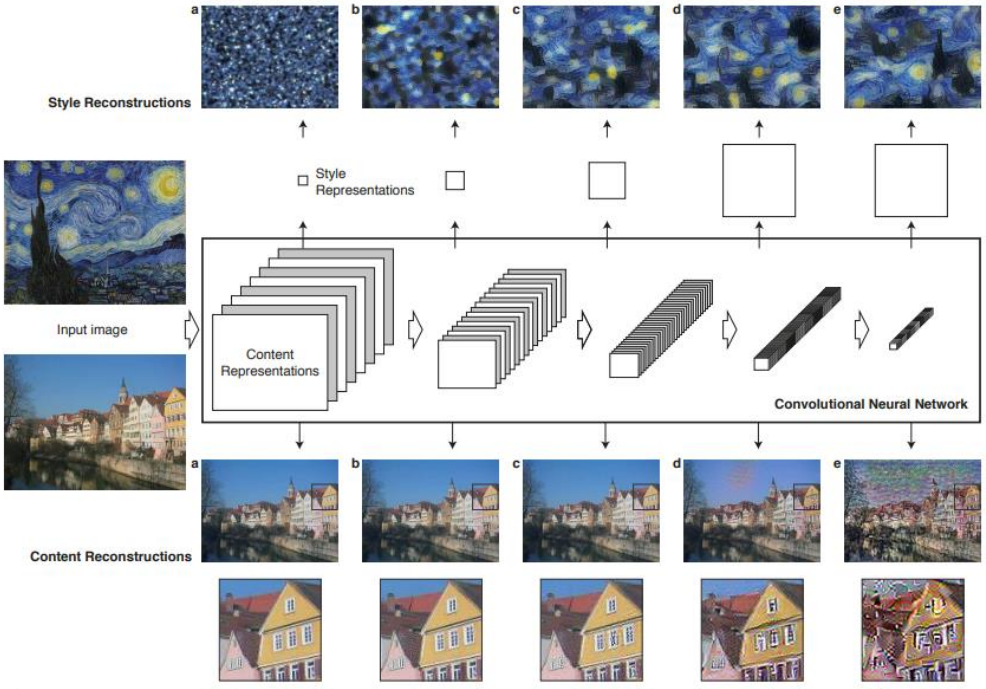


Figure 8. Image representations in a Convolutional Neural Network [60]

Neural style transfer (NST) has gained popularity in recent years [59–62]. After the emergence of neural style transfer (NST) in 2015 [60], some studies devoted to extending current NST algorithms to a variety of research branches [61–64]. There are four main topics widely studied by current NST research: image style transfer, audio style transfer, video style transfer, and text style transfer.

Image style transfer aims to transform images into synthetic artworks automatically. Numerous investigations strive to enhance general Neural Style Transfer (NST) algorithms

for specific image categories [62, 63]. One noteworthy study, instance style transfer, leverages instance segmentation techniques to exclusively stylize a single user-designated object within an image [64, 65]. Distinguishing itself from image style transfer, video style transfer necessitates a distinctive approach, given the imperative of ensuring seamless transitions between successive video frames [66, 67]. Meanwhile, audio style transfer extends the principles from image style transfer to the auditory domain, generating novel sounds by infusing the desired style from a target audio source [68, 69].

Text style transfer, which involves reshaping an input sentence into a desired style while retaining its original content, has garnered growing interest in recent times [70, 71]. The widespread utilization of Generative Adversarial Networks (GANs) [73] in image and voice generation [71, 72] has also catalyzed advancements in text style transfer. The proliferation of these deep generative models has fostered innovation in this realm. In this trajectory, Hu et al. [76] harnessed a neural generative model to attain a disentangled latent representation, enabling control over the sentiment and tense of the generated text. Simultaneously, Ficler and Goldberg [77] orchestrated a recurrent neural network language model to manipulate multiple linguistic style facets, achieved by conditioning the model on specific style and content parameters. Though adversarial network approaches have demonstrated their prowess across various research domains, their reliance on copious data can hinder performance in scenarios with limited data availability. Furthermore, the mechanics of applying adversarial networks for controlled text generation remain largely uncharted [72].

Text style transfer is a promising approach to transform a sentence with a target style meanwhile keeping the original content unchanged. There have been a few studies

conducting text style transfer approach that address automatic intervention [18]. The use of large pre-trained language models, such as BERT [78], GPT [79] and transformer [80] has become widespread in many downstream NLP tasks [81], for instance, text summarization [82] and offensive language detection [83]. Thus, exploring text style transfer for meeting this challenge is a promising but under explored research branch. To fill this gap, this study proposes a new text style transfer-based approach.

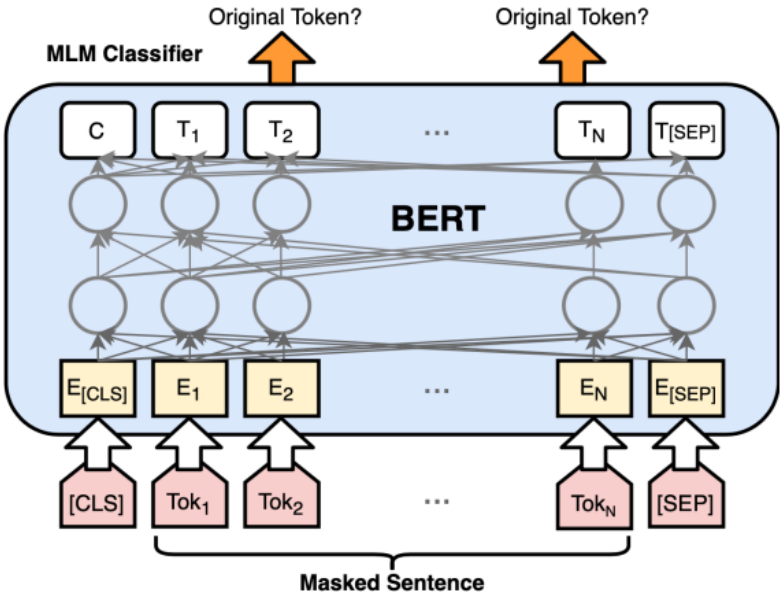


Figure 9. The fBERT masked language model [83]

To train a text style transfer-based offensive intervention model, we need a parallel offensive/non-offensive language dataset. Inspired by [95], we adopt a rule-based rephrasing method to collect a parallel offensive/non-offensive language dataset for training the model. The details of the dataset and data preprocessing will be introduced in the next chapter. The rule-based method is a simple yet effective method for generating data samples of corresponding offensive language data. We do this by first collecting a list of the top 200 offensive words from Hasebase, which is the world's largest authoritative

structured repository of hate speech [30]. in English as the keywords to collect candidate offensive sentences. Then we rephrase those candidate offensive sentences with the rules to generate corresponding sentences without offensiveness one by one. We identify and adopt the substitutions from [95]. For instance, “fxxk” is one of the top offensive words commonly used in social media posts. This word is usually used as a single word or in combination with other nouns (i.e., “fxxk” + nouns). In this case, attackers want to express their anger against others. A rule may use the term “darn” to substitute this word [95]. In another rule, we eliminate the offensiveness from some offensive words like “Bxtch” and “Nxgger”, which are used to attack specific groups of people, by substituting with words’ synonymous. For example, “Bxtch” will be substituted with “woman”. “Nxgger” will be substituted with “African American”. Some rules for rephrasing offensive words are listed in Table 3.

Table 3. Selected rules of offensive words substitution

Rules	Offensive Word	Substitution
1	fxxk	darn
2	shxt	darn
3	bxtch	woman
4	freak	strange
5	nxgger	black guy
6	ixiot	guy
7	sxupid	guy
8	ugly	unattractive
9	dxmn	omg
10	asxhole	unwise guy
11	loser	guy
12	pxssy	woman
13	silly	unwise
14	useless	unworkable
15	bullshxt	darn

3.2.3 A New Pipeline for Generating a Parallel Dataset

Despite the fact that a rule-based method is effective and simple, it could be improved by the generation of corresponding non-offensive sentences using large language models (LLMs). Recent LLMs are explicitly trained in a self-supervised way with the objective of predicting parts of the text. Given a sentence with one missing word, language models such as BERT [78], can suggest multiple words that match the context of the sentence. Using the useful property of language models, we can therefore automatically create corresponding non-offensive sentences without the need for prior knowledge of the specific rules. Thus, we proposed a new pipeline for generating parallel offensive/non-offensive datasets. There are three main components of our proposed pipeline of parallel datasets generation.

The first phase is defined as “find & replace”. In this step, we first use Detoxify [97], a pre-trained offensive language classifier to find the target offensive keywords in each input source text. Detoxify was trained on three large-scale offensive language datasets. After the target offensive keywords are found, those tokens will be replaced with “[Mask]” token in the source text. We input the new sentence to the masked language model which outputs 10 new candidates for the target keywords. For example, the bullshit was parsed and found in the sentence: “The whole article is bullshit.” We mask out this work to get: “The whole article is [Mask].” The model then predicts several candidate words: satire, fake, garbage, compelling, attractive, informative, incomplete, boring, timely and comprehensive. The masked language model is a Distil-Roberta model. This model is a distilled version of the RoBERTa-based model [89]. It follows the same training procedure as DistilBERT [98]. The model has 6 layers, 768 dimensions and 12 heads,

totalizing 82M parameters compared to 125M parameters for RoBERTa-base model that is twice as fast as Roberta-base model. The workflow for the first phase is described in Figure 10.

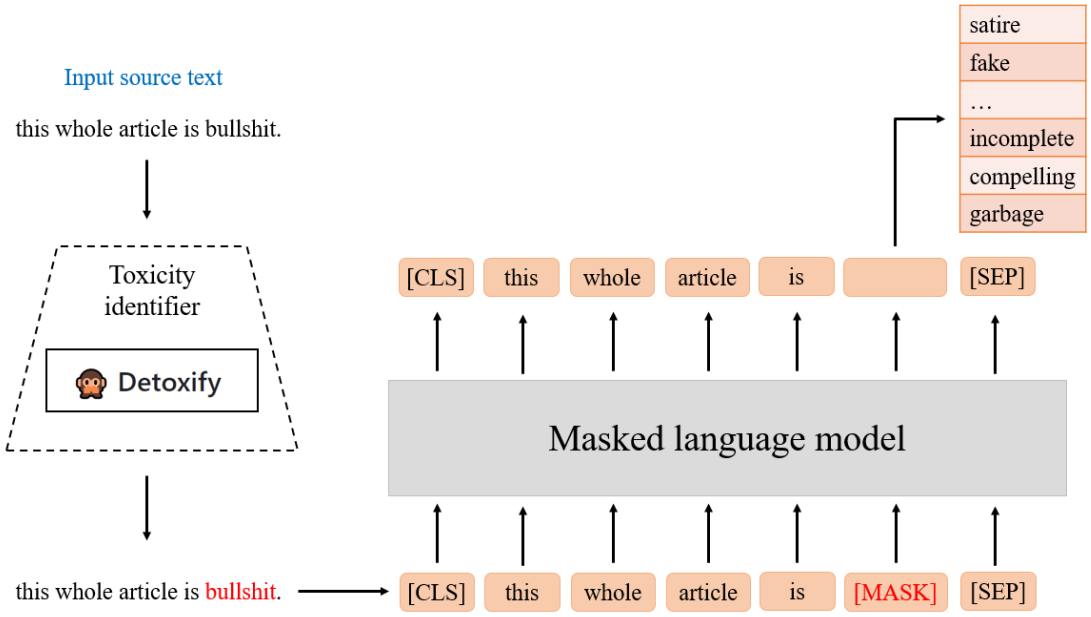


Figure 10. Generation of new candidate keywords from a masked LM

The second phase is defined as “refill & rank”. We fill in the source text with the new candidates to generate 10 new sentences. Using the same example in the first phase, we can get: “This whole article is satire.”, “This whole article is fake.”, “This whole article is incomplete.”, “This whole article is compelling.” and so on. We then use Detoxify to compute an offensiveness score for each candidate sentence. The range of the output score from Detoxify is between 0 and 1. If a score is higher, it indicates that the input text is more offensive. According to the scores that each sentence receives, we can rank all the candidate sentences. The workflow of the second phase is described in Figure 11.

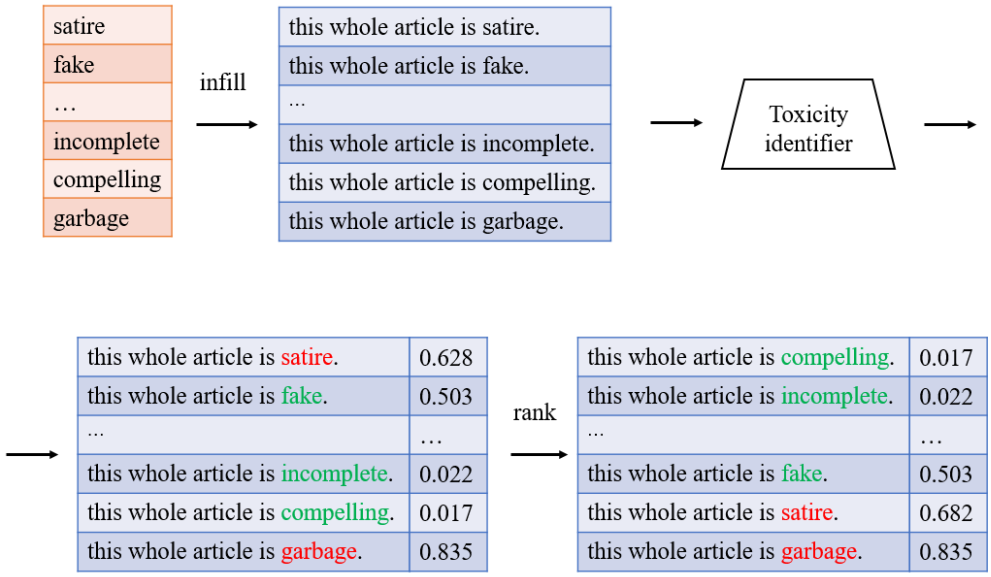


Figure 11. Generation of new candidate sentences and ranking them

The third phase is defined as “filter & drop”. In this step, the purpose is to select the best corresponding candidate sentence for the input source sentence. We first filter out the candidate sentences which are offensive. Then we use BERTScore [99] to measure the similarity between the non-offensive candidate sentences and the input source sentence.

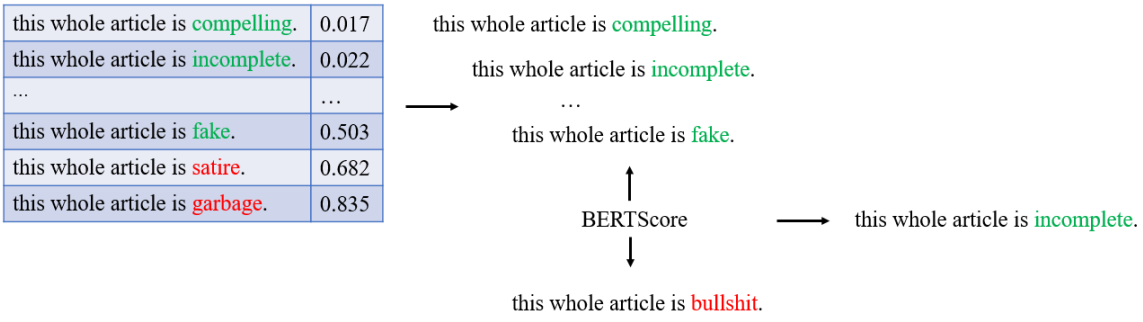


Figure 12. Selection of the best corresponding candidate sentence

BERTScore is a promising new metric for evaluating language generation models. It is a more robust metric than BLEU [100] and METEOR[101], which can underestimate the performance of a language generation model if the generated sentence is semantically correct but differs from the reference sentence in surface form [99]. BERTScore uses contextualized token embeddings to compute the similarity of two sentences [78]. Figure 13 shows the computation process of BERTScore.

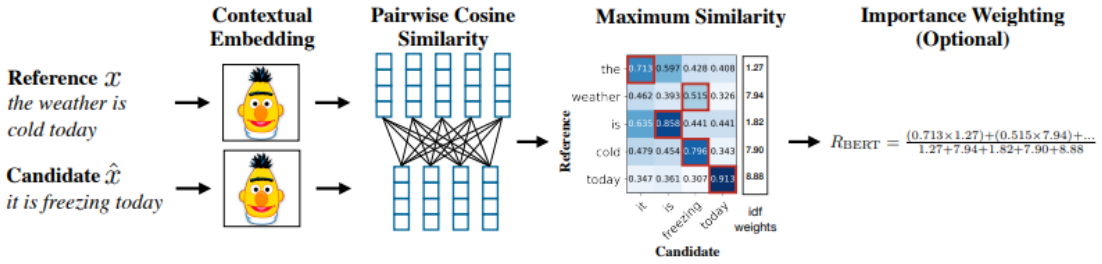


Figure 13. Description of the computation of the BERTScore [99]

Through the previous three steps, we can generate the corresponding non-offensive output target text for the input target text. After repeating these steps, we finally collect a parallel offensive/non-offensive dataset.

3.2.4 TST-IOC

As a sequence-to-sequence task, automatic offensive intervention can be performed with an encoder-decoder model trained with parallel data. Since a parallel offensive/non-offensive dataset does not occur in the wild. Such parallel datasets are extremely rare. When the parallel dataset is available, the majority of researchers use machine translation tools and large language models to perform style transfer [102]. We follow this practice by

training a vanilla transformer on our collected parallel dataset. The model training step is shown in Figure 14.

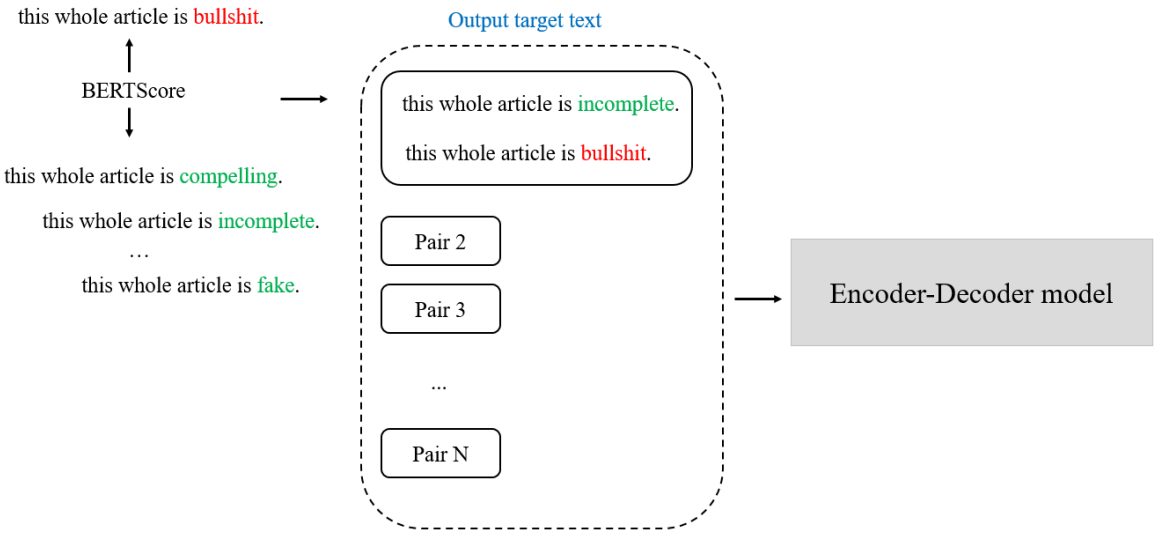


Figure 14. Training an encoder-decoder model on a parallel offensive/non-offensive dataset

The vanilla transformer model [80] is an encoder-decoder architecture that is trained end-to-end. It does not use any recurrent layers, instead relying on positional embedding to encode word order. Positional embedding is a vector that is added to each token in a sequence, encoding its absolute position in the sequence. This allows the model to learn how to attend to different tokens in the sequence, regardless of their order.

The encoder of the vanilla transformer is composed of 6 identical layers, each of which has two sub-layers. The first sub-layer is a multi-head self-attention mechanism. This mechanism allows the model to attend to all of the tokens in the sequence, simultaneously. The second sub-layer is a position-wise fully connected feed-forward network. This network applies a non-linear transformation to the output of the self-attention layer.

The two sub-layers of each encoder layer are connected by a residual connection and followed by a normalization layer. This helps to stabilize the training process and improve the performance of the model [103].

The decoder of the vanilla transformer is also composed of 6 identical layers. Each layer has the same three sub-layers as the encoder layer: 1) A multi-head self-attention mechanism, which allows the decoder to attend to its own output tokens; 2) A position-wise fully connected feed-forward network, which applies a non-linear transformation to the output of the self-attention layer; 3) A multi-head attention mechanism over the output of the encoder stack. This allows the decoder to attend to the input sequence, in order to generate the output sequence. The three sub-layers of each decoder layer are connected by residual connections and followed by normalization layers. This helps to stabilize the training process and improve the performance of the model.

To transform an offensive input sentence into a non-offensive output sentence, TST-IOC first uses the encoder of a vanilla transformer model to map the input sequence of word representations $T_i = (t_1, t_2, \dots, t_k, \dots, t_m)$ to a sequence of continuous representations $Z_i = (z_1, z_2, \dots, z_k, \dots, z_m)$. The decoder of the pre-trained model then generates an output sequence $Y_i = (t_1, t_2, \dots, t_k', \dots, t_m)$. The overview of our TST-IOC approach is shown in Figure 15.

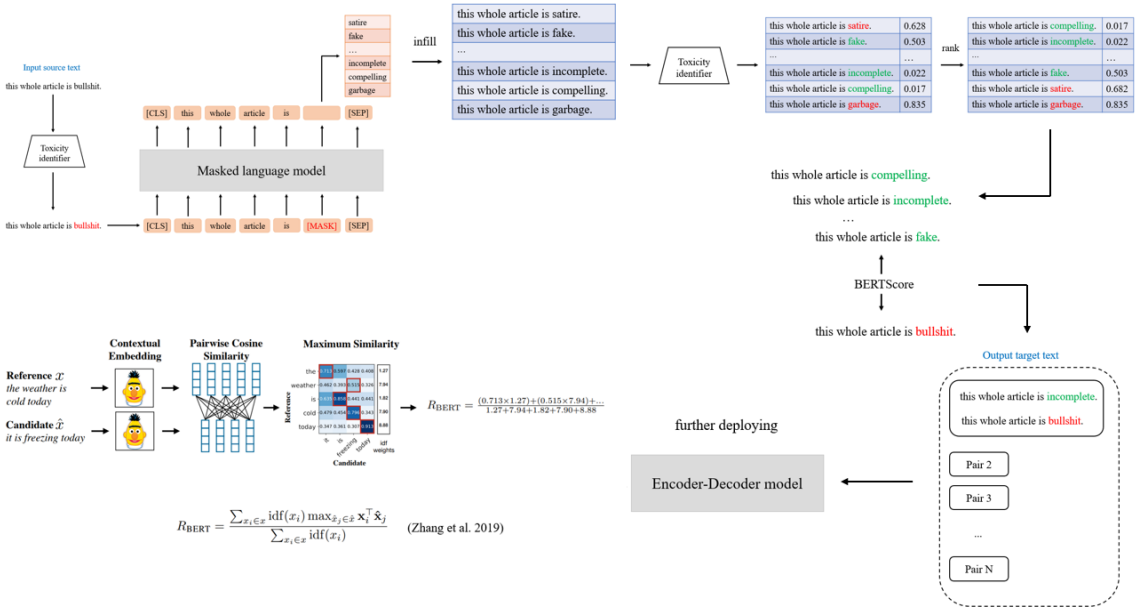


Figure 15. The overview of our TST-IOC approach

3.3 Summary

In this section, we first formulate the automatic offensive intervention problem and introduce a text style transfer-based approach for this task when a parallel dataset is not available. Secondly, we formulate the automatic offensive intervention problem when a parallel dataset is available and discuss the potential of utilizing the parallel corpora to address the automatic offensive intervention problem. Thirdly, we also introduce a rule-based rephrasing approach for the generation of a parallel offensive/non-offensive language dataset. Finally, we propose a novel text style transfer-based approach, TST-IOC step by step.

CHAPTER 4: EVALUATION

In this chapter, we first introduce the datasets that are collected and used for training the proposed models. Then we introduce the evaluation methods for the automatic intervention of offensive language, which consists of two parts. The first part is the objective evaluation of system performance. We describe the baselines, evaluation method, and performance metrics that are used to assess the performance of our proposed TST-IOC approach. The second part is subjective evaluation via a user study. In this user study, we also assess several new aspects, such as user satisfaction, acceptance, and perceived usefulness of the proposed models that the objective evaluation is unable to derive.

4.1 Data Collection

A parallel offensive/non-offensive language dataset contains multiple sentence pairs. Each sentence pair consists of an offensive sentence and its corresponding non-offensive sentence. Due to a lack of qualified parallel datasets for automatic intervention research, it is difficult to find an existing parallel dataset for supervised training. Preparing a quality dataset for this specific automatic intervention task is challenging. To overcome this challenge, this study introduces an adapted rule- and an LLM-based approach for the generation of a parallel offensive/non-offensive language dataset.

Utilizing an integrative data and analytics platform developed by the School of Data Science (SDS) at UNC Charlotte, we gathered Twitter data. This dataset comprises tweets related to Covid-19 that were generated from March 29, 2020 to April 15, 2020. As the 2020 pandemic evolved into a global crisis, social media users' sentiments and responses

have painted a vivid linguistic landscape in the realm of digital communication. The ripple effects of the pandemic have given rise to worldwide dialogues, within which prejudiced discourses linked to Covid-19 have taken root and persisted.

An illustrative instance lies in the aftermath of the former President Trump's initial usage of the racially charged term "Chinese virus" to refer to the coronavirus in a tweet dated March 17, 2020. Subsequently, an escalating number of individuals adopted this offensive phrase across social media platforms, thereby exacerbating hostility and promoting discrimination against Chinese Americans and other Asian communities. This troubling trend has led to an upsurge in hate crimes targeting these groups. Furthermore, the lockdown and quarantine also accelerate the number of people expressing their negative emotions or even provoke them to abuse others with offensive language on social media. Thus, we selected the period between March 29, 2020 and April 15, 2020 to collect our offensive language posts dataset. We first used 200 offensive keywords from Hatebase [31] to search for relevant tweets written in English. The collected tweets were preprocessed by removing punctuations, hashtags, retweets, and URLs in the textual data. We then used the cleaned offensive textual data to generate the pairwise non-offensive textual data by deploying our proposed LLM-based approach introduced in Chapter 3. Following this approach, we obtained a parallel offensive/non-offensive language dataset, which contains 88,278 pairs in which each offensive sentence has its pairwise non-offensive sentence, and the total size of the dataset is 176,556. This dataset is used for training the vanilla transformer model. We split the dataset into a training set (80%), a validation set (10%), and a test set (10%).

To prepare a non-parallel offensive/non-offensive language dataset for training the encoder-decoder model when a parallel dataset is not available, we reuse the offensive textual data collected for the generation of a parallel offensive/non-offensive language dataset. The non-offensive textual data were collected from two public datasets. The OLID dataset [96] addresses the challenge of detecting offensive language. This dataset consists of 13,241 training samples and 860 test samples. The size of non-offensive textual data in OLID is 10,080. To balance the sample size of two categories in the dataset, we collected non-offensive textual data from an offensive comment classification challenge dataset [104] which contains comments from Wikipedia annotated with the labels (offensive or non-offensive). This dataset contains 143,928 non-offensive sentences. The total size of the non-parallel offensive/non-offensive language dataset is 176,556. The information about the two public datasets is shown in Table 4.

Table 4. The distribution of Categories and tweets in two public datasets

Dataset	Categories	Total # of Tweets
OLID	Non-Offensive: 10,080	14,960
	Offensive: 4,880	
Toxic Comment Classification Challenge	Non-Offensive: 143,948	159,571
	Offensive: 15,623	

4.2 Objective Evaluation

4.2.1 The Framework of Systematic Evaluation Paradigm

Evaluating the performance of automatic intervention systems is crucial. Objective evaluation methods involve quantitative metrics and measurements that can be easily analyzed and compared. These methods are crucial for assessing various technical aspects

of automatic intervention systems and can provide an initial insight into the system's performance, helping developers identify potential bottlenecks or areas of improvement. However, due to the lack of automatic intervention studies, evaluation methods and performance measures of automatic intervention systems have been severely under explored. Although prior studies have proposed several automatic metrics, such as classification accuracy and content preservation, to evaluate automatic intervention methods from a quantitative perspective [18], a few studies have assessed the performance of automatic intervention systems and designed human evaluation via a user study. As a result, most of the prior studies assessed their models solely based on automatic evaluation metrics. Despite that many studies have conducted an automatic evaluation on text generation tasks [98,100,101], given the limitations of automatic evaluation, much previous work has involved human evaluation metrics for text generation, especially in machine translation [105]. In this study, we extend current automatic intervention research by proposing a systematic paradigm for evaluating the generated results of intervention methods from various perspectives and conducting a user study to investigate the user acceptance of the proposed automatic intervention systems.

For an automatic intervention task, one major purpose is to generate successful results, which are generated sentences without offensiveness from the original resources. The modified text must preserve the meaning of an original sentence while eliminating the offensiveness simultaneously. In [18], the study applied three quantitative metrics: 1) classification accuracy, which can measure if the generated sentences are transformed into non-offensive sentences successfully [31]; 2) content preservation, a metric that computes the content similarity between generated and original sentences; and 3) perplexity, which

is computed by a word-level LSTM language model trained by non-offensive training sentences. While evaluating these three aspects is not enough, a successful automatic offensive intervention process needs to remove offensiveness and keep generated results understandable. For instance, the core standard is the offensiveness degree of the generated results. Online users may also care about the understandability of the generated results after offensiveness is eliminated from the original sentences. Thus, this study considers the quality of the generated results of automatic intervention systems in the following perspectives:

- *Offensiveness removal*: To what degree is the offensiveness of the resource eliminated?
- *Content preservation*: To what degree the meaning of an original sentence is preserved?
- *Readability*: Is the generated alternative sentence fluent/grammatically, correct?

For each performance measure, we present metrics to evaluate the generated results. Table 5 lists the metrics.

Table 5. The performance metrics

Performance metrics	Metrics
Offensiveness removal	Transformed accuracy (TA)
Content preservation	BLEU
Readability	Grammatical accuracy (GA)

4.2.2 Performance Metrics

Evaluation is one of the key challenges in NLP text generation tasks such as machine translation, machine summarization, and text style transfer [97, 98]. The studies on the evaluation of style transfer are still very limited and need more comprehensive metrics. There exist no commonly accepted benchmark metrics to evaluate the performance of automatic intervention systems due to the lack of related research in this area. As we discussed in the previous paragraph, based on our proposed framework of a systematic evaluation paradigm, we will evaluate the quality of the generated results of automatic intervention systems from four perspectives, including offensiveness removal, content preservation and readability. For each perspective, we will introduce several metrics in the following paragraphs.

Offensiveness removal: To evaluate the effectiveness of the automatic offensive intervention, we treat this challenge as a binary classification task. Thus, we utilize a pre-trained offensive language classifier, Detoxify [97], to achieve the assessment. This classifier is a pre-trained state-of-the-art language model, and the model was trained on a large-scale offensive dataset which was released by Google in 2018. We calculate the average offensive score of the original input sentences and the generated output sentences with this classifier to show if the original offensive input is transformed to a non-offensive output successfully. We also calculate the accuracy of transformed input sentences (TA). Transformation accuracy is calculated as the following:

$$TA = \frac{Num\ of\ non-offensive\ output\ sentences}{Num\ of\ offensive\ input\ sentences} \quad (5)$$

Content preservation: To evaluate content preservation, we use BLEU score. BLEU [100] is among the most popular metrics that are used to compute the similarity

between model outputs and the ground truth based on word overlapping in many NLP studies [98, 100, 101]. A BLEU score is computed using a couple of n-grams modified precisions. Specifically,

$$BLEU = BP \cdot exp\left(\sum_{n=1}^{N} w_n \, log \, p_n\right) \quad (6)$$

Where p_n is the modified precision for n-grams; w_n is a weight between value 0 and 1; BP is the brevity penalty.

$$BP = \begin{cases} 1, & if \ c > r \\ e^{(1-r/c)}, & if \ c \leq r \end{cases} \quad (7)$$

In the formula, *c* is the length of the candidate sentence and *r* is the length of the reference sentence. Usually, BLEU uses $N = 4$ and $w_n = 1/N$.

Readability: To evaluate readability, we use grammatical accuracy (GA). Grammatical error correction [109] and context-free grammar [110] have been used to predict the grammaticality of a sentence in text generation tasks. In this study, we use a Python wrapper for language tool [111], an open-source grammar, style and spell corrector to check the accuracy of English grammar in an output non-offensive sentence. This language tool can calculate the number of errors in text.

4.2.3 Baselines

To investigate the performance of our proposed approach, we compare it to other automatic offensive intervention approaches:

- ***Fully removal (FR)/Partial removal (PR)*** [17] – As discussed in Chapter 2, these methods are widely used by some platforms, like YouTube and

World of Warcraft websites. The offensive words in a sentence are partially replaced with special characters (e.g., “f***”) or completely removed.

- ***Content control (CC)*** [112] – With this method, a sentence will be removed if it contains more offensive words than a pre-determined threshold[17]. We define three offensive words as a threshold.
- ***TST with a non-parallel dataset*** – We implement a non-parallel text style transfer-based model follows the previous text style research in other domain tasks [76, 86].
- ***T5 paraphraser*** [113] – Krishna et al. conduct a study on paraphrase generation and suggest that a general-purpose paraphraser could eliminate a style signal from input text.
- ***ParaGeDi*** [114] – ParaGeDi is the SOTA model which fine-tunes a T5 paraphraser with style-informed language models for automatic offensive intervention tasks.

4.3 A User Study

However, solely relying on objective metrics might not capture the full complexity of real-world scenarios and user experiences. Human evaluation methods, on the other hand, bring the critical human perspective into the assessment process. Integrating human evaluation ensures that automatic intervention systems align with human values and preferences, ethical norms, and legal regulations. It also addresses the limitations of objective metrics by accounting for factors that are challenging to quantify, such as user trust, system transparency, and the potential for unintended consequences. Beside the

objective evaluation, to understand users' acceptance of the generated results from the proposed automatic intervention models, we also assessed human performance [108] on the evaluation. In this study, we collected crowdsourced human judgments of generated results from Amazon Mechanical Turk (AMT). Amazon Mechanical Turk serves as a platform where virtual tasks demanding human intelligence are undertaken. This array of AMT services affords businesses the opportunity to tap into a scalable workforce as needed while providing workers with an array of numerous tasks to engage with at their convenience. It is an online labor market where employees are called workers. The workers are recruited for the execution of task, which is called Human Intelligence task (HIT) [115]. Many text generation tasks have been conducted on the AMT platform to minimize time and cost [105, 111, 112]. To receive high qualified evaluation results from AMT workers, we restricted AMT workers' qualifications, by recruiting workers with a HIT approval rate larger than 85%. The recruited workers are all fluent in English speaking.

We asked recruited workers to assess the generated results of the baselines and our proposed approach on the same aspects that are used in automatic evaluation: offensiveness removal, content preservation and readability. We also assessed additional aspects: user satisfaction, user acceptance, and user perception of the generated results. We selected 50 sentences from the test sets. We collected human judgments per sentence from the following six aspects:

Offensiveness removal: workers were asked to read the original input offensive sentence and the generated sentence from all intervention methods. Then they were asked to rate the agreement of if the generated sentence was non-offensive on a seven-point Likert

scale: (1) Strongly disagree; (2) Disagree; (3) More or less disagree; (4) Undecided; (5) More or less agree; (6) Agree; (7) Strongly agree.

Content preservation: workers were asked to read the original input offensive sentence and the generated sentence from all intervention methods. Then they were asked to rate the agreement of content preservation/if the content in original sentence was equivalent to the content in the generated sentence on a seven-point Likert scale: (1) Strongly disagree; (2) Disagree; (3) More or less disagree; (4) Undecided; (5) More or less agree; (6) Agree; (7) Strongly agree.

Readability: workers were asked to read the original input offensive sentence and the generated sentence from all intervention methods. Then they were asked to rate the agreement of Readability/if the generated sentence was readable on a seven-point Likert scale: (1) Strongly disagree; (2) Disagree; (3) More or less disagree; (4) Undecided; (5) More or less agree; (6) Agree; (7) Strongly agree.

User satisfaction: workers were asked to read the original input offensive sentence and the generated sentence from all intervention methods. Then they were asked to rate the agreement of if they were satisfied with the generated result based on a seven-point Likert scale: (1) Strongly disagree; (2) Disagree; (3) More or less disagree; (4) Undecided; (5) More or less agree; (6) Agree; (7) Strongly agree.

User acceptance: workers were asked to read the original input offensive sentence and the generated sentence from all intervention methods. Then they were asked to rate the agreement of if the generated result was accepted on a seven-point Likert scale: (1) Strongly disagree; (2) Disagree; (3) More or less disagree; (4) Undecided; (5) More or less agree; (6) Agree; (7) Strongly agree.

User perception: workers were asked to read the original input offensive sentence and the generated sentence from all intervention methods. Then they were asked to rate the agreement of if the generated result was perceived usefulness on a seven-point Likert scale: (1) Strongly disagree; (2) Disagree; (3) More or less disagree; (4) Undecided; (5) More or less agree; (6) Agree; (7) Strongly agree.

To understand which evaluation aspect is the most crucial to users for automatic intervention tasks, we also asked workers to rate all aspects: offensiveness removal, content preservation, readability, user satisfaction, user acceptance of the transformed content and user perception from 7 to 1. We then computed the average results from all workers as the final human performance for all baselines and our proposed approach. The larger values represent the results are better.

This human evaluation can not only measure the effectiveness of the proposed approach with metrics that objective evaluation is not able to measure but also has two objectives: first, we want to explore which perspectives of rephrased results are the most concerned from users' side. This result of human evaluation can be a guidance for designing a more comprehensive automatic intervention system than current one in the future study. Second, this result helps this study investigate if the proposed automatic offensive intervention method has achieved users' expectations which might have profound implications for designing a human-centric automatic intervention system in the future.

CHAPTER 5: RESULTS

5.1 Results of Objective Evaluation

In the objective evaluation section, we conducted a comprehensive evaluation from three perspectives: offensiveness removal, content preservation and readability. Our evaluation encompassed three fundamental perspectives: Offensiveness removal, Content preservation and Readability. Table 6 shows the objective evaluation scores for all baselines and our proposed method. Through an examination of each method's performance in these aspects, we can gain valuable insights into their strengths and limitations.

The offensiveness removal assesses how well an original offensive input is transformed to a non-offensive output. High offensiveness removal scores imply that a method can eliminate offensiveness from an original offensive input text successfully. Fully removal (FR) demonstrated exceptional performance for offensiveness removal, achieving an impressive score of 98.58%. This indicates that FR successfully removed the offensiveness in the original offensive input text, which is not surprising as the offensive keywords are completely removed. Our proposed approach, TST-IOC showed the best performance on offensiveness removal, with a score of 99.73%. This suggests that TST-IOC excelled in eliminating offensiveness for automatic intervention tasks. Conversely, T5 paraphraser obtained significantly lower offensiveness removal scores of 10.59%, respectively. This outcome indicates possible challenges faced by these methods in accurately transforming offensive text into non-offensive text.

The preservation score is a commonly employed metric for gauging the likeness between those crafted by human references and translations generated by machines. A

higher preservation score suggests a closer resemblance to human-like translations, though it is essential to consider the semantic quality and fluency of the output as well. In our evaluation analysis, we measure the preservation score between the rephrased output text and the source input text. On the other hand, the T5 paraphraser, TST non-parallel and ParaGeDi received a relatively lower preservation score of 63.97%, 0.61% and 18.23%, respectively. It could also imply that these text generation-based methods focused on generating more diverse paraphrases, potentially offering novel interpretations but could hurt the content preservation. Our proposed TST-IOC achieved a relatively high preservation score compared to those three methods.

The Grammatical Accuracy metric gauges the grammatical correctness of the translated or paraphrased output. A lower Readability score signifies better grammar accuracy, which is crucial for producing coherent and understandable text. The method TST non-parallel showcased the most favorable result in terms of grammar accuracy, achieving a score of 2.035. This implies that TST non-parallel excelled in generating grammatically sound rephrased text, which is a vital aspect in any NLP task. Meanwhile, methods FR and PR exhibited higher Readability scores of 5.098 and 5.083, respectively, suggesting potential challenges in maintaining grammatical correctness during the rephrasing process. TST-IOC beats FR and PR in the perspective of grammar accuracy.

Considering the collective analysis of all metrics, the TST-IOC method emerged as the top performer in this evaluation. With high scores in Offensiveness removal and Content preservation, coupled with a relatively low Readability score, TST-IOC demonstrated a well-rounded performance in translation and paraphrasing tasks.

Table 6. Objective evaluation results of different intervention methods (Numbers in bold indicate the best results)

Method	Offensiveness Removal	Content Preservation	Readability
FR	98.58	79.58	5.098
PR	34.37	79.28	5.083
CC	1.163	**97.12**	4.274
TST non-parallel	17.31	0.6075	2.035
T5 paraphraser	10.59	63.97	2.223
ParaGeDi	95.86	18.23	**1.713**
TST-IOC (our method)	**99.73**	78.54	4.345

5.2 Results of Human Evaluation

In the human evaluation section, we evaluated the performance of all baselines and our proposed method from six perspectives: offensiveness removal, content preservation, readability, user acceptance, user satisfaction and user perception. Table 7 offers an insightful analysis of various methods based on their performance in Offensiveness Removal, Content Preservation, and Readability. Upon examining the results, several notable patterns and trends emerge.

The Fully removal (FR) method, despite achieving an impressive Offensiveness Removal score of 5.60, seems to sacrifice Content Preservation score of 3.88 and Readability score of 2.34. This suggests that while it effectively filters out offensive language, it might inadvertently alter the intended meaning of the text, leading to a loss of coherence and comprehension.

Similarly, the Partial removal (PR) method performs well in Offensiveness Removal score of 5.44, but like the Fully removal (FR) method, it also experiences challenges in Content Preservation score of 3.34 and Readability score of 3.14. This indicates that it may struggle to strike a balance between removing offensive elements and preserving the original message's clarity and coherence. The Content control (CC) method demonstrates a notable Content Preservation score of 6.04, implying that it effectively retains the essence of the input text. However, it lags in Offensiveness Removal score of 2.02 and Readability score of 2.88, implying that it may not be as successful in removing offensive language and could produce text that is less fluent and more challenging to read.

The T5 paraphraser method achieves a relatively low Content Preservation score of 2.68 and Offensiveness Removal score of 1.34. This suggests that it may not be as effective in removing offensive content and maintaining the core meaning of the original text reasonably well.

The ParaGeDi method stands out as a strong performer with a high Content Preservation score of 4.98 and an Offensiveness Removal score of 5.06, indicating that it excels in both retaining the essence of the text and effectively filtering out offensive language. Additionally, it maintains a commendable Readability score of 3.86, implying that the processed text remains coherent and easy to comprehend.

Finally, the TST-IOC method emerges as a promising solution, boasting both an impressive Offensiveness Removal score of 5.82 and a Content Preservation score of 5.98. It strikes a commendable balance between filtering out offensive language while preserving the original message's essence. Moreover, it maintains a moderate Readability score of 3.46, suggesting that it produces text that is relatively fluent and understandable.

Table 7. Part 1 Human evaluation results of different intervention methods (Numbers in bold indicate the best results)

Method	Offensiveness Removal	Content Preservation	Readability
FR	5.60	3.88	2.34
PR	5.44	3.34	3.14
CC	2.02	**6.04**	2.88
TST non-parallel	2.22	1.46	3.76
T5 paraphraser	1.34	2.68	3.66
ParaGeDi	5.06	3.64	**3.86**
TST-IOC (our method)	**5.82**	5.98	3.46

For the other three perspectives, we can see from Table 8 that presents the results of various methods based on: Acceptance, Satisfaction, and Perception. These metrics offer valuable insights into users' feedback and overall experience with each method.

The Fully removal (FR) method, on the other hand, paints a contrasting picture. It receives the lowest Acceptance score of 1.78 and Satisfaction score of 1.64, implying that users are less inclined to adopt or appreciate this approach. Surprisingly, the Perception score of 2.78 is higher, indicating that users may perceive this method differently than their level of acceptance and satisfaction would suggest. Further investigation is needed to understand the reasons behind this discrepancy.

In contrast, the Partial removal (PR) method proves to be a notable performer with a high Acceptance score of 3.86 and a Satisfaction score of 3.54. Users appear to embrace this method and find it satisfactory in meeting their requirements. Moreover, the Perception score of 4.22 reinforces the positive sentiment, indicating that users hold this method in high regard and have a favorable perception of its performance.

Similarly, the Content control (CC) method demonstrates a moderate Acceptance score of 3.22 and a Satisfaction score of 3.02. However, its lower Perception score of 2.34 suggests that users may not perceive it as positively as the acceptance and satisfaction scores would indicate. This disconnect could prompt researchers to explore potential reasons behind this discrepancy and find ways to enhance user perception.

The TST non-parallel method yields a moderate Acceptance score of 2.92 and a Satisfaction score of 3.22. Interestingly, its Perception score (4.32) is substantially higher, indicating that users may perceive this method more positively than the acceptance and satisfaction scores would imply. This incongruity calls for further investigation into the factors driving user perception.

The T5 paraphraser method garners a relatively high Acceptance score of 4.15 and a moderately high Satisfaction score of 3.66, suggesting that users are generally willing to adopt and appreciate this approach. Moreover, the Perception score of 3.88 aligns with the overall positive sentiment, signifying that users have a favorable perception of this method's efficacy.

The ParaGeDi method stands out as a strong performer, receiving a high Acceptance score of 4.76 and a Satisfaction score of 4.98. Users highly embrace and find this method satisfactory, reflecting its effectiveness in meeting user needs. Furthermore, the Perception score of 4.66 emphasizes the positive user perception, reaffirming its value and potential as a preferred solution.

Our proposed method, TST-IOC excels across all metrics, with high Acceptance (4.88), Satisfaction (4.76), and Perception (4.84) scores. This indicates that users highly accept and are extremely satisfied with this method, aligning with their perception of its

exceptional performance. This promising outcome suggests that the TST-IOC method holds significant potential and could be a preferred choice among users for offensive intervention tasks.

Table 8. Part 2 Human evaluation results of different intervention methods (Numbers in bold indicate the best results)

Method	Acceptance	Satisfaction	Perception
FR	1.78	1.64	2.78
PR	3.86	3.54	4.22
CC	3.22	3.02	2.34
TST non-parallel	2.92	3.22	4.32
T5 paraphraser	4.14	3.66	3.88
ParaGeDi	4.76	**4.98**	4.66
TST-IOC (our method)	**4.88**	4.76	**4.84**

5.3 Summary

In this chapter, we first discussed the results and implications of the evaluation of various methods for offensive language intervention. The TST-IOC method stands out as a highly promising solution, excelling in offensive content filtering and content preservation while maintaining decent readability. Second, we presented the results of human evaluation which reveals trade-offs between Offensiveness Removal, Content Preservation, and Readability, with some methods excelling in specific aspects but falling short in others. Our proposed TST-IOC method is highlighted as a standout performer, offering potential as a preferred solution. Researchers can draw insights from the analysis

to optimize methods and enhance user satisfaction and perception. Future research should focus on addressing identified limitations and improving performance in areas such as transformed accuracy, content consistency, and grammatical correctness. The detailed analysis provides valuable guidance for researchers and practitioners in offensive language intervention, aiding in informed decision-making for specific application needs.

CHAPTER 6: DISCUSSION

6.1 Major Findings

In objective evaluation, T5 paraphraser displayed potential for generating diverse paraphrases but could benefit from improvements in grammar accuracy. Additionally, the ParaGeDi method showcased a high level of Transformed Accuracy, but its relatively low BLEU and GA scores suggest areas where optimization could be explored to enhance the overall performance. The TST-IOC method proves to be a highly promising solution, with an impressive Offensiveness Removal score of 5.82 and a Content Preservation score of 5.98. It excels in striking a commendable balance between effectively filtering out offensive language while preserving the essential meaning of the original message. Additionally, the method maintains a moderate Readability score of 3.46, indicating that the generated text is both fluent and understandable. This combination of outstanding offensive content filtering and content preservation, coupled with decent readability, positions the TST-IOC method as a strong candidate for various real-world applications.

In human evaluation, the analysis of these methods reveals the inherent trade-offs between Offensiveness Removal, Content Preservation, and Readability. While some methods excel in specific aspects, they may fall short in others. Our proposed TST-IOC method appears to be a promising solution, addressing multiple dimensions of the problem and offering the potential for more effective and balanced offensive content filtering. However, further exploration and testing are necessary to determine the best approach for specific applications and contexts. In addition, the analysis reveals diverse user experiences and perceptions with the evaluated methods. While some methods exhibit strong user acceptance and satisfaction, their perception might not be as positive. Conversely, other

methods receive more favorable user perceptions than the acceptance and satisfaction scores would imply. The ParaGeDi method and the TST-IOC method emerge as standout performers, garnering high scores across all metrics, suggesting their effectiveness and potential as preferred solutions. Researchers can draw valuable insights from these findings to further refine and optimize methods, ultimately enhancing user satisfaction and perception in real-world applications.

We also conducted a correlation analysis on the six human evaluation measures. The findings suggest that acceptance, satisfaction, and perception are closely related. Satisfaction has a high positive correlation with acceptance (0.8446) and perceived usefulness (0.8968), indicating that higher satisfaction levels are associated with greater acceptance and positive perception. It's important to note that correlation doesn't imply causation and further analysis is needed to understand their implications for the specific context.

The detailed analysis presented in this study provides valuable guidance to researchers and practitioners in the field of offensive language intervention. Depending on the specific requirements of their tasks, they can make informed decisions about the most suitable method for offensive language intervention applications.

Moreover, future research could focus on addressing the identified limitations of the evaluated methods, aiming to improve their performance in areas such as transformed accuracy, content consistency, and grammatical correctness. Fine-tuning and adaptation techniques may further elevate the overall effectiveness of these methods, making them even more valuable assets for offensive language intervention tasks.

6.2 Research Contributions

The major contributions of this study to offensive language study can be summarized as follows:

- The scarcity of naturally occurring parallel offensive/non-offensive datasets has been a persistent challenge. These types of parallel datasets, vital for offensive intervention research, are indeed a rarity. Addressing this gap, we have introduced an innovative pipeline designed to create parallel datasets encompassing offensive and non-offensive content. This novel approach not only addresses the problem in scarcity of parallel datasets, but also opens up new avenues for advancing offensive intervention studies.
- Through implementing a novel text style transfer-based approach to extend current offensive language research and improve the comprehension of the generalizability of automatic intervention models which has been rarely explicitly explored in previous research.
- In this study, evaluating the performance of automatic intervention methods is important. However, automatic intervention systems have suffered from a notable deficiency in studies examining intervention techniques, evaluation methodologies, and metrics to gauge system performance. We delve into a systematic evaluation paradigm and human assessment to gauge the effectiveness of automatic intervention systems. Our research expands the current knowledge on offensive language analysis by comprehensively exploring the evaluation of automatic offensive intervention methods, considering both quantitative and qualitative perspectives.
- This study conducts a user study to investigate the difference of evaluation between human judgment and automatic evaluation when assessing the rephrased results.

Human evaluation stands as an indispensable cornerstone in assessing the effectiveness of automatic offensive intervention methods. This user study enables researchers to investigate the user acceptance of the proposed automatic offensive intervention methods. In addition, this user study not only provides a guidance for designing an effective and comprehensive automatic offensive intervention system but also has profound implications for designing and evaluating a human-centric automatic offensive intervention system in the future.

6.3 Practical Implications

The ubiquity of social media has provided a free space for users to express their opinions and thoughts. Despite its advantages, the relentless use of social media has created certain issues that hamper the essence of it. Crowds such as offensive language pertaining to racism, sexism, ethnicity, and religion make it difficult to maintain the intricate balance between freedom of expression and the defense of human dignity. It is imperative to prevent the propagation of such offensive content on social media.

Social media users generate massive volumes of content at extraordinary speeds every day. Existing intervention strategies used by social media platforms to mitigate the propagation of offensive language are not efficient. The current automatic filtering system can break the readability of the original posts when removing offensive content directly. The other manual filtering system relies on a slow process of human moderation to remove offensive content is time and labor consuming. This study develops an efficient automatic intervention method for online communities, which could mitigate the spreading of offensive language on social media by leveraging advanced deep learning techniques. Thus, this research not only protects online users to avoid potential harm from offensive content

but also benefits online communities that provide a substitute to them instead of current inefficient content filtering strategies.

The field of automatic offensive intervention remains relatively unexplored, and there is a scarcity of research on the performance evaluation of automatic intervention systems. To address these gaps, we present a systematic paradigm for evaluating automatic intervention systems, aiming to pave the way for more effective and comprehensive developments in this area. This systematic evaluation paradigm will serve as a valuable guide for future studies.

Additionally, this study includes a user study to assess the effectiveness of automatic intervention systems. This investigation not only offers insights into user acceptance and perceptions but also holds significant implications for the future development of human-centric automatic intervention systems. By considering users' perspectives, we can ensure the creation of more tailored and user-friendly intervention solutions.

6.4 Limitations and Future Research

Future work in the field of offensive language intervention should focus on addressing the identified limitations of the evaluated methods to enhance their overall performance. The current objective evaluation metrics are borrowed from Machine Translation research. Optimization efforts on exploring a better evaluation method to access the overall performance of rephrasing output results are required. In human evaluation, the trade-offs between Offensiveness Removal, Content Preservation, and Readability underscore the need for more balanced approaches. While the TST-IOC method emerges as a promising solution, addressing multiple dimensions of the problem,

further exploration and testing are necessary to determine its applicability in specific contexts. Understanding user perception is crucial, as some methods may excel in acceptance and satisfaction but fall short in perceived effectiveness. The ParaGeDi and TST-IOC methods are notable performers, highlighting the potential of leveraging advanced language models as preferred solutions that warrant further investigation and optimization in the future.

We proposed a new pipeline for generating parallel offensive/non-offensive datasets. However, further user studies are needed to assess the quality of data generated by this pipeline. In addition, our proposed TST-IOC method processes offensive sentences that only contain a single offensive keyword. We will extend our proposed method with the capability to replace multiple offensive keywords simultaneously in the future.

In this dissertation, we evaluated the proposed method from six dimensions: offensiveness removal, content preservation, readability, user satisfaction, user acceptance, and user perception. In future studies, we may conduct additional user studies to obtain insights on the importance of those individual dimensions to offensive language intervention. This investigation can provide guidance on the design of evaluation methods for assessing offensive language intervention approaches in future.

The study's detailed analysis provides valuable guidance to researchers and practitioners in offensive language intervention. Tailoring the choice of method based on specific task requirements becomes crucial for successful implementation. To improve the evaluated methods, future research should focus on refining transformed accuracy, content consistency, and grammar accuracy. Fine-tuning and adaptation techniques offer promising avenues for elevating the overall effectiveness of these methods in real-world

applications, making them valuable assets for addressing offensive language challenges. By addressing these limitations and capitalizing on the findings, researchers can contribute to enhancing user satisfaction and perception while combating offensive language in diverse contexts.

REFERENCES

[1] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, 2012, pp. 71–80.

[2] E. Ferrara, "' Manipulation and abuse on social media' by Emilio Ferrara with Ching-man Au Yeung as coordinator," *ACM SIGWEB Newsl.*, no. Spring, pp. 1–9, 2015.

[3] M. Wiegand, J. Ruppenhofer, A. Schmidt, and C. Greenberg, "Inducing a lexicon of abusive words–a feature-based approach," 2018.

[4] B. Auxier and M. Anderson, "Social media use in 2021," *Pew Res. Cent.*, 2021.

[5] M. J. George *et al.*, "Young adolescents' digital technology use, perceived impairments, and well-being in a representative sample," *J. Pediatr.*, vol. 219, pp. 180–187, 2020.

[6] S. Hinduja and J. W. Patchin, "Connecting adolescent suicide to the severity of bullying and cyberbullying," *J. Sch. Violence*, vol. 18, no. 3, pp. 333–346, 2019.

[7] S. M. Croucher, T. Nguyen, and D. Rahmani, "Prejudice Toward Asian Americans in the Covid-19 Pandemic: The Effects of Social Media Use in the United States," *Front. Commun.*, vol. 5, 2020, doi: 10.3389/FCOMM.2020.00039.

[8] N. Yu, S. Pan, C. C. Yang, and J. Y. Tsai, "Exploring the Role of Media Sources on COVID-19-Related Discrimination Experiences and Concerns Among Asian People in the United States: Cross-Sectional Survey Study.," *J. Med. internet Res.*, vol. 22, no. 11, 2020, doi: 10.2196/21684.

[9] J. Wen, J. Aston, X. Liu, and T. Ying, “Effects of misleading media coverage on public health crisis: a case of the 2019 novel coronavirus outbreak in China,” vol. 31, no. 2. pp. 331–336, 2020, doi: 10.1080/13032917.2020.1730621.

[10] E. Macguire, “Anti-Asian hate continues to spread online amid Covid-19 pandemic,” *Aljazeera. Available online https//www. aljazeera. com/news/2020/04/anti-asian-hate-continues-spread-online-covid-19-pandemic-200405063015286. html (accessed May 6, 2020)*, 2020.

[11] J. D. Aten, “Long-term Covid-19 mental health effects for Asian Americans,” *Psychol. Today*, 2020.

[12] G. Xiang, B. Fan, L. Wang, J. Hong, and C. Rose, “Detecting offensive tweets via topical feature discovery over a large scale twitter corpus,” in *Proceedings of the 21st ACM international conference on Information and knowledge management*, 2012, pp. 1980–1984.

[13] W. Warner and J. Hirschberg, “Detecting hate speech on the world wide web,” in *Proceedings of the second workshop on language in social media*, 2012, pp. 19–26.

[14] I. Kwok and Y. Wang, “Locate the hate: Detecting tweets against blacks,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2013, vol. 27, no. 1.

[15] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, “Abusive language detection in online user content,” in *Proceedings of the 25th international conference on world wide web*, 2016, pp. 145–153.

[16] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate speech detection with comment embeddings," in *Proceedings of the 24th international conference on world wide web*, 2015, pp. 29–30.

[17] Z. Xu and S. Zhu, "Filtering offensive language in online communities using grammatical relations," in *Proceedings of the Seventh Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*, 2010, pp. 1–10.

[18] C. N. dos Santos, I. Melnyk, and I. Padhi, "Fighting offensive language on social media with unsupervised text style transfer," *arXiv Prepr. arXiv1805.07685*, 2018.

[19] T. Jay and K. Janschewitz, "The pragmatics of swearing," *J. politeness Res. Lang. Behav. Cult.*, vol. 4, no. 2, pp. 267–288, 2008, doi: 10.1515/JPLR.2008.013.

[20] N. Tarasova, "Classification of Hate Tweets and Their Reasons using SVM." 2016.

[21] Y. Chen, "DETECTING OFFENSIVE LANGUAGE IN SOCIAL MEDIAS FOR PROTECTION OF ADOLESCENT ONLINE SAFETY." 2011.

[22] N. Thompson, *Anti-Discriminatory Practice: Equality, Diversity and Social Justice*. 2012.

[23] S. Sood, J. Antin, and E. Churchill, "Profanity use in online communities," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 1481–1490.

[24] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran, "Deceiving google's perspective api built for detecting toxic comments," *arXiv Prepr. arXiv1702.08138*, 2017.

[25] R. Guermazi, M. Hammami, and A. Ben Hamadou, “Using a Semi-automatic Keyword Dictionary for Improving Violent Web Site Filtering,” *Signal-Image Technology and Internet-Based Systems*, no. 3. pp. 337–344, 2007, doi: 10.1109/SITIS.2007.137.

[26] L. G. McNamee, B. L. Peterson, and J. Peña, “A Call to Educate, Participate, Invoke and Indict: Understanding the Communication of Online Hate Groups,” *Commun. Monogr.*, vol. 77, no. 2, pp. 257–280, 2010, doi: 10.1080/03637751003758227.

[27] S. Agarwal and A. Sureka, “Using KNN and SVM Based One-Class Classifier for Detecting Online Radicalization on Twitter,” *International Conference on Distributed Computing and Internet Technology*. pp. 431–442, 2015, doi: 10.1007/978-3-319-14977-6_47.

[28] P. Fortuna and S. Nunes, “A Survey on Automatic Detection of Hate Speech in Text,” *acm Comput. Surv.*, vol. 51, no. 4, 2018, doi: 10.1145/3232676.

[29] C. Wu, Y. Qian, and R. Wilkes, “Anti-Asian discrimination and the Asian-white mental health gap during COVID-19,” *Ethn. Racial Stud.*, vol. 44, no. 5, pp. 819–835, 2021.

[30] J. Salminen *et al.*, “Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media,” in *Proceedings of the International AAAI Conference on Web and Social Media*, 2018, vol. 12, no. 1.

[31] T. Davidson, D. Warmsley, M. Macy, and I. Weber, “Automated hate speech detection and the problem of offensive language,” in *Proceedings of the International AAAI Conference on Web and Social Media*, 2017, vol. 11, no. 1.

[32] F. Del Vigna12, A. Cimino23, F. Dell’Orletta, M. Petrocchi, and M. Tesconi, “Hate me, hate me not: Hate speech detection on facebook,” in *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, 2017, pp. 86–95.

[33] S. Agarwal and A. Sureka, “Characterizing Linguistic Attributes for Automatic Classification of Intent Based Racist/Radicalized Posts on Tumblr Micro-Blogging Website.” 2017.

[34] P. Burnap and M. L. Williams, “Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making,” 2014.

[35] E. F. Unsvåg and B. Gambäck, “The effects of user features on twitter hate speech detection,” in *Proceedings of the 2nd workshop on abusive language online (ALW2)*, 2018, pp. 75–85.

[36] S. Liu and T. Forss, “Combining N-gram based Similarity Analysis with Sentiment Analysis in Web Content Classification,” *International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. pp. 530–537, 2014, doi: 10.5220/0005170305300537.

[37] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, “Reading text in the wild with convolutional neural networks,” *Int. J. Comput. Vis.*, vol. 116, no. 1, pp. 1–20, 2016.

[38] A. Rakhlin, “Convolutional neural networks for sentence classification,” *GitHub*, 2016.

[39] R. Johnson and T. Zhang, “Effective use of word order for text categorization with convolutional neural networks,” *arXiv Prepr. arXiv1412.1058*, 2014.

[40] S. Yuan, X. Wu, and Y. Xiang, “A Two Phase Deep Learning Model for Identifying Discrimination from Tweets.,” in *EDBT*, 2016, pp. 696–697.

[41] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, “Deep learning for hate speech detection in tweets,” in *Proceedings of the 26th international conference on World Wide Web companion*, 2017, pp. 759–760.

[42] P. Mishra, M. Del Tredici, H. Yannakoudakis, and E. Shutova, “Author Profiling for Abuse Detection,” *International Conference on Computational Linguistics*. pp. 1088–1098, 2018, [Online]. Available: https://pure.uva.nl/ws/files/49953886/C18_1093.pdf LK - https://academic.microsoft.com/paper/2839648288.

[43] J. H. Park and P. N. Fung, “One-step and Two-step Classification for Abusive Language Detection on Twitter,” *Meeting of the Association for Computational Linguistics*. pp. 41–45, 2017, doi: 10.18653/V1/W17-3006.

[44] V. Singh, A. Varshney, S. S. Akhtar, D. Vijay, and M. Shrivastava, “Aggression Detection on Social Media Text Using Deep Neural Networks.” pp. 43–50, 2018, doi: 10.18653/V1/W18-5106.

[45] C. Wang, “Interpreting Neural Network Hate Speech Classifiers.” pp. 86–92, 2018, doi: 10.18653/V1/W18-5111.

[46] R. (Roger) Hu, W. Dorris, N. Vishwamitra, F. Luo, and M. Costello, “On the Impact of Word Representation in Hate Speech and Offensive Language Detection and Explanation,” *Conference on Data and Application Security and Privacy*. pp. 171–173, 2020, doi: 10.1145/3374664.3379535.

[47] H. Qassim, A. Verma, and D. Feinzimer, “Compressed residual-VGG16 CNN model for big data places image recognition,” *IEEE Annual Computing and Communication Workshop and Conference*. pp. 169–175, 2018, doi: 10.1109/CCWC.2018.8301729.

[48] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, 1997.

[49] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2009.

[50] N. D. Gitari, Z. Zuping, Z. Zhang, H. Damien, and J. Long, “A Lexicon-based Approach for Hate Speech Detection,” *Multimedia and Ubiquitous Engineering*, vol. 10, no. 4. pp. 215–230, 2015, doi: 10.14257/IJMUE.2015.10.4.21.

[51] D. N. Perkins and G. Salomon, “Transfer of learning,” *Int. Encycl. Educ.*, vol. 2, pp. 6452–6457, 1992.

[52] J. M. Royer, “Theories of the transfer of learning,” *Educ. Psychol.*, vol. 14, no. 1, pp. 53–69, 1979.

[53] T. J. Nokes, “Testing three theories of knowledge transfer,” in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2004, vol. 26, no. 26.

[54] D. G. K. J. H. Boicho and N. Kokinov, *The analogical mind: Perspectives from cognitive science*. MIT press, 2001.

[55] J. R. Anderson, J. G. Greeno, P. J. Kline, and D. M. Neves, “Acquisition of problem-solving skill,” *Cogn. Ski. their Acquis.*, vol. 191, p. 230, 1981.

[56] S. Ohlsson, “Learning from performance errors.,” *Psychol. Rev.*, vol. 103, no. 2, p. 241, 1996.

[57] P. Mathur, R. Sawhney, M. Ayyar, and R. R. Shah, “Did you offend me? Classification of Offensive Tweets in Hinglish Language.” pp. 138–148, 2018, doi: 10.18653/V1/W18-5118.

[58] G. Wiedemann, E. Ruppert, R. Jindal, and C. Biemann, “Transfer Learning from LDA to BiLSTM-CNN for Offensive Language Detection in Twitter.” 2018, [Online]. Available: https://arxiv.org/pdf/1811.02906.pdf LK - https://academic.microsoft.com/paper/2900160274.

[59] M.-A. Rizoiu, T. Wang, G. Ferraro, and H. Suominen, “Transfer Learning for Hate Speech Detection in Social Media.” 2019, [Online]. Available: https://arxiv.org/pdf/1906.03829.pdf LK - https://academic.microsoft.com/paper/2948937448.

[60] L. A. Gatys, A. S. Ecker, and M. Bethge, “Image Style Transfer Using Convolutional Neural Networks,” *Computer Vision and Pattern Recognition*. pp. 2414–2423, 2016, doi: 10.1109/CVPR.2016.265.

[61] F. Luan, S. Paris, E. Shechtman, and K. Bala, “Deep Photo Style Transfer,” *Computer Vision and Pattern Recognition*. pp. 6997–7005, 2017, doi: 10.1109/CVPR.2017.740.

[62] Y. Li, N. Wang, J. Liu, and X. Hou, “Demystifying neural style transfer,” *arXiv Prepr. arXiv1701.01036*, 2017.

[63] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual Losses for Real-Time Style Transfer and Super-Resolution,” *European Conference on Computer Vision*. pp. 694–711, 2016, doi: 10.1007/978-3-319-46475-6_43.

[64] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, “Universal style transfer via feature transforms,” *arXiv Prepr. arXiv1705.08086*, 2017.

[65] C. Castillo, S. De, X. Han, B. Singh, A. K. Yadav, and T. Goldstein, “Son of zorn’s lemma: Targeted style transfer using instance-aware semantic segmentation,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 1348–1352.

[66] A. Selim, M. Elgharib, and L. Doyle, “Painting style transfer for head portraits using convolutional neural networks,” *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–18, 2016.

[67] M. Ruder, A. Dosovitskiy, and T. Brox, “Artistic style transfer for videos,” in *German conference on pattern recognition*, 2016, pp. 26–36.

[68] H. Huang *et al.*, “Real-time neural style transfer for videos,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 783–791.

[69] P. Verma and J. O. Smith, “Neural style transfer for audio spectograms,” *arXiv Prepr. arXiv1801.01589*, 2018.

[70] P. K. Mital, “Time domain neural audio style transfer,” *arXiv Prepr. arXiv1711.11160*, 2017.

[71] J. Li, R. Jia, H. He, and P. Liang, “Delete, retrieve, generate: A simple approach to sentiment and style transfer,” *arXiv Prepr. arXiv1804.06437*, 2018.

[72] Z. Fu, X. Tan, N. Peng, D. Zhao, and R. Yan, “Style transfer in text: Exploration and evaluation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, vol. 32, no. 1.

[73] I. J. Goodfellow *et al.*, “Generative Adversarial Networks,” *ArXiv*, vol. abs/1406.2, 2014.

[74] S. Wen, W. Liu, Y. Yang, T. Huang, and Z. Zeng, “Generating realistic videos from keyframes with concatenated GANs,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2337–2348, 2018.

[75] A. Siarohin, E. Sangineto, S. Lathuiliere, and N. Sebe, “Deformable gans for pose-based human image generation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3408–3416.

[76] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, “Toward controlled generation of text,” *International Conference on Machine Learning*. pp. 1587–1596, 2017, [Online]. Available: http://proceedings.mlr.press/v70/hu17e/hu17e.pdf LK - https://academic.microsoft.com/paper/2964222296.

[77] J. Ficler and Y. Goldberg, “Controlling Linguistic Style Aspects in Neural Language Generation.” pp. 94–104, 2017, doi: 10.18653/V1/W17-4912.

[78] J. Devlin, M.-W. Chang, K. Lee, and K. N. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *North American Chapter of the Association for Computational Linguistics*. pp. 4171–4186, 2018, doi: 10.18653/V1/N19-1423.

[79] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[80] A. Vaswani *et al.*, “Attention is all you need,” in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[81] J. Zhu *et al.*, “Incorporating BERT into Neural Machine Translation,” *International Conference on Learning Representations*. 2020, [Online]. Available: https://www.microsoft.com/en-us/research/uploads/prod/2020/04/2002.06823.pdf LK - https://academic.microsoft.com/paper/2994928925.

[82] N. Kalchbrenner and P. Blunsom, “Recurrent Continuous Translation Models,” *Empirical Methods in Natural Language Processing*. pp. 1700–1709, 2013, [Online]. Available: https://www.aclweb.org/anthology/D13-1176.pdf LK - https://academic.microsoft.com/paper/1753482797.

[83] D. Sarkar, M. Zampieri, T. Ranasinghe, and A. Ororbia, “FBERT: A Neural Transformer for Identifying Offensive Content,” *arXiv Prepr. arXiv2109.05074*, 2021.

[84] M. Han, O. Wu, and Z. Niu, “Unsupervised automatic text style transfer using lstm,” in *National CCF Conference on Natural Language Processing and Chinese Computing*, 2017, pp. 281–292.

[85] S. Prabhumoye, Y. Tsvetkov, R. Salakhutdinov, and A. W. Black, “Style transfer through back-translation,” *arXiv Prepr. arXiv1804.09000*, 2018.

[86] T. Shen, T. Lei, R. Barzilay, and T. S. Jaakkola, “Style Transfer from Non-Parallel Text by Cross-Alignment,” *Neural Information Processing Systems*, vol. 30. pp.

6830–6841, 2017, [Online]. Available: https://papers.nips.cc/paper/2017/file/2d2c8394e31101a261abf1784302bf75-Paper.pdf LK - https://academic.microsoft.com/paper/2963366196.

[87] G. Liu and J. Guo, “Bidirectional LSTM with attention mechanism and convolutional layer for text classification,” *Neurocomputing*, vol. 337, pp. 325–338, 2019.

[88] M. R. Kibria and M. A. Yousuf, “Context-driven Bengali Text Generation using Conditional Language Model,” *Stat. Optim. Inf. Comput.*, vol. 9, no. 2, pp. 334–350, 2021.

[89] Y. Liu *et al.*, “Roberta: A robustly optimized bert pretraining approach,” *arXiv Prepr. arXiv1907.11692*, 2019.

[90] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V Le, “Xlnet: Generalized autoregressive pretraining for language understanding,” *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[91] D. Araci, “Finbert: Financial sentiment analysis with pre-trained language models,” *arXiv Prepr. arXiv1908.10063*, 2019.

[92] I. Beltagy, K. Lo, and A. Cohan, “Scibert: A pretrained language model for scientific text,” *arXiv Prepr. arXiv1903.10676*, 2019.

[93] T. Caselli, V. Basile, J. Mitrović, and M. Granitzer, “Hatebert: Retraining bert for abusive language detection in english,” *arXiv Prepr. arXiv2010.12472*, 2020.

[94] O. J. Romero, A. Wang, J. Zimmerman, A. Steinfeld, and A. Tomasic, “A Task-Oriented Dialogue Architecture via Transformer Neural Language Models and

Symbolic Injection," in *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2021, pp. 438–444.

[95] H.-P. Su, Z.-J. Huang, H.-T. Chang, and C.-J. Lin, "Rephrasing profanity in chinese text," in *Proceedings of the First Workshop on Abusive Language Online*, 2017, pp. 18–24.

[96] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the Type and Target of Offensive Posts in Social Media," *North American Chapter of the Association for Computational Linguistics*. pp. 1415–1420, 2019, doi: 10.18653/V1/N19-1144.

[97] A. F. Akyürek, M. Y. Kocyigit, S. Paik, and D. Wijaya, "Challenges in measuring bias via open-ended language generation," *arXiv Prepr. arXiv2205.11601*, 2022.

[98] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv Prepr. arXiv1910.01108*, 2019.

[99] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv Prepr. arXiv1904.09675*, 2019.

[100] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," *Meeting of the Association for Computational Linguistics*. pp. 311–318, 2002, doi: 10.3115/1073083.1073135.

[101] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[102] V. Logacheva *et al.*, “Paradetox: Detoxification with parallel data,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 6804–6818.

[103] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” *arXiv Prepr. arXiv1607.06450*, 2016.

[104] S. Rosenthal, P. Atanasova, G. Karadzhov, M. Zampieri, and P. Nakov, “SOLID: A Large-Scale Semi-Supervised Dataset for Offensive Language Identification,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 915–928.

[105] K. Sakaguchi and B. Van Durme, “Efficient online scalar annotation with bounded support,” *arXiv Prepr. arXiv1806.01170*, 2018.

[106] K. Carlson, A. Riddell, and D. N. Rockmore, “Zero-Shot Style Transfer in Text Using Recurrent Neural Networks.” 2017.

[107] M. Artetxe, G. Labaka, E. Agirre, and K. Cho, “Unsupervised neural machine translation,” *International Conference on Learning Representations*. 2018.

[108] S. Rao and J. R. Tetreault, “Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer,” *North American Chapter of the Association for Computational Linguistics*, vol. 1. pp. 129–140, 2018, doi: 10.18653/V1/N18-1012.

[109] C. Napoles, K. Sakaguchi, and J. Tetreault, “There’s No Comparison: Reference-less Evaluation Metrics in Grammatical Error Correction,” *arXiv Prepr. arXiv1610.02124*, 2016.

[110] E. Charniak, “Statistical parsing with a context-free grammar and word statistics,” *National Conference on Artificial Intelligence*. pp. 598–603, 1997.

[111] N. G. Aramouni *et al.*, “Automating Customer Experience Agents’ Evaluation with Natural Language Processing,” *Memorias las JAIIO*, vol. 8, no. 1, pp. 66–69, 2022.

[112] J. Seering, T. Fang, L. Damasco, M. Chen, L. Sun, and G. Kaufman, “Designing user interface elements to improve the quality and civility of discourse in online commenting behaviors,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–14.

[113] K. Krishna, J. Wieting, and M. Iyyer, “Reformulating unsupervised style transfer as paraphrase generation,” *arXiv Prepr. arXiv2010.05700*, 2020.

[114] D. Dale *et al.*, “Text detoxification using large pre-trained neural models,” *arXiv Prepr. arXiv2109.08914*, 2021.

[115] G. Paolacci, J. Chandler, and P. G. Ipeirotis, “Running experiments on amazon mechanical turk,” *Judgm. Decis. Mak.*, vol. 5, no. 5, pp. 411–419, 2010.

[116] M. Karpinska, N. Akoury, and M. Iyyer, “The Perils of Using Mechanical Turk to Evaluate Open-Ended Text Generation,” *arXiv Prepr. arXiv2109.06835*, 2021.

[117] S. Levy, M. Saxon, and W. Y. Wang, “The Truth is Out There: Investigating Conspiracy Theories in Text Generation,” *arXiv Prepr. arXiv2101.00379*, 2021.

APPENDIX A: EXAMPLES OF OFFENSIVE LANGUAGE

ID	Offensive Language on Twitter
1	*You one fucking hypocrite your Twitter handle shows you write for washington post but your own newspaper is saying it as chinese virus*
2	*hey dipshit if you dont want people from NY flying to other places like your beloved Florida, maybe fucking shut down airports across the nation This is ridiculous.*
3	*I feel pretty fucked then cuz I started out fine.*
4	*After 3 days of feeling like shit. My doctor said to go get a COVID19 test at the MDVeip site.*
5	*Mnuchin will pull the same shit he did in 2008 during.*
6	*Twitter are you trying to keep us from looking at that shit.*

APPENDIX B: OTHER RELATED OFFENSIVE LANGUAGE DATA

Year	Resource	Size	Categories of Posts	References
2017	Twitter	24,802	Hate speech, offensive but not hate speech, or neither	[31]
2017	Facebook	17,567	No hate, weak hate, strong hate	[32]
2018	Twitter	1,650	Abusive (551) or non-abusive (1,099)	[3]
2018	YouTube, Facebook	137,098	Hateful (2,364), non-hateful comments (1,357)	[30]
2018	Twitter English	16,907	Racist, sexist or neither	[35]
	Twitter Portuguese	5,668	Hate speech, none	
2016	Tumblr, microblog	3,228	Racist or radicalized intent	[33]
2015	Blogs posting from Hate websites	100	Not hateful, weakly hateful and strongly hateful	[50]
2014	Web pages	165,000	Violence, racism, racist and hate	[36]
2017	Twitter	16K	Sexist (3,383), racist (1,972) and neither (rest)	[41]
2018	Twitter	16K	Sexist (3,383), racist (1,972) and neither (rest)	[42]
2017	Twitter	16K	Sexist (3,383), racist (1,972) and neither (rest)	[43]
	Twitter	6,909	Sexism, racism, neither and both	
2018	Twitter	24,802	Hate speech, offensive but not hate speech, or neither	[45]
2020	Twitter	24,802	Hate speech, offensive but not hate speech, or neither	[46]

APPENDIX C: COLLECTED OFFENSIVE KEYWORDS FROM HATEBASE

ahole	gayboy	pakie
anus	gaygirl	paky
ash0le	gays	pecker
ash0les	gayz	peeenus
asholes	God-damned	peeenusss
ass	h00r	peenus
Ass Monkey	h0ar	peinus
Assface	h0re	pen1s
assh0le	hells	penas
assh0lez	hoar	penis
asshole	hoor	penis-breath
assholes	hoore	penus
assholz	jackoff	penuus
asswipe	jap	Phuc
azzhole	japs	Phuck
bassterds	jerk-off	Phuk
bastard	jisim	Phuker
bastards	jiss	Phukker
bastardz	jizm	polac
basterds	jizz	polack
basterdz	knob	polak
Biatch	knobs	Poonani
bitch	knobz	pr1c
bitches	kunt	pr1ck
Blow Job	kunts	pr1k
boffing	kuntz	pusse
butthole	Lesbian	pussee
buttwipe	Lezzian	pussy
c0ck	Lipshits	puuke
c0cks	Lipshitz	puuker
c0k	masochist	queer
Carpet	masokist	queers
Muncher	massterbait	queerz
cawk	masstrbait	qweers
cawks	masstrbate	qweerz
Clit	masterbaiter	qweir
cnts	masterbate	recktum
cntz	masterbates	rectum
cock	Motha Fucker	retard
cockhead	Motha Fuker	sadist
cock-head	Motha Fukkah	scank
cocks		schlong
CockSucker		screwing
cock-sucker		

crap	Motha	semen
cum	Fukker	sex
cunt	Mother	sexy
cunts	Fucker	shit
cuntz	Mother	shits
dick	Fukah	shitter
faigs	Mother	Shitty
fart	Fuker	Shity
flipping the	Mother	shitz
bird	Fukkah	bitch
fuck	Mother	blowjob
fucker	Fukker	clit
fuckin	motherfucker	arschloch
fucking	Mutha	fuck
fucks	Fucker	shit
Fudge	Mutha Fukah	ass
Packer	Mutha Fuker	asshole
fuk	Mutha	gay
Fukah	Fukkah	oriface
Fuken	Mutha	orifice
fuker	Fukker	orifiss
Fukin	n1gr	packi
Fukk	nastt	packie
Fukkah	nigger	packy
Fukken	nigur	paki
Fukker	niiger	orgasim
Fukkin	niigr	orgasm
g00k	orafis	orgasum

APPENDIX D: IRB APPROVAL FOR USER STUDY

CHARLOTTE
DIVISION OF RESEARCH

To:	Zhihui Liu University of North Carolina at Charlotte
From:	Office of Research Protections and Integrity
Approval Date:	27-Jun-2023
RE:	Notice of Determination of Exemption
Exemption Category:	2
Study #:	IRB-23-0877
Study Title:	A Deep Learning Approach to Automatic Intervention of Online Offensive Content on Social Media to Improve Online Safety

This submission has been reviewed by the Office of Research Protections and Integrity (ORPI) and was determined to meet the Exempt category cited above under 45 CFR 46.104(d). This determination has no expiration or end date and is not subject to an annual continuing review. However, you are required to obtain approval for all changes to any aspect of this study before they can be implemented and to comply with the Investigator Responsibilities detailed below.

Your approved consent forms (if applicable) and other documents are available online at Submission Page.

APPENDIX E: MATERIALS OF HUMAN EVALUATION COLLECTION

Consent to Participate in a Research Study

Title of the Project: A Deep Learning Approach to Automatic Intervention of Online Offensive Content on Social Media to Improve Online Safety
Principal Investigator: Zhihui Liu, UNC Charlotte
Faculty Advisor: Dongsong Zhang, UNC Charlotte

You are invited to participate in a research study (IRB-23-0877). Participation in this research study is voluntary. The information provided is to give you key information to help you decide whether or not to participate.

- Our goal is to develop a deep learning model to rephrase offensive content. To evaluate the quality of the rephrased offensive content by our proposed deep learning model, we will conduct a user study to collect feedback from online users. The results of this study may protect online users from harmful content, and eventually build a safer online environment for everyone.
- You must have a HIT approval rate larger than 85% and are located in the US. Participants also must be age 18 or older to participate in this study.
- You are asked to complete an evaluation task. In this evaluation task, you will review a text pair. After reading the text pair, you need to answer several questions based on your best knowledge and give your best agreement from a seven-point Likert scale (1-strongly disagree; 2-disagree;3-somewhat disagree; 4-neither disagree nor agree; 5-somewhat agree; 6-agree; 7-strongly agree).
- The text pairs may have offensive words such as abusive, hateful, and/or toxic language. You should not participate in this study if you will be upset by or offended by reading phrases and words that are offensive.
- It will take you less than 3 minutes to complete the task.
- We do not believe that you will experience any physical risk from participating in this study.
- You will not benefit personally by participating in this study. What we learn about how people are evaluated the generated text from a deep learning model may be beneficial to others.
- You will receive compensation of $0.05 on AMTurk after you finish the entire task (one text pair for all the evaluation question). You may withdraw from the study; if you do, you will NOT receive any compensation.

- To ensure you are paying attention and completing the tasks carefully and thoughtfully, your work will be denied without payment if any of these two conditions ***are met/satisfied***: 1) you did not complete the entire task, ***including the task questions***; 2) you failed to give the correct answer to all the verification questions.

Your privacy will be protected, and confidentiality will be maintained to the extent possible. Your responses will temporarily be linked to your MTurk worker ID. We will protect your MTurk worker ID by storing it securely. The MTurk Worker ID information will be destroyed once we finish our data collection. Task responses will be stored separately with access to this information controlled and limited only to people who have approval to have access. After this study is complete, the data will be deleted.

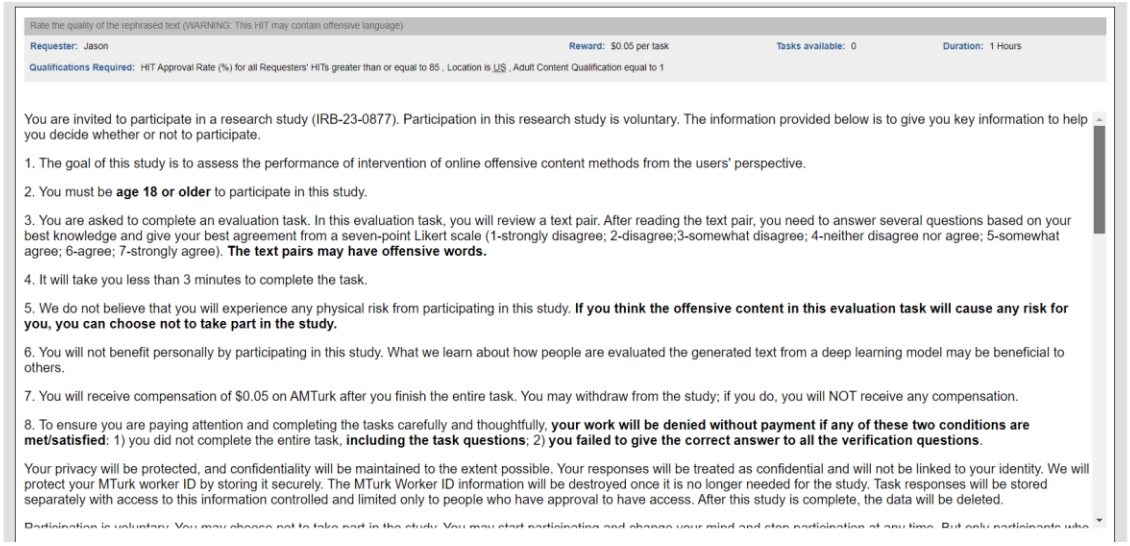
Participation is voluntary. You may choose not to take part in the study. You may start participating and change your mind and stop participation at any time, the incentive is only applicable for those who complete the study.

If you have questions concerning the study, contact the principal investigator, Zhihui Liu by email at zliu15@uncc.edu or Dr. Dongsong Zhang at dzhang15@uncc.edu. If you have further questions or concerns about your rights as a participant in this study, contact the Office of Research Protections and Integrity at (704) 687-1871 or uncc-irb@uncc.edu.

You may print a copy of this form. If you are 18 years of age or older, have read and understood the information provided and freely consent to participate in the study, you may proceed to the task by clicking the "**Agree**" button.


You may print a copy of this form. If you are 18 years of age or older, have read and understood the information provided and freely consent to participate in the study, you may proceed to the task by selecting the "**Agree**" item in the below box. **Only participants who select "Agree" will be involved.**

Agree



Rate the quality of the rephrased text (WARNING: This HIT may contain offensive language)

Requester: Jason **Reward:** $0.05 per task **Tasks available:** 0 **Duration:** 1 Hours

Qualifications Required: HIT Approval Rate (%) for all Requesters' HITs greater than or equal to 85 , Location is US , Adult Content Qualification equal to 1

You are invited to participate in a research study (IRB-23-0877). Participation in this research study is voluntary. The information provided below is to give you key information to help you decide whether or not to participate.

1. The goal of this study is to assess the performance of intervention of online offensive content methods from the users' perspective.

2. You must be **age 18 or older** to participate in this study.

3. You are asked to complete an evaluation task. In this evaluation task, you will review a text pair. After reading the text pair, you need to answer several questions based on your best knowledge and give your best agreement from a seven-point Likert scale (1-strongly disagree; 2-disagree;3-somewhat disagree; 4-neither disagree nor agree; 5-somewhat agree; 6-agree; 7-strongly agree). **The text pairs may have offensive words.**

4. It will take you less than 3 minutes to complete the task.

5. We do not believe that you will experience any physical risk from participating in this study. **If you think the offensive content in this evaluation task will cause any risk for you, you can choose not to take part in the study.**

6. You will not benefit personally by participating in this study. What we learn about how people are evaluated the generated text from a deep learning model may be beneficial to others.

7. You will receive compensation of $0.05 on AMTurk after you finish the entire task. You may withdraw from the study; if you do, you will NOT receive any compensation.

8. To ensure you are paying attention and completing the tasks carefully and thoughtfully, **your work will be denied without payment if any of these two conditions are met/satisfied**: 1) you did not complete the entire task, **including the task questions**; 2) **you failed to give the correct answer to all the verification questions**.

Your privacy will be protected, and confidentiality will be maintained to the extent possible. Your responses will be treated as confidential and will not be linked to your identity. We will protect your MTurk worker ID by storing it securely. The MTurk Worker ID information will be destroyed once it is no longer needed for the study. Task responses will be stored separately with access to this information controlled and limited only to people who have approval to have access. After this study is complete, the data will be deleted.

Participation is voluntary. You may choose not to take part in the study. You may start participating and change your mind and stop participation at any time. But only participants who complete the entire task can receive the rewards.

If you have questions concerning the study, contact the principal investigator, Zhihui Liu by email at zliu15@uncc.edu. If you have further questions or concerns about your rights as a participant in this study, contact the Office of Research Protections and Integrity at (704) 687-1871 or uncc-irb@uncc.edu.

You may print a copy of this form. If you are 18 years of age or older, have read and understood the information provided and freely consent to participate in the study, you may proceed to the task by selecting the **"Agree"** item in the below box. **Only participants who select "Agree" will be involved.**

Select

In this evaluation task, you will read a text pair that may contain offensive language. The text pair contains source text **(the original offensive sentence)** and its rephrased text. If you are not familiar with offensive language. Please refer to the definition of offensive language as below:

Abusive language: The term abusive language was used to refer to hurtful language and includes hate speech, derogatory language and also profanity.

Hateful language: includes any communication outside the law that disparages a person or a group on the basis of some characteristics such as race, color, ethnicity, gender, sexual orientation, nationality, and religion.

Toxic language: Toxic comments are rude, disrespectful or unreasonable messages that are likely to make a person leave a discussion.

Here is an example of a text pair:

Source text: f*ck you, I won't do what you tell me.

Rephrased text: I won't do what you tell me.

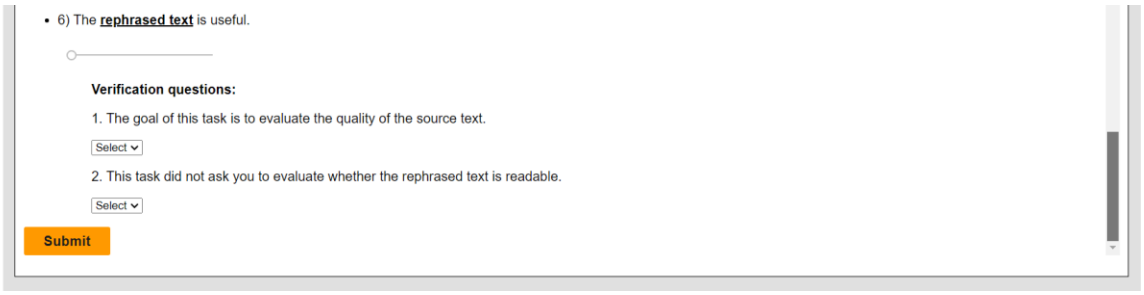
Task instruction:

Please read the two pieces of text below and use the sliders below indicate how much you agree with the statements (1 = Strongly disagree, 7 = Strongly agree)

Source Text: ${original_text}

Rephrased Text: ${rephrased_text}

- 1) The **rephrased text** is no longer offensive.
- 2) The content of the **rephrased text** is unchanged compared with the source text.
- 3) The **rephrased text** is readable.
- 4) I am satisfied with the **rephrased text**.
- 5) The changes to the **rephrased text** are acceptable.


• 6) The rephrased text is useful.
Verification questions:
1. The goal of this task is to evaluate the quality of the source text.
Select
2. This task did not ask you to evaluate whether the rephrased text is readable.
Select
Submit

APPENDIX F: THE CORRELATION MATRIX FOR HUMAN EVALUATION

	offensiveness	Content preservation	readability	acceptance	Satisfaction	perception
offensiveness	1					
Content preservation	0.6448	1				
readability	0.1294	-0.0720	1			
acceptance	-0.0184	-0.0404	0.1673	1		
Satisfaction	0.0321	0.0148	0.1692	0.8446	1	
perception	-0.0451	-0.0654	0.0398	0.8474	0.8968	1