EVALUATING ADVERSARIAL STYLOMETRY USING TEXTFOOLER

A COMPARATIVE ANALYSIS OF ADVERSARIAL ATTACK ON GENDER AND AGE USING THE REDDIT DATASET

YEXIN YANG

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

444535

COMMITTEE

dr. Chris Emmery
Westlake

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 24th, 2024

WORD COUNT

7279

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my thesis supervisor, Dr. Chris Emmery, for his invaluable guidance and support throughout this research. His expertise, patience, and insightful feedback have been instrumental in shaping this thesis and enhancing my academic and research skills. I would also like to extend my heartfelt thanks to my family for their unwavering support and encouragement during my studies, especially my partner, Kevin, whose constant support and understanding have been invaluable. Thank you all for your support and encouragement.

EVALUATING ADVERSARIAL STYLOMETRY USING TEXTFOOLER

A COMPARATIVE ANALYSIS OF ADVERSARIAL ATTACK ON GENDER AND AGE USING THE REDDIT DATASET

YEXIN YANG

Abstract

The advancement of author identification technology poses significant privacy risks on social media. Adversarial stylometry, which seeks to hide an author's identity by altering their writing style, is a strategy for privacy protection. However, existing research on adversarial stylometry is limited by the databases used. This study aims to evaluate the performance of TextFooler, a validated adversarial writing model, in challenging the widely-used pre-trained classification model BERT using the Reddit dataset. The results show that the BERT model achieved an accuracy of 0.648 for age prediction and 0.76 for gender prediction. Following the TextFooler attack, the accuracy of the target model decreased by 0.319 and 0.389 for gender and age prediction, respectively, demonstrating the effectiveness of the attack. The study also found that simply replacing individual words with synonyms is not always effective, as it can generate suspicious text due to contextual inconsistencies, suggesting that antonym replacements may sometimes be necessary. Future research in adversarial stylometry will require training on more diverse, cross-domain datasets and employing combined methods.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

The dataset is unpublished and owned by Dr. Chris Emmery, the supervisor of this thesis. This project does not involve collecting data from human participants or animals. A data agreement has been signed by the author of this thesis to use this dataset. The original owner of the data and code used in this thesis retains ownership of the data and code during and after the completion of this thesis. All figures were created by the author. All libraries and frameworks used are listed in the thesis. Grammarly was used for checking spelling and grammar. ChatGPT was used to improve the author's original content through paraphrasing and enhancing language. No other typesetting tools or services were used.

2 INTRODUCTION

Writing is a means of free expression, and each person's unique writing style enriches the diversity of human expression. Writing style includes vocabulary, syntax, structure, tone, rhythm, and more. These features not only reflect the thought process, expressive habits, and literary style of the author but also enable people to distinguish between different authors (Zheng et al., 2006).

With the development of natural language processing(NLP) and machine learning technologies, computers have expanded their focus from analysing text itself to identifying the authors of the text. For example, Ricciardi concluded through lexical analysis using several machine learning models that the real author behind Elena Ferrante's bestselling novel "L'amica geniale" (Ferrante, 2017) is Domenico Starnone.

These techniques, known as author profiling or author attribution, aim to reveal information about the author or determine the attribution. They are applied in many fields, from literature study to cybersecurity. The widespread use of the above techniques has brought new challenges to the protection of authors' privacy. The Internet provides a convenient platform for people to create, and people publish all kinds of information on it. Everyone can be an author of any kind. It also means that everyone's privacy protection may be affected by the author profiling technology.

Reddit is one of the most popular social media websites in the world, where registered users can anonymously post content.The total number of posts made by users on the site has reached millions. This massive traffic poses a significant threat to user privacy. Some users who have faced cyberbullying have had their identities exposed, leading to blackmail and even death threats in real life (Y. C. Zhang, 2020).

Adversarial writing involves modifying original texts to make text analysis and detection techniques ineffective, thereby protecting the author's privacy. The challenge of adversarial writing is to ensure an author's privacy and anonymity while preserving the original semantics and readability of the text. Texts on social media, with their informal language, ambiguity, and dynamic trends, pose challenges for traditional obfuscation techniques. This study aims to explore and enhance the robustness of existing adversarial writing models for texts on social media.

2.1 *Relevance*

Adversarial writing is crucial for privacy protection and cybersecurity. By changing lexical and syntactic features, individuals can hide their identities, maintaining anonymity and protecting sensitive information from unautho-

rized access. This technique is valuable in defending against cyber threats and ensuring the privacy of personal data. Research on adversarial writing aims to address these privacy concerns comprehensively. It investigates methods to effectively disguise identifying information, making it more challenging to trace or exploit personal data. Additionally, this research provides critical insights and guidance for developing more resilient author profiling models. By enhancing the robustness of these models, we can significantly improve cybersecurity measures, ensuring they can withstand sophisticated attempts at identity deception and data breaches. Overall, advancements in adversarial writing research contribute not only to individual privacy but also to the broader field of cybersecurity, making digital environments safer and more secure.

The scientific significance of this study lies in its exploration of TextFooler's performance on a previously unexplored, sizable dataset. This helps uncover additional potential of the model, addressing the research gap caused by the limited datasets used in earlier studies. By investigating how obfuscations can manipulate or deceive machine learning models, researchers can identify weaknesses and develop more resilient algorithms. To defend against adversarial attacks such as TextFooler-style word perturbations, Mein et al. developed FireBERT. Their approach protects against 95% of pre-generated adversarial samples while maintaining 98% of the original benchmark performance. This research is essential for enhancing the reliability and robustness of artificial intelligence systems. Research into adversarial writing contributes to our understanding of natural language processing, machine learning, and cybersecurity domains.

3 RELATED WORK

The literature review in this paper will cover four aspects: stylometry, author profiling on social media, adversarial stylometry and evaluation of obfuscation. The goal is to comprehensively explore the advancement of adversarial stylometry on social media and related research progress.

3.1 *Stylometry*

Through the study of writing styles, stylometry has been applied to areas such as author attribution, author profiling, and author identification. In the last century, at the early stages of its development, stylometry primarily focused on analysing classical literary works. Statistical models played a role in this period, addressing the authorship problem by exploring the distribution of different words counts (Mosteller & Wallace, 1963). With the integration of stylometry and machine learning, deep learning, stylometry

improves the ability to make high-level generalisations from textual data (Tweedie et al., 1996). Stylometry has been further developed in the digital age. With large amounts of accessible data, stylometry researches expand its focus from literature to various applications across different fields (Neal et al., 2017). In collaboration with forensic science, computational authorship analysis assists investigators in evaluating clues and provides linguistic evidence for judicial proceedings (Argamon, 2018). In another case, traditional plagiarism detection primarily relies on the lexical features of the text, allowing it to identify plagiarism solely based on lexical text similarity. However, when combined with deep learning algorithms, it becomes possible to detect plagiarism based on both lexical and semantic similarity, thereby enhancing accuracy (Ali & Taqa, 2022).

In addition to de-authorship anonymization, stylometry is also applied to determine demographic features of authors which known as author profiling, including gender, age, geographical origin, and even intelligence quotient (IQ) (Adebayo & Yampolskiy, 2022). This is also the focus of this research. Author profiling theories are based on the idea that the natural language people choose reflects their social identity and mental state. Demographic groups sharing similar characteristics tend to exhibit a certain level of consistency in their word use (Pennebaker et al., 2003). For instance, older people are more likely to employ obsolete expressions, syntax, or spelling, whereas younger people tend to use contemporary slang and abbreviations (Ehrhardt & Visconti, 2018). The differences in language style between men and women are evident in their word choices. Men tend to use profane language more frequently, whereas women often employ negative and emotionally charged words (Savoy, 2020). Researchers began author profiling studies using formal text corpora, such as Schler et al. exploring the British National Corpus, achieving an 80% accuracy in predicting author gender.

3.2 *Author Profiling on Social Media*

In the past 15 years, there has been a growing interest in applying author profiling to social media. The exploration of user data from these platforms holds significant commercial potential, particularly for large corporations aiming to gain deeper insights into user demographics. It enables them to design marketing strategies and advertisements for specific user groups, ultimately enhancing advertising effectiveness and overall user satisfaction (Chen & Skiena, 2014). Given that the dataset used in this study is sourced from Reddit, this part of the review focuses primarily on author profiling about social media.

In the beginning, most of the literature focused on feature engineering and traditional machine learning classifiers, asMiller et al. used n-gram feature representations with Perceptron and Naïve Bayes algorithms to predict gender on Twitter, achieving accuracies ranging from 90% to 100%. Also in the PAN author profiling challenge tasks of 2013 and 2015, participants used supervised machine learning methods such as decision trees, Support Vector Machines, and logistic regression to predict the gender and age of blog posts and Twitter posts(Rangel et al., 2013, 2015). The highest accuracy achieved was 95%. Stylistic features and content-based features played significant roles in the prediction process.

In recent years, some authors have been turning to deep learning. In the PAN author profiling task 2018 and 2019, some participants used word embeddings and character embeddings techniques to fully capture semantic and syntactic information, and few participants approached the task with Convolutional Neural Networks, Recurrent Neural Networks (RNNs), a voted LSTM and a BERT model (Rangel & Rosso, 2019; Rangel et al., 2018). Barlas and Stamatatos combined a multi-headed neural network language model with pre-trained language models on cross-domain attribution, and Bert and ELMo pre-trained models got the best results. Onikoyi et al. in their 2023 study, compared the performance of different word embedding models (such as GLOVE, BERT, GPT2, and Word2Vec) when combined with a machine learning model. They found that using advanced word embedding techniques like GloVE, BERT, and GPT2 significantly improved classification accuracy, yielding an accuracy range of 60-70%. This highlights the importance of employing such advanced techniques for enhancing the accuracy of classification models.

These studies indicate the increasing prevalence of deep learning in Stylometry. Beyond enhancing word embedding efficiency, more pre-trained models trained on large-scale corpora are being used.

In summary, due to variations in dataset sources, languages, and scales, researchers obtained differing accuracy results. Classical machine learning methods have demonstrated strong performance in gender and age classification, yet there is no definitive optimal feature and classifier combination. The formal research showed that in informal texts, content and stylistic features prove most effective. This insight inspires advancements in adversarial writing. The introduction of deep learning techniques brings new changes to stylometry, however further training on larger, more diverse datasets across different languages is needed to optimize deep learning models.

3.3 *Adversarial Stylometry*

At the same time, the extensive use of stylometry on social media has also raised concerns. Any information posted on social networks may be collected and used to reveal author attributes such as gender and age, and even to identify the author (Casimiro & Digiampietri, 2022). This raises serious privacy and anonymity issues for users, especially for minority ethnicities and whistleblowers, who may face heightened risks (Balakrishnan et al., 2021). While delving into stylometry research, scholars must conscientiously acknowledge the potential privacy threats associated with technological advancements.

Adversarial stylometry has become a new field of stylometry research. It involves attacking the author analysing models by changing the original texts. The framework for adversarial stylometry, consisting of obfuscation, imitation, and translation, was initially proposed by Brennan et al. Obfuscation involves using various methods to make an author's writing style less recognizable. Imitation entails altering the writing style to make it identifiable as that of a specific author. Machine translation involves translating back and forth between different languages to change the expression style. This study primarily focuses on obfuscation techniques, as they offer a more targeted approach to combating authorship profiling models and contribute to understanding methods for enhancing these models.

Under Brennan's frame, more research focuses on building obfuscations (Emmery et al., 2021; Lepekhin & Sharoff, 2021; Xing et al., 2024). There are four primary types of stylometric obfuscation: lexical, syntactic, morphological, and homograph obfuscation (Uchendu et al., 2023). Syntax obfuscation refers to modifying the structure of text to make it harder to understand. For example, Mutant-X (Mahmood et al., 2019) used mutation and crossover techniques to change the original text, and reduced the attribute accuracy by 37%.

Lexical obfuscation is the most extensively studied method achieved by words substitution. Word not only influences writing style but also reveals key content of the text, making it the most important feature in author profiling models. Many studies have built synonym substitution models, including Reddy and Knight successfully confused machine learning algorithms' gender predictions on Twitter and Yelp data using lexical substitution. TextFooler(Jin et al., 2020) used word importance ranking and word transformer, making it more efficient than other word substitution generators. EmotionFooler (Yang et al., 2022) builds upon TextFooler by incorporating part-of-speech and similarity score checks, improving the quality of generated samples and resulting in more natural output. Addi-

tionally, Emmery et al. used a transformerbased extension on TextFooler to achieve high transferability in wild.

Morphological obfuscation and homograph obfuscation are character-level attacks that disrupt the processing of analysis models by changing individual characters. While simple and effective (Gagiano et al., 2021), these methods are easily detected, leading to suspicion. Additionally, they can be easily countered by preprocessing techniques such as spell-checking (Wolff & Wolff, 2020).

3.4 *Evaluation of Obfuscation*

All these methods face the primary challenge of adversarial stylometry: how to obfuscate predictive models while minimizing the significant changes introduced by the obfuscation process(Gröndahl & Asokan, 2019). Potthast et al. defined the three dimensions of obfuscation evaluation as follows: safe, sound, and sensible. This implies that a good obfuscation method should effectively hide the author's identity, maintain semantic consistency, and remain inconspicuous. These dimensions will be emphasized when evaluating the performance of obfuscation in this study.

The effectiveness of obfuscation is relatively easy to measure by observing changes in the accuracy of target models. However, measuring the safety and sensibility of obfuscation is more complex. Both automatic machine evaluation and human evaluation have been used by researchers to address this.Various metrics, such as MAUVE, METEOR, USE Cosine Similarity, and BERTScore, have been employed to measure semantic preservation (Banerjee & Lavie, 2005; Jin et al., 2020; Mahmood et al., 2019; Xing et al., 2024), and language tools are used to detect grammatical errors.

Human evaluation remains a crucial method for assessing obfuscation, especially when the ultimate target of the obfuscation is humans rather than models. For instance, while Emmery et al. simplified the evaluation of semantic preservation using human reader.

In contrast to traditional stylometry, adversarial stylometry introduces an additional challenge beyond the algorithm, determining its application scenarios. For example, how would non-technical users apply obfuscation techniques(Z. Wang et al., 2022)? In recent years, the development of large-scale language models(LLM) has led to the proliferation of neural generated text. Adversarial stylometry can provide insights into machine generated text (MGT) detection (Macko et al., 2024). The future development of adversarial stylometry can draw inspiration from the models of adversarial attacks in image and audio (W. E. Zhang et al., 2020), but it also requires more practical application research on textual data.

3.5 *Summary*

This review has summarized the development of stylometry and its application on social media, with a focus on the evolution of adversarial stylometry and the evaluation of obfuscation techniques. The progress in those fields is closely linked to advancements in natural language processing (NLP) technology. The potential of adversarial stylometry in protecting data privacy is significant.

Unlike traditional stylometry, adversarial stylometry introduces an additional challenge: determining its application scenarios. For example, how can non-technical users apply obfuscation techniques (H. Wang et al., 2020)? The proliferation of large-scale language models (LLM) has led to an increase in neural-generated text. Adversarial stylometry can provide insights into machine-generated text (MGT) detection (Macko et al., 2024). Future development of adversarial stylometry can draw inspiration from adversarial attack models in image and audio (W. E. Zhang et al., 2020), but it also requires more practical application research on textual data.

These hopes for the future development of adversarial stylometry necessitate training, testing, and optimizing existing obfuscation techniques on larger, cross-domain, and multilingual datasets. Currently, research on text obfuscation techniques often relies on relatively small and similar datasets, such as Yelp, Twitter, and IMDB(Modupe et al., 2022; Potthast et al., 2016, 2018; Uchendu et al., 2023; Weinsberg et al., 2012). This study aims to address this research gap by using our Reddit dataset.

3.6 *Motivation of Method*

3.6.1 *Attack Model*

Based on the literature, future studies in adversarial writing will focus more on lexical obfuscation. This obfuscation method shows better performance in preserving semantic and grammatical accuracy while avoiding detection. Therefore, this study will focus on an adversarial method based on lexical obfuscation.

TextFooler, introduced by Jin et al., is one of the most popular lexical obfuscation techniques. It employs two key mechanisms for generating adversarial samples. Firstly, it selects words based on their impact on prediction changes when removed. Secondly, it replaces words by selecting candidates with similar meanings using cosine similarity and verifies them based on part-of-speech and semantic similarity. These replacements are chosen to maintain the original sentence's semantic and grammatical coherence until they influence the predictions of the target model.

TextFooler's capability to attack effectively without requiring access to the target model's architecture, parameters, or training data makes it suitable for various applications.

This study chooses TextFooler as the attack model due to its extensive research and application potential. TextFooler has demonstrated strong performance in non-target classification (Crothers et al., 2022; Lepekhin & Sharoff, 2021; Li et al., 2022; Morris et al., 2020). Additionally, researchers have developed various extended models based on TextFooler. For example,Kwon introduced Friend-Guard TextFooler and dual-targeted TextFooler, and Yang et al. provided EmotionFooler, which improved part-of-speech and similarity checking. This means the research on Textfooler may contribute more to the further research.

3.6.2 *Target Model*

Given that this research emphasizes the performance of the attack model on a larger, novel dataset, the selection of the target model prioritizes widely-adopted models over novel ones. Well-established models ensure the broader applicability and relevance of our findings within the field.

BERT (Bidirectional Encoder Representations from Transformers) is a Transformer-based model introduced by Google AI in 2018(Devlin et al., 2018). BERT uses multi-head self-attention mechanisms and feed-forward networks to simultaneously consider the relationships between each word and all other words in a sentence, thereby enhancing its ability to capture long-range dependencies. During the pre-training phase, BERT uses masked language modeling (MLM) and next sentence prediction (NSP) tasks, which enable it to capture bidirectional contextual relationships in text and generate high-quality contextual embeddings. These characteristics make BERT perform well in tasks such as text classification, text generation, and question answering systems, making it one of the key models in the field of natural language processing (NLP).

3.7 *Research Questions*

Combining the research gap and research objectives, the research question that this study aims to answer is:

> *How effectively does TextFooler protect the gender and age information of authors in our Reddit dataset?*

The following sub-questions need to be considered to answer the main research question:

RQ1 *How accurately can the initial author profiling model predict the gender and age of the author?*

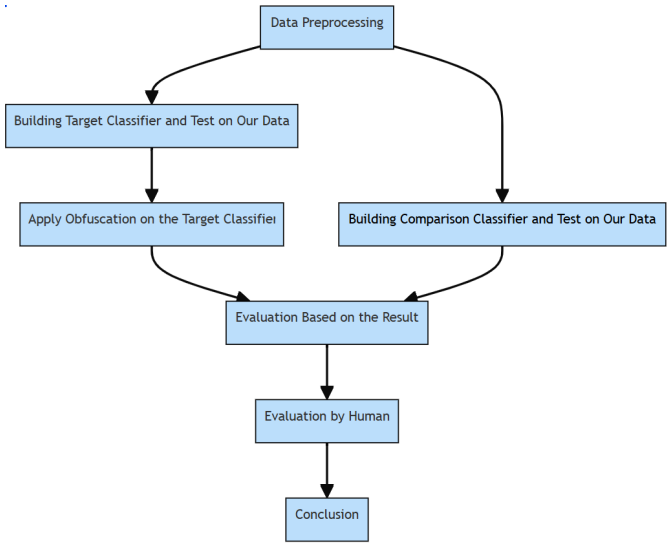

Figure 1: The experimental pipeline

RQ2 *How does the accuracy of the initial author profiling model change after the application of TextFooler?*

RQ3 *To what extent does TextFooler keep the original semantics and readability of the text?*

4 METHOD

This section outlines the specific methods employed in this study, including data collection and preprocessing, software and packages used, and detailed experimental procedures. Figure 1 below illustrates the experimental pipeline.

4.1 *Data*

The dataset used for this research is an unpublished dataset comprising historical data from Reddit, covering a period of two year in total, from 2020 to 2022. Reddit is among the most widely-used social media platforms in the world. As of 2023, it records an average of 73.1 million daily active users and 268 million weekly active users. The platform is driven by posts from registered users who contribute to discussions across over 100,000 individual communities known as subreddits. With its vast array of content, Reddit has amassed over 1 billion posts to date.

The data is organized in a structured format, and stored as a CSV file like table 1. It consists of over 40,000 records, each representing an

Table 1: The structure of the gender dataset

Author_ID	Post	Female
t2_rnjzutp	plants already and the fence on.....	1
t2_bbe41	my muscle memory. And noticed I could.....	0
t2_10dakq	they have disadvantage from fear.....	0

individual post, accompanied by the author ID and labelled features. For example: age and gender, nationality, personality, and political leaning. The nationality, personality, and political leaning are extracted from self-reports and tags in the subreddits. Gender and age are extracted from self-reports in posts. For example, "I (23f) went on a date..." indicates that the author is a 23-year-old female. Due to data from different years, to maintain consistency, the reported age is subtracted from the year of the post to determine the birth year. The author ID as an identifier, represents a unique author. Given the possibility of multiple posts from a single author, there could be multiple posts associated with the same author ID. Table 1 shows the basic structure of the raw data, with only partial text from the posts shown due to length limitation. In this dataset, gender is binary, comprising only male and female categories, without inclusion of LGBTQ individuals. The last column displays gender, which has been encoded using boolean coding, where 1 represents female and 0 represents male.

4.2 *Preprocessing*

Previous studies on author profiling have employed various data preprocessing methods, such as removing tags, stop words, noisy words, and URLs to reduce noise and improve accuracy (Rangel et al., 2013). The text data used in this study is sourced from Reddit, where user posts contain many informal language elements, such as abbreviations, grammar errors, and emojis. This study adopts minimal preprocessing in an attempt to preserve the original style of the authors' texts as much as possible. Additionally, Alzahrani and Jololian found that the BERT model achieves the highest accuracy when no preprocessing is applied to the Twitter text.

In processing the textual data, an initial step involved converting all posts to lowercase and removing any URLs. As previously mentioned, gender and age information were extracted from users' self-report within their posts, potentially revealing their attributes directly through the text. Thus, the segments in the AAG or GAA format within the posts were deleted.

Due to varying post counts per author, measures were taken to balance their impact on the dataset. A maximum cap of 20 posts per author was

enforced, with the selection process randomizing the retention of posts. Furthermore, to standardize the dataset, each post's character count was restricted to a maximum of 500 characters. The dataset is split into training and testing sets, with 80% for training model and 20% for testing.

For the age attribute, the dataset was divided into three age groups: youth(0-25 years), adult(26–45 years), and middle age and above (46 years and above). These age ranges correspond to different life stages, which may influence the language and topics discussed in the text.

4.3 *Algorithms and Software*

The implementation was carried out using Python language. Key libraries included scikit-learn(Pedregosa et al., 2011) for Naive Bayes and evaluation, and Hugging Face's Transformers for BERT-base uncased(Wolf et al., 2020). Some common packages used were pandas(McKinney, 2010), NumPy(Harris et al., 2020) for data manipulation. Matplotlib(Hunter, 2007) was used for data visualization and plotting results. Transformers and PyTorch(Paszke et al., 2019) were used for building BERT-base uncased model. TextAttack(Morris et al., 2020) and TensorFlow(Abadi et al., 2016) were used for adversarial attacks.

The entire pipeline, from data preprocessing to model training and evaluation, was implemented and executed in VS Code and Google Colab with GPU support for efficient computation.

4.4 *Target Classifier*

The preprocessed data was used to train and test the target classifier.

The Naive Bayes classifier was chosen as the baseline model due to its established efficiency and simplicity in previous studies (Aborisade & Anwar, 2018; Jockers & Witten, 2010). Textual data extracted from posts was transformed into feature vectors using CountVectorizer with an n-gram range of (1,2). The dataset was partitioned into training (80%), and test (20%) sets. Hyperparameter tuning was performed using Grid Search with 5-fold cross-validation to identify the optimal parameters for the Naive Bayes model.

BERT (Bidirectional Encoder Representations from Transformers) is renowned for its effectiveness across various natural language processing (NLP) tasks. The BERT base uncased model, comprising 12 transformer blocks, 768 hidden layers, and 12 attention heads with a total of 110 million parameters, was employed in this study. Token extraction was performed using the BERT tokenizer. The models were fine-tuned for ten epochs, with a batch size of 16, a learning rate of 2e-05, and a maximum sequence

length of 128. Early stopping with patience of 2 epochs was employed to prevent overfitting. The training process included regularization through the addition of a dropout layer with a dropout rate of 0.3. Optimization was performed using the Adam W optimizer.

4.5 *Obfuscation*

The obfuscation method involves altering the original text to obscure the writing style of the author, making it challenging for the target model to correctly classify the author's demographic attributes. This study uses the TextAttack(Morris et al., 2020) framework to implement and evaluate the obfuscation techniques. TextAttack provides a comprehensive set of tools and algorithms for generating adversarial examples in NLP. TextAttack has four components, a goal function, a set of constraints, a transformation, and a search method. These four components can be combined with 16 adversarial attacks, which are called "recipes". This versatility enables TextAttack to interact with various models and datasets. Through such flexible combinations, researchers and developers can easily customize and implement adversarial attacks tailored to different application scenarios and requirements.

TextFooler was selected as the primary obfuscation model. The WordSwapEmbedding technique was used to replace words by their embeddings to generate adversarial samples. To ensure that the generated adversarial samples remain syntactically and semantically plausible, the following constraints were applied to the replacement words: restrictions on modifying the part-of-speech tags of words, limitations on modifying stopwords, constraints on repetitive modifications to the text, enforcement of a minimum cosine similarity of 0.9 between the word embeddings of replacement words and the original words, and a maximum word modification rate of 10%.

4.6 *Evaluation Method*

4.6.1 *Target Model Evaluation*

The task for the target model in this study is classification, focusing on gender and age classification tasks.

For the gender classification task, a binary classification problem, the dataset shows a near-balanced distribution. The classes are distributed as follows, male (23,777 samples, 53.27%) and female (20,858 samples, 46.73%). Given this near-balance, the metric for evaluating performance on the gender classification task is prediction accuracy.

The age classification task presents a different challenge due to the unbalanced nature of the dataset. The age groups are divided into three categories with the following distributions, young adult: 25,932 samples (61.93%), teenager: 10,248 samples (24.47%), senior: 5,693 samples (13.60%). Given this imbalance, it is crucial to use metrics that can provide a more nuanced evaluation of model performance. Therefore, in addition to accuracy, the F1 score is also employed as an evaluation metric. The F1 score, which considers both precision and recall, provides a better measure of the model's performance across the different age groups, particularly addressing the issues arising from the unbalanced dataset.

4.6.2 *Obfuscation Evaluation*

To evaluate the effectiveness of obfuscation techniques, accuracy drop is the primary metric. Accuracy drop measures the reduction in accuracy of the target model when subjected to adversarial examples, indicating the effectiveness of the obfuscation method in disrupting the target model's performance.

Additional metrics include attack success rate, perturbation rate, and number of queries. Attack success rate is the percentage of adversarial examples that cause the target model to make incorrect predictions, assessing the efficacy of the adversarial attack. Perturbation rate measures the extent of changes made to the original text to create adversarial examples. It is calculated as the proportion of words in the text that have been altered. It helps in evaluating the subtlety of the obfuscation technique, ensuring that the text remains as close to the original as possible while still being effective in the attack. Number of queries counts the number of queries made to the target model during the generation of adversarial examples. A lower number of queries indicates a more efficient attack method.

Human evaluation is also considered to assess the semantics and readability of the text (Jin et al., 2020). Due to the noise present in Reddit texts, such as grammar and spelling errors, participants' judgments may be influenced. Therefore, this study will use an alternative approach to evaluate whether the changes made by the attacks on the input are natural (Emmery et al., 2021). Randomly selected texts before and after the attacks will be mixed and presented to participants. Participants will determine whether sentences have been altered and identify which words raise their suspicions.

By employing these metrics, the study aims to provide a thorough and detailed evaluation of both the classification performance of the target model and the effectiveness of the adversarial obfuscation techniques.

5 RESULTS

This section presents the results of all experiments. First, the performance of the baseline model and the BERT base uncased model in predicting gender and age after training on the dataset will be introduced. Next, the effectiveness of the obfuscation attacks is discussed. Finally, the results of the human evaluation of the quality of the texts after the attacks will be presented.

5.1 *Target Classifier*

For age prediction, the Naive Bayes model achieved an accuracy of 0.615. However, due to the imbalanced distribution of age groups, it is essential to consider the F1 score, which was 0.500 for this model. The Naive Bayes model struggled to accurately classify the less represented teenage and senior age groups. In contrast, BERT-base uncased model performed slightly better than the Naive Bayes model. BERT-base uncased model achieved an accuracy of 0.648, with F1 score as 0.616, indicating its better classification ability.

For gender, the Naive Bayes model achieved an accuracy of 0.694, while the BERT base uncased model achieved a higher accuracy of 0.76.

In summary, BERT-base uncased model demonstrated better classification capabilities than the baseline Naive Bayes model for both gender and age datasets, making it a more suitable choice for further adversarial attack evaluations.

Table 2: Classifier Performance Comparison

Model	**Task**	**Accuracy**	**F1 Score**
Naive Bayes	Age	0.615	0.500
BERT	Age	**0.648**	0.616
Naive Bayes	Gender	0.694	-
BERT	Gender	**0.760**	-

5.2 *Obfuscation*

5.2.1 *Obfuscation Result*

Due to limited computational resources, we randomly selected 10% of examples from the test set for the attack test. Out of the attempted attacks on age dataset, 97 were successful, 69 failed, and 50 were skipped. The

original accuracy of the model was 76.85%, which dropped to 31.94% under attack. The attack success rate was 58.43%, with an average of 8.9% of words perturbed per input. These results indicate that the attack was able to significantly reduce the model's accuracy, demonstrating the vulnerability of the model to adversarial attacks. For the attack on the gender classifier, 96 attacks were successful, 91 failed, and 47 were skipped. The original accuracy of the model was 79.91%, which decreased to 38.89% under attack. The attack success rate was 51.34%, and on average, 7.05% of words were perturbed per input.

Attribute	**Accuracy Drop**	Success Rate	% Perturbed Words	Number of Queries
Age	0.319	0.584	8.94%	87.04
Gender	0.389	0.514	7.05%	87.79

Table 3: Results of TextFooler attacks on age and gender prediction models.

Table 4 illustrates a successful attack example on age classification. After the attack, a part of the original text transforms from "whole" to "entire," "starting" to "initiates," and "oh" to "ah." The model classifies the original text as category 2 (adult), while the classification of the perturbed text is 1 (teenager). The true category of the original text is 2. The original score is 0.286, whereas the score of the perturbed text is 0.541. This indicates that the model's confidence in classifying the original text is relatively low, whereas its confidence significantly increases after the attack. It's worth noting that the modified sentences retain their original meaning, even though the grammar is not entirely correct.

Original	a **whole** new set for all the family, **starting** with my **oh**.
Perturbed	a **entire** new set for all the family, **initiates** with my **ah**.

Table 4: An example of original and perturbed texts

5.2.2 *Error Analysis of Obfuscation*

The error analysis is used to examine where and why the attacks failed, and identify the limitations and weaknesses of the adversarial attack methods used in the experiment.

> Original Text: "**Though** because I'm a first-time **mom** and it seems like those pregnancies tend to go beyond the due **date** by at **least** a week off topic, but I read this in Val Kilmer's voice as Doc Holiday from Tombstone and now I **want** to watch it again. 'It **appears** my hypocrisy knows no **bounds**.' Ahhh dd

> buddies!!! It's so **crazy** that it's so close now. I'm not sure which would be less **scary**, having the kid earlier or having to **wait** so long to **get** him. I think my **husband needs** more time to prepare himself lol."
>
> Perturbed Text: "**Albeit** because I'm a first-time **momma** and it seems like those pregnancies tend to go beyond the due **dating** by at **fewest** a week off topic, but I read this in Val Kilmer's voice as Doc Holiday from Tombstone and now I **wants** to watch it again. 'It **appear** my hypocrisy knows no **limitations**.' Ahhh dd buddies!!! It's so **madman** that it's so close now. I'm not sure which would be less **spooky**, having the kid earlier or having to await so long to **got** him. I thought my **hubby need** more time to prepare himself lol."

In the above example, the attack failed. The obfuscation successfully selected the "important words" influencing the classifier's prediction. However, changing "mom" to "momma" and "my husband" to "my hubby" did not change the perception that the writer is a woman. In this case, synonym substitution was ineffective; only antonyms should be used to mislead the prediction model into drawing the opposite conclusion.

5.3 *Human Evaluation*

In the human evaluation phase, the goal was to assess whether the text perturbed by the attack model appeared natural and whether it would arouse suspicion. We randomly selected 10 texts, comprising both original and post-attack versions, and presented them to four participants. The participants were not informed whether the texts were original or altered. Their task was to identify which texts had been perturbed among the 10 presented samples. The overall accuracy of participants in correctly identifying whether a text had been altered was 40%. Participants were only able to correctly identify 33.3% of the texts that had been perturbed. This suggests that the perturbed texts maintained a level of naturalness and coherence, making them less conspicuous.

There is an example where participants identified the text as perturbed. Participants noted that "aid" is not an appropriate verb in this context. Although "help" and "aid" have similar meanings, "aid" is too formal considering the whole context, making it seem suspicious.

> Original Text: "I am on a relapse now and tapering but that one, if you're into vodka, will get you through some uncomfortable situations. Best of luck friend, don't do like me and burn it all

> down. But this may actually **help** you, as insane as it is to even know that. Good job friend!"
>
> Perturbed Text: "I am on a relapse now and tapering but that one, if you're into vodka, will get you through some uncomfortable situations. Best of luck friend, don't do like me and burn it all down. But this may actually **aid** you, as insane as it is to even know that. Good job friend!"

In another example, shown below, all participants identified the text as perturbed due to multiple suspicious errors.

> Original Text: "In an interview with Miami's Channel 7, Sunrise **police** Chief **Anthony Rosa** said Pullease's behavior was 'disgusting,' **adding** that 'the video **speaks** for itself.' But also, **asked** why Pullease isn't facing criminal charges, Rosa **said**: 'so there's some **details** of the **investigation** that I've not **disclosed**, that I'm **unable** to **disclose** right now ..."
>
> Perturbed Text: "In an interview with Miami's Channel 7, Sunrise **policemen** Chief **Antoni Rossa** said Pullease's behavior was 'disgusting,' **adds** that 'the video **talk** for itself.' But also, **enquired** why Pullease isn't facing criminal charges, Rosa **say**: 'so there's some **detail** of the **probe** that I've not **divulged**, that I'm **incapable** to **divulge** right now ..."

Participants pointed out errors such as inconsistencies in names (e.g., "Rossa" vs. "Rosa"), mismatched tenses (e.g., "said" vs. "say"), and inappropriate word choices (e.g., "disclosed" vs. "divulge"). These discrepancies indicate that higher perturbation rates make the text more likely to raise suspicion.

Additionally, the human evaluation revealed a 20% error rate where original sentences were mistakenly identified as perturbed. This highlights the noisy nature of Reddit data, which complicates the evaluation process and underscores the challenge of creating perturbations that blend seamlessly with the original text.

6 DISCUSSION

6.1 *Answering Research Questions*

The first sub-research question addresses the accuracy of the initial author profiling model in predicting the gender and age of the author. Our

findings indicate that for age prediction, the Naive Bayes model achieved an accuracy of 0.615, while the BERT-base uncased model achieved a slightly higher accuracy of 0.648. For gender prediction, the Naive Bayes model reached an accuracy of 0.694, whereas the BERT-base uncased model achieved a higher accuracy of 0.76. These results underscore the ability of author profiling models to extract significant information about authors from textual data on social media platforms. Integrating additional sources, such as image analysis (including profile pictures) (Liu et al., 2016) and timestamps (Rocha et al., 2016) from social media, could potentially enhance prediction accuracy. These insights highlight significant privacy risks associated with such practices. The BERT-base uncased model demonstrated superior classification capabilities compared to the baseline Naive Bayes model for both gender and age datasets, showcasing its potential in the field of author profiling. This is consistent with other research findings, where gender prediction accuracy typically exceeds that of age prediction. Studies using BERT models for gender prediction report accuracies ranging between 70% and 90%, while age prediction accuracies often vary from 30% to 70%(Abdul-Mageed et al., 2019; Bsir et al., 2024). Clearly, the results are influenced by the dataset used, and other experiments suggest various methods to further optimize the BERT model. Future research might also investigate the integration of additional contextual information from text, such as linguistic features and user behaviour patterns, to enhance model accuracy.

The second sub-research question investigates how the accuracy of the initial author profiling model changes after applying TextFooler. The original accuracy of the age prediction model was 76.85%, which dropped by 31.94% under attack. The original accuracy of the gender classifier was 79.91%, which decreased by 38.89% under attack. These results demonstrate the effectiveness of TextFooler in perturbing the predictions of the author profiling models. The substantial decrease in accuracy highlights the vulnerability of these models to adversarial attacks, raising important concerns about their robustness and reliability in real-world applications. Compared to other studies, our experiment showed a lower accuracy drop due to the additional constraints we implemented to ensure that the generated text retained maximum similarity to the original, reducing the likelihood of detection. Besides accuracy, the attack success rate, perturbation rate, and number of queries yielded similar results for both tasks, indicating that the TextFooler model maintains efficiency under consistent constraints. Further investigation into different types of adversarial attacks and defences could provide a more comprehensive understanding of the model's resilience and potential weaknesses.

The third sub-research question examines the extent to which TextFooler preserves the original semantics and readability of the text. For this evaluation, we used human judgment. The overall accuracy of participants in correctly identifying whether a text had been altered was 40%. Participants correctly identified only 33.3% of the perturbed texts, indicating that 66.7% of the perturbed texts were not detected as altered. This suggests that the texts retained a significant degree of the original semantics and readability. It was observed that higher perturbation rates made the text more suspicious to participants. TextFooler demonstrated effectiveness in preserving original semantics, but there is room for improvement in readability. Discrepancies in tense and formality, especially in longer texts, can render the perturbed text less natural and more detectable. Future enhancements to TextFooler could focus on maintaining contextual coherence and improving the naturalness of the perturbed texts to make them less detectable while preserving the intended obfuscation.

Returning to the main research question of how effectively TextFooler protects the gender and age information of authors in our Reddit dataset, our analysis of the first two sub-questions shows that TextFooler demonstrates robust performance even against strong author profiling models trained on extensive datasets. The predictive accuracy of these models significantly declines post-attack, consistent with findings from prior studies(Mozes et al., 2021; Neshaei et al., 2024). This indicates that TextFooler is an effective obfuscation tool, capable of significantly disrupting the identification of personal privacy on social media. However, TextFooler is not perfect.

In the discussion of the third sub-question, we noted that TextFooler's strong ability to selectively replace words while maintaining the original meaning of the text is commendable, but preserving readability remains a challenge. If replacements focus only on individual words without considering their context within the sentence, maintaining sentence coherence and conciseness becomes difficult. This can lead to decreased readability, thereby reducing the effectiveness of the model. We also identified some weaknesses in the TextFooler model. Synonym substitution of keywords is not always effective, particularly when the keywords indicate the author's identity or are topic-related. Using antonyms instead of synonyms could be more effective in misleading the model.

6.2 *Impact*

This study contributes to the field of author profiling by demonstrating the vulnerabilities of current models to adversarial attacks. The findings reveal that even sophisticated models like BERT, which perform well in

predicting gender and age, can have their accuracy substantially reduced through targeted obfuscation methods like TextFooler. This highlights critical privacy concerns, as it shows that personal information extracted from social media can be effectively protected against profiling attempts. Moreover, the study underscores the need for improved adversarial robustness in author profiling models and provides insights into balancing semantic preservation and readability in text perturbation. These contributions are crucial for developing more secure and privacy-respecting AI applications in social media analysis. The implications of these findings extend to various domains including cybersecurity, data privacy regulations and ethical AI deployment where protecting user identity and sensitive information is paramount. By demonstrating the potential of adversarial attacks to safeguard privacy, this research paves the way for more robust privacy-preserving techniques in the ever-evolving landscape of digital communication.

6.3 *Limitations and Future Work*

This study acknowledges several limitations. First, the author attributes in the database were extracted from self-reports, which may not accurately reflect the true attributes of the authors, potentially introducing noise. The preprocessing steps were limited, and there was no cleaning of spam users or posts, nor any filtering of other languages, which might have introduced additional noise. Due to computational and memory constraints, only partial text fragments were used. Specifically, only the first 500 characters were selected without stratified sampling, which could affect the accuracy of the experiment. Additionally, the external validity of the data from Reddit still needs to be tested to ensure the results are generalizable. Future work should explore cross-domain and cross-type data to further optimize the obfuscation model. Furthermore, achieving attack success with fewer perturbations remains a critical challenge. Future research should focus on preserving the semantics of the original text while achieving effective obfuscation and ensuring the generated text remains unrecognizable. This balance is essential for maintaining the readability and naturalness of perturbed texts, which is vital for practical applications. Enhancements in computational techniques, including the use of more advanced natural language processing models and increased computational resources, could provide more accurate and generalizable results. Additionally, exploring other social media platforms and types of textual data could offer broader insights and applications of the findings.

7 CONCLUSION

The objective of this study was to evaluate how effectively TextFooler protects the gender and age information of authors in our Reddit dataset. To address this, we employed the BERT base model as the target model and utilized the TextAttack framework with TextFooler as the attack method. Naïve Bayes was used as a baseline model to demonstrate the effectiveness of the BERT base model in predicting the gender and age of authors. After the TextFooler attack, the accuracy of the target model decreased by 0.319 and 0.389 for gender and age prediction, respectively. Human evaluation revealed that participants were only able to correctly identify 33.3% of the texts that had been perturbed. This demonstrates the effectiveness and subtlety of TextFooler as an adversarial writing tool. However, TextFooler also has limitations. Simply replacing individual words without considering the specific context can generate text that raises suspicion or leads to ineffective attacks. In summary, this study shows that author profiling models can accurately predict the gender and age of authors based on text from social media, highlighting significant privacy risks. Although adversarial writing tools like TextFooler can effectively reduce the accuracy of prediction models, their use still carries the risk of detection. Future research should focus on improving these tools to preserve the semantics and readability of the text while enhancing their ability to obfuscate author information. On top of that, exploring other adversarial techniques and combining them with TextFooler could potentially yield more robust solutions. Expanding the scope of datasets and incorporating diverse linguistic features might also help in developing more comprehensive and versatile privacy-preserving models. Ultimately, this research contributes to the ongoing discourse on privacy and security in the digital age, emphasizing the need for continuous innovation and ethical considerations in AI development.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 265–283.

Abdul-Mageed, M., Zhang, C., Rajendran, A., Elmadany, A., Przystupa, M., & Ungar, L. (2019). Sentence-level bert and multi-task learning of age and gender in social media. *arXiv preprint arXiv:1911.00637*.

Aborisade, O., & Anwar, M. (2018). Classification for authorship of tweets by comparing logistic regression and naive bayes classifiers. *2018*

IEEE International Conference on Information Reuse and Integration (IRI), 269–276.

Adebayo, G. O., & Yampolskiy, R. V. (2022). Estimating intelligence quotient using stylometry and machine learning techniques: A review. *Big Data Mining and Analytics*, *5*(3), 163–191.

Ali, A., & Taqa, A. Y. (2022). Analytical study of traditional and intelligent textual plagiarism detection approaches. *Journal of Education and Science*, *31*(1), 8–25.

Alzahrani, E., & Jololian, L. (2021). How different text-preprocessing techniques using the bert model affect the gender profiling of authors. *arXiv preprint arXiv:2109.13890*.

Argamon, S. (2018). Computational forensic authorship analysis: Promises and pitfalls. *Language and Law/Linguagem e Direito*, *5*(2), 7–37.

Balakrishnan, R., Sloan, S., & Aswani, A. (2021). Protecting anonymous speech: A generative adversarial network methodology for removing stylistic indicators in text. *arXiv preprint arXiv:2110.09495*.

Banerjee, S., & Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Barlas, G., & Stamatatos, E. (2021). A transfer learning approach to cross-domain authorship attribution. *Evolving Systems*, *12*(3), 625–643.

Brennan, M., Afroz, S., & Greenstadt, R. (2012). Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security (TISSEC)*, *15*(3), 1–22.

Bsir, B., Khoufi, N., & Zrigui, M. (2024). Prediction of author's profile basing on fine-tuning bert model. *Informatica*, *48*(1).

Casimiro, G. R., & Digiampietri, L. A. (2022). Authorship attribution with temporal data in reddit. *Proceedings of the XVIII Brazilian Symposium on Information Systems*, 1–8.

Chen, Y., & Skiena, S. (2014). Building sentiment lexicons for all major languages. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 383–389.

Crothers, E., Japkowicz, N., Viktor, H., & Branco, P. (2022). Adversarial robustness of neural-statistical features in detection of generative transformers. *2022 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ehrhardt, S., & Visconti, J. (2018). Authorship attribution analysis. *Handbook of communication in the legal sphere*, 169–200.

Emmery, C., Kádár, Á., & Chrupała, G. (2021). Adversarial stylometry in the wild: Transferable lexical substitution attacks on author profiling. *arXiv preprint arXiv:2101.11310*.

Ferrante, E. (2017). *L'amica geniale. edizione completa*. E/O Edizioni.

Gagiano, R., Kim, M. M.-H., Zhang, X. J., & Biggs, J. (2021). Robustness analysis of grover for machine-generated news detection. *Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association*, 119–127.

Gröndahl, T., & Asokan, N. (2019). Text analysis in adversarial settings: Does deception leave a stylistic trace? *ACM Computing Surveys (CSUR)*, *52*(3), 1–36.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., et al. (2020). Array programming with numpy. *Nature*, *585*(7825), 357–362.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95.

Jin, D., Jin, Z., Zhou, J. T., & Szolovits, P. (2020). Is bert really robust? a strong baseline for natural language attack on text classification and entailment. *Proceedings of the AAAI conference on artificial intelligence*, *34*(05), 8018–8025.

Jockers, M. L., & Witten, D. M. (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, *25*(2), 215–223.

Kwon, H. (2021). Dual-targeted textfooler attack on text classification systems. *IEEE Access*, *11*, 15164–15173.

Lepekhin, M., & Sharoff, S. (2021). Experiments with adversarial attacks on text genres. *arXiv preprint arXiv:2107.02246*.

Li, L., Song, D., & Qiu, X. (2022). Text adversarial purification as defense against adversarial attacks. *arXiv preprint arXiv:2203.14207*.

Liu, L., Preotiuc-Pietro, D., Samani, Z. R., Moghaddam, M. E., & Ungar, L. (2016). Analyzing personality through social media profile picture choice. *Proceedings of the International AAAI Conference on Web and Social Media*, *10*(1), 211–220.

Macko, D., Moro, R., Uchendu, A., Srba, I., Lucas, J. S., Yamashita, M., Tripto, N. I., Lee, D., Simko, J., & Bielikova, M. (2024). Authorship obfuscation in multilingual machine-generated text detection. *arXiv preprint arXiv:2401.07867*.

Mahmood, A., Ahmad, F., Shafiq, Z., Srinivasan, P., & Zaffar, F. (2019). A girl has no name: Automated authorship obfuscation using mutant-x. *Proceedings on Privacy Enhancing Technologies*.

McKinney, W. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, *445*(1), 51–56.

Mein, G., Hartman, K., & Morris, A. (2020). Firebert: Hardening bert-based classifiers against adversarial attack.

Miller, Z., Dickinson, B., Deitrick, W., Hu, W., & Wang, A. H. (2014). Twitter spammer detection using data stream clustering. *Information Sciences*, *260*, 64–73.

Modupe, A., Celik, T., Marivate, V., & Olugbara, O. O. (2022). Post-authorship attribution using regularized deep neural network. *Applied Sciences*, *12*(15), 7518.

Morris, J. X., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., & Qi, Y. (2020). Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. *arXiv preprint arXiv:2005.05909*.

Mosteller, F., & Wallace, D. L. (1963). Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, *58*(302), 275–309.

Mozes, M., Bartolo, M., Stenetorp, P., Kleinberg, B., & Griffin, L. D. (2021). Contrasting human-and machine-generated word-level adversarial examples for text classification. *arXiv preprint arXiv:2109.04385*.

Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., & Woodard, D. (2017). Surveying stylometry techniques and applications. *ACM Computing Surveys (CSuR)*, *50*(6), 1–36.

Neshaei, S. P., Boreshban, Y., Ghassem-Sani, G., & Mirroshandel, S. A. (2024). The impact of quantization on the robustness of transformer-based text classifiers. *arXiv preprint arXiv:2403.05365*.

Onikoyi, B., Nnamoko, N., & Korkontzelos, I. (2023). Gender prediction with descriptive textual data using a machine learning approach. *Natural Language Processing Journal*, *4*, 100018.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*, 8024–8035.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, *12*(Oct), 2825–2830.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, *54*(1), 547–577.

Potthast, M., Hagen, M., & Stein, B. (2016). Author obfuscation: Attacking the state of the art in authorship verification. *CLEF (Working Notes)*, 716–749.

Potthast, M., Schremmer, F., Hagen, M., & Stein, B. (2018). Overview of the author obfuscation task at pan 2018: A new approach to measuring safety. *CLEF (Working Notes)*.

Rangel, F., Celli, F., Rosso, P., Martin, P., Stein, B., Daelemans, W., et al. (2015). Overview of the 3rd author profiling task at pan 2015. *CLEF2015 Working Notes. Working Notes of CLEF 2015-Conference and Labs of the Evaluation forum.*

Rangel, F., & Rosso, P. (2019). Overview of the 7th author profiling task at pan 2019: Bots and gender profiling in twitter. *Working notes papers of the CLEF 2019 evaluation labs*, *2380*, 1–7.

Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., & Inches, G. (2013). Overview of the author profiling task at pan 2013. *CLEF conference on multilingual and multimodal information access evaluation*, 352–365.

Rangel, F., Rosso, P., Montes-y-Gómez, M., Potthast, M., & Stein, B. (2018). Overview of the 6th author profiling task at pan 2018: Multimodal gender identification in twitter. *Working notes papers of the CLEF*, *192*.

Reddy, S., & Knight, K. (2016). Obfuscating gender in social media writing. *Proceedings of the First Workshop on NLP and Computational Social Science*, 17–26.

Ricciardi, A. (2021). *Finding ferrante: Authorship and the politics of world literature*. Columbia University Press.

Rocha, A., Scheirer, W. J., Forstall, C. W., Cavalcante, T., Theophilo, A., Shen, B., Carvalho, A. R., & Stamatatos, E. (2016). Authorship attribution for social media forensics. *IEEE transactions on information forensics and security*, *12*(1), 5–33.

Savoy, J. (2020). Machine learning methods for stylometry. *Cham: Springer*.

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006). Effects of age and gender on blogging. *AAAI spring symposium: Computational approaches to analyzing weblogs*, *6*, 199–205.

Tweedie, F. J., Singh, S., & Holmes, D. I. (1996). Neural network applications in stylometry: The federalist papers. *Computers and the Humanities*, *30*, 1–10.

Uchendu, A., Le, T., & Lee, D. (2023). Attribution and obfuscation of neural text authorship: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, *25*(1), 1–18.

Wang, H., Wang, S., Xu, D., Zhang, X., & Liu, X. (2020). Generating effective software obfuscation sequences with reinforcement learning. *IEEE Transactions on Dependable and Secure Computing*, *19*(3), 1900–1917.

Wang, Z., Le, T., & Lee, D. (2022). Upton: Preventing authorship leakage from public text release via data poisoning. *arXiv preprint arXiv:2211.09717*.

Weinsberg, U., Bhagat, S., Ioannidis, S., & Taft, N. (2012). Blurme: Inferring and obfuscating user gender based on ratings. *Proceedings of the sixth ACM conference on Recommender systems*, 195–202.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Davison, J. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.

Wolff, M., & Wolff, S. (2020). Attacking neural text detectors. *arXiv preprint arXiv:2002.11768*.

Xing, E., Venkatraman, S., Le, T., & Lee, D. (2024). Alison: Fast and effective stylometric authorship obfuscation. *arXiv preprint arXiv:2402.00835*.

Yang, F., Purwanto, E., & Man, K. L. (2022). Emotionfooler: An effective and precise textual adversarial attack method with part of speech and similarity score checking. *Professor Ka Lok Man, Xi'an Jiaotong-Liverpool University, China Professor Young B. Park, Dankook University, Korea Chairs of CICET 2022*, 22.

Zhang, W. E., Sheng, Q. Z., Alhazmi, A., & Li, C. (2020). Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *11*(3), 1–41.

Zhang, Y. C. (2020). "cyber bullying stays with you": A textual analysis of a subreddit involving cyberbullying victims. *A Closer Look in Unusual Times*, 81.

Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American society for information science and technology*, *57*(3), 378–393.