BACHELOR'S THESIS / ENGINEERING 2024

Generating Adversarial Examples for Misinformation Detection Using Automated Text Simplification

Martina Gómez Martín

Universitat Pompeu Fabra Barcelona

Generating Adversarial Examples for Misinformation Detection Using Automated Text Simplification

TREBALL DE FI DE GRAU DE
Martina Gómez Martín

Piotr Przybyla

Grau en Enginyeria Matemàtica en Ciència de Dades

Curs 2023 - 2024

Universitat Pompeu Fabra Barcelona
Escola d'Enginyeria

Abstract

Text classification techniques have been thoroughly employed to identify unreliable content, such as fake news and propaganda. These models, usually based on Deep Neural Networks, are crucial for managing online platforms with user-created content. In reaction to these actions, adversarial examples have been created to bypass these algorithms and exploit their vulnerabilities.

Various methods have emerged to create adversarial examples, testing the robustness of widely used text classification algorithms. These examples reveal weaknesses, showing that minor text perturbations, without altering meaning, can lead to misleading classifications. This paper focuses on using Text Simplification, an evolving NLP method for its implications in enhancing accessibility and comprehensibility. Given its nature of modifying text, this technique will be used to generate adversarial examples for evaluating the robustness of text classification algorithms.

Specifically, we employ MUSS (an unsupervised automated sentence simplification technique) to create adversarial attacks and use the BODEGA framework to evaluate the attack model across four misinformation detection tasks. Results show that, despite a few successful cases, it manages to create effective examples with few queries.

Resum

Les tècniques de classificació de text s'han emprat exhaustivament per identificar contingut poc fiable, com notícies falses i propaganda. Aquests models, generalment basats en Xarxes Neuronals Profundes, són crucials per gestionar plataformes en línia amb contingut creat pels usuaris. En resposta, s'han creat exemples adversarials per eludir aquests algorismes i explotar-ne les vulnerabilitats.

Han sorgit diversos mètodes per crear exemples adversarials, provant la robustesa dels algorismes de classificació de text més utilitzats. Aquests exemples revelen debilitats, mostrant que petites pertorbacions en el text, sense alterar-ne el significat, poden portar a classificacions enganyoses. Aquest article se centra en l'ús de la Simplificació de Text, un mètode en evolució en el camp del PLN per les seves implicacions en la millora de l'accessibilitat i la comprensibilitat. Dada la seva naturalesa de modificar el text, aquesta tècnica s'utilitzarà per generar exemples adversarials per avaluar la robustesa dels algorismes de classificació de text.

Específicament, fem servir MUSS (una tècnica de simplificació de frases automatitzada no supervisada) per crear atacs adversarials i utilitzem el marc BODEGA per avaluar el model d'atac en quatre tasques de detecció de desinformació. Els resultats mostren que, encara que els casos exitosos no són nombrosos, s'aconsegueixen exemples efectius amb poques consultes.

Resumen

Las técnicas de clasificación de texto se han empleado exhaustivamente para identificar contenido poco fiable, como noticias falsas y propaganda. Estos modelos, generalmente basados en Redes Neuronales Profundas, son cruciales para gestionar plataformas en línea con contenido creado por los usuarios. En respuesta a estas acciones, se han creado ejemplos adversariales para eludir estos algoritmos y explotar sus vulnerabilidades.

Han surgido varios métodos para crear ejemplos adversariales, probando la robustez de los algoritmos de clasificación de texto más utilizados. Estos ejemplos revelan debilidades, mostrando que pequeñas perturbaciones en el texto, sin alterar su significado, pueden llevar a clasificaciones engañosas. Este artículo se centra en el uso de la Simplificación de Texto, un método en evolución en el campo del PLN por sus implicaciones en la mejora de la accesibilidad y la comprensibilidad. Dada su naturaleza de modificar el texto, esta técnica se utilizará para generar ejemplos adversariales para evaluar la robustez de los algoritmos de clasificación de texto.

Específicamente, empleamos MUSS (una técnica de simplificación de oraciones automatizada no supervisada) para crear ataques adversariales y utilizamos el marco BODEGA para evaluar el modelo de ataque en cuatro tareas de detección de desinformación. Los resultados muestran que, aunque los casos exitosos no son numerosos, se logran ejemplos efectivos con pocas consultas.

Contents

List of Tables

Chapter 1

INTRODUCTION

The exponential increase in data volume in recent years presents a challenge in terms of manual management. This massive influx of data has highlighted the pressing need for Artificial Intelligence (AI) solutions that can effectively handle such vast amounts of information. This has boosted the popularity of Natural Language Processing (NLP) models [Dwivedi and Dwivedi, 2022].

One of the crucial use cases of NLP models is to regulate user-generated content on various platforms to avoid the dissemination of misinformation. To do so we use text classification algorithms which are used as a filter to classify between credible and manipulative text [Ahmed et al., 2022].

However, the limited understanding of Machine Learning (ML) models and their behavior, especially those based on Deep Neural Networks (DNNs), exposes them to vulnerabilities. Researchers continuously seek ways to make algorithms more robust and enhance their ability to discern between genuine and deceptive content [Flores and Hao, 2022].

One of the main methods employed to expose and rectify algorithmic vulnerabilities is through the creation of adversarial examples (AEs) [Oseni et al., 2021]—crafted inputs designed to mislead or deceive AI systems. The efforts of finding AEs are relatively new for NLP.

This paper aims to contribute to this evolving field by focusing on the application of Automatic Text Simplification (ATS) for generating AEs. The research includes the following:

1. Conducting a comprehensive study of state-of-the-art simplification techniques to understand their potential in generating AEs.

2. Selecting a suitable simplification technique and applying simplification to four distinct misinformation tasks.

3. Evaluating model performance to assess its effectiveness in simplifying text

while preserving meaning.

4. Utilizing the chosen simplification technique to craft adversarial attacks against a target model across the same four tasks, aiming to exploit vulnerabilities and test victim model robustness.

5. Analyzing the successful outcomes of the adversarial attack to assess the efficacy in undermining model robustness and reliability and proposing strategies to mitigate adversarial vulnerabilities.

By undertaking these steps, we aim to shed light on the potential of simplification as a tool for generating adversarial attacks and ultimately enhancing the resilience of AI algorithms in real-world applications.

1.1 Motivation

The motivation behind this research arises from the critical need to improve and boost the usability of NLP models. One of the primary goals is to reduce repetitive human repetitive tasks and automate them efficiently by applying trustworthy algorithms. The automation of routine tasks not only enhances productivity but also allows human resources to focus on more complex and value-added activities, leading to overall operational efficiency and effectiveness.

Another compelling motivation is the critical need to reduce the spread of misinformation. In today's digital age, the rapid dissemination of fake news and misleading information poses significant challenges to societal discourse and decision-making processes [Akers et al., 2018, Tucker et al., 2018]. By developing robust algorithms and techniques that can accurately discern between credible and deceptive content, we can contribute to fostering a more informed and trustworthy society [Ahmed et al., 2022].

Additionally, this research seeks to address the ongoing challenge of AEs in the NLP field. Adversarial attacks pose a substantial threat to the reliability and integrity of AI systems, potentially leading to erroneous decisions and outcomes [Carter et al., 2021, Zhang et al., 2020]. By exploring new techniques and methodologies to combat AEs, we aim to enhance the resilience and trustworthiness of NLP models, ultimately advancing the state-of-the-art in AI security and reliability.

The significance of this research lies in its potential to drive advancements in AI technology that are not only impactful, but also address real-world challenges. By tackling these pressing issues, we can contribute to the development of more robust, reliable, and trustworthy AI systems, with far-reaching implications across various domains and industries.

Chapter 2

RALATED WORK

2.1 Adversarial Examples

Adversarial examples are altered inputs designed to deceive models into making incorrect predictions. They are directly related to the robustness and reliability of AI applications across a wide range of industries and domains.

- **Adversarial Examples**: inputs to machine learning models that are intentionally designed to cause the model to make a mistake [Zhang et al., 2020]. Often, these are crafted by introducing specific types of "noise" or subtle manipulations that exploit vulnerabilities in the model. The concept extends to various types of data, including images, text, and other media that can be digitally processed.

2.1.1 Mechanisms and Techniques

Adversarial attacks emerged in the computer vision [Zhang et al., 2023] field. Within this domain, adversarial attacks focus on inducing misclassification while minimizing perceptual distortion. Techniques such as the Fast Gradient Sign Method (FGSM), Jacobian-based Saliency Map Attack (JSMA), and others that leverage gradients are the most commonly used to find the most effective alterations [Zhang et al., 2020]. For instance, FGSM uses gradients to determine the direction to modify pixel values to maximize prediction error, maintaining the perturbation's subtlety.

When it comes to textual data, utilizing these computer vision attacks presents various challenges as textual data is symbolic and discrete in nature, which poses difficulties in defining perturbations. Moreover, when it comes to images, humans typically struggle to detect small changes in pixel composition; however, altering the semantics of a sentence is more complicated as there is no equivalent of imperceptible noise in text [Liang et al., 2017]. For example, removing a negation

word can significantly impact its sentiment [Zhang et al., 2020]. There are various strategies that have been explored to address these challenges. One approach involves making changes at different levels of the text:

- At the word level: For example, substituting words while maintaining meaning by utilizing thesauri [Ren et al., 2019] or language models [Li et al., 2020].

- At the character level: Introducing modifications that are unlikely to be noticed by a reader (e.g., inserting typographical errors) [Boucher et al., 2021].

- At the sentence level: Utilizing paraphrasing techniques [Barzilay and McKeown, 2001].

Additionally, alternative solutions have been suggested, such as inserting semantically empty phrases [Zhang et al., 2020] identifying the optimal position in the embedding space and selecting its closest real-world neighbor [Gong et al., 2018], or generating text samples from a distribution defined by continuous parameters [Guo et al., 2021].

2.2 Automated Simplification

2.2.1 Background

In an era characterized by a constant increase in the volume of information, Automatic Text Simplification (ATS) has emerged as a crucial area of research within NLP. ATS tools automatically transform complex texts to make them easier to understand while retaining the essential meaning and content integrity [Al-Thanyyan and Azmi, 2021].

The importance of ATS extends beyond mere convenience; a significant portion of the population faces barriers to accessing and understanding complex texts because of factors such as language proficiency, cognitive impairments, or educational background [Tajner, 2021]. Therefore, Text Simplification is important to promote inclusivity and equity by ensuring that individuals with diverse linguistic backgrounds, cognitive abilities, or educational levels can participate more fully in social, economic, and educational activities. Moreover, ATS has the potential to enhance communication between professionals and lay audiences, improving the dissemination of knowledge across various domains, and have implications for research in other fields.

2.2.2 Challenges

The main obstacle in the field of Text Simplification lies in the task of preserving the original meaning and coherence of the text while also making it more comprehensible [Zhang and Lapata, 2017]. ATS systems face the challenge of finding the right equilibrium between simplifying complex language and retaining crucial information, all while maintaining syntactic fluency and grammatical accuracy.

2.2.3 Approaches

The rewrite approaches that give rise to simplified text include the identification and substitution of complex words, recognizing and addressing complex syntactic structures such as coordination or subordination, and the deletion of elements of the original text [Al-Thanyyan and Azmi, 2021].

Earlier work on Sentence Simplification (SS) focused on these individual aspects of the simplification problem. For example, some models based on language rules performed syntactical simplification by trying to understand the context [Bott et al., 2012] of complex words [Paetzold and Specia, 2016] and choosing the best word substitution. Other models, such as YATS [Ferrés et al., 2016] combine different aspects of simplification. This tool uses a context-vector model to find suitable substitutions by ranking them by word frequency and also applies two stages of syntactic simplification to simplify structures such as appositive phrases, coordinated correlatives, and relative clauses.

More recent approaches have treated the SS task as monolingual machine translation [Wubben et al., 2012, Zhu et al., 2010]. This approach requires a large parallel corpus of aligned complex-simple sentence pairs [Saggion, 2017]. WikiLarge [Zhang and Lapata, 2017, Naskar et al., 2019, Alva-Manchego et al., 2020] and Newsla [Zhang and Lapata, 2017, Xu et al., 2015, Jiang et al., 2020] have been widely used. Those methods often employ sequence-to-sequence as the backbone, then integrate different submodules into it, such as reinforcement learning [Zhang and Lapata, 2017], external simplification rules databases [Zhao et al., 2018], adding a new loss [Nishihara et al., 2019], or lexical complexity features [Martin et al., 2021, Feng et al., 2023].

Since supervised datasets are limited, other methods have been proposed to generate unsupervised datasets, which often consist of mined paraphrases. One such method is backtranslation, where a sentence is translated into another language and then translated back into the original language [Lu et al., 2021]. Another approach involves using heuristics such as embedding similarity to identify semantically similar sentence pairs [Martin et al., 2021].

To effectively train models on unsupervised parallel data, the use of control tokens has proven to be beneficial. These tokens allow models to focus on specific

features that are correlated with sentence simplicity. For instance, the ACCESS method adds tokens at the beginning of each input sentence, specifying the desired output length, similarity between the output and input, output word rank, and output tree depth [Martin et al., 2021]. Since these tokens are typically added in plain text before tokenization, they essentially serve as a form of prompt learning [Chi et al., 2023].

The latest efforts toward SS have focused on Large Language Models (LLMs) [Feng et al., 2023, Kew et al., 2023]. LLMs have a much larger scale in terms of model parameters and training data and can be prompted zero-shot or few-shot to solve a task. Experiments performed by [Feng et al., 2023] using GPT3.5 and ChatGPT showed that performance in SS tasks outperformed current state-of-the-art methods in the domain of multilingual SS tasks [Chi et al., 2023].

2.3 Adversarial examples with Automated Simplification Techniques

The process of generating AEs involves modifying the text through techniques such as word substitution or sentence paraphrasing, with the aim of creating AEs [Zhang et al., 2020]. In this paper, we propose a novel approach to the generation of AE by utilizing ATS techniques. Our approach involves using a simplifier function to modify input texts and then examining how these modified texts are labeled by a classifier model. Through this study, we aim to investigate the robustness of simplification as a means of generating AEs. Specifically, we will focus on the domain of misinformation detection by labeling texts based on their veracity.

This research presents a novel approach and marks the initial steps in evaluating the effectiveness of ATS as a method to generate successful AEs. The absence of prior exploration in this specific area underscores the need for further investigation into the potential of ATS techniques for enhancing the resilience of NLP models against adversarial attacks.

Chapter 3

METHODS

The research methodology comprises four primary stages: choosing an appropriate simplification technique, implementing the chosen technique across various tasks to assess the outcomes of the simplification process, attacking victim models using the selected method, and evaluating the achieved results.

3.1 Selecting a simplification tool

To conduct our research effectively, it is essential to choose a tool that possesses specific attributes. Firstly, the tool must have the ability to support the English language. Secondly, it should be open-source, providing the advantage of customization and adaptability to meet our requirements. Furthermore, integration with Python is a vital aspect, as the attack evaluation is performed with the BODEGA framework. Lastly, we take into consideration the potential conflicts that may arise with existing libraries to ensure compatibility.

Given the prior research in SS and the constraints related to resources found for utilizing LLMs for our experiments, we opted to employ MUSS (Multilingual Unsupervised Learning by Mining Paraphrases). MUSS achieves strong performance, is available as an open-source Python implementation, and supports simplification in English [Martin et al., 2021].

3.1.1 Multilingual Unsupervised Learning by Mining Paraphrases (MUSS)

MUSS addresses the availability limitations of parallel simplified corpora by training controllable models using sentence-level paraphrase data—parallel sentences with the same meaning but different phrasing. Paraphrased data is more readily

available, enabling the training of flexible models adaptable to various simplification scenarios. Mining paraphrases have been shown to lead to better simplification performance compared to directly mining simplifications, as they are more straightforward and require fewer prior assumptions [Martin et al., 2021].

The MUSS method is classified as unsupervised since it does not rely on labeled simplification data for its training. This model employs control tokens to facilitate the focus on particular features, including output length, similarity between the output and input, output word rank, and output tree depth.

In order to utilize MUSS effectively, it was necessary to preprocess texts by segmenting them into sentences. To accomplish this, we utilized LAMBO (Layered Approach to Multi-Level Boundary Identification) [Przybyla, 2022], a segmentation tool that has been trained to identify the boundaries of tokens and sentences in real-world text. LAMBO, which is implemented as a PyTorch deep neural network, successfully divides the text into sentences before simplification.

3.2 Description Misinformation Tasks

The four tasks for detecting misinformation in our experiments are outlined in [Przybyła et al., 2023].

- Style-based news bias assessment (HN): Hyperpartisan News Detection task aims to detect fake news based on writing style. The corpus [Kiesel et al., 2019] defines as non-credible articles those from sources annotated as hyperpartisan, both right- and left-wing.

- Propaganda detection (PR): The objective of the propaganda detection task is to identify written content in which the writer attempts to influence the reader's opinion. While this does not always indicate the presence of false information, within the field of journalism, it is often linked to deceptive practices. The corpus used has 14 propaganda techniques annotated in 371 newspaper articles by professional annotators [Da San Martino et al., 2020].

- Fact checking (FC): Fact-checking tasks center on the verification of claims within the fact-checking process. The task requires a pair of texts as input: the claim being examined and the corresponding evidence. The output label determines if the evidence corroborates or contradicts the claim [Thorne et al., 2018].

- Rumor detection (RD): Rumor is information spreading between people despite not having a reliable source. The input of this task is a dataset of rumors and non-rumors [Han et al., 2019], created from Twitter threads

relevant to six real-world events (2013 Boston marathon bombings, 2014 Ottawa shooting, 2014 Sydney siege, 2015 Charlie Hebdo Attack, 2014 Ferguson unrest, 2015 Germanwings plane crash) and follow-ups from other social media users.

3.3 Simplification Evaluation

In order to assess the efficacy of MUSS models in addressing misinformation, a thorough manual analysis of different examples is undertaken. Fifty simplified instances of each misinformation task are chosen and simplified using the MUSS method for this purpose. The instances are then classified according to the extent of their deviation from the original text while achieving the desired task.

The metrics used for evaluation are the following:

- Simplification (S): The degree of simplification in relation to the original text. In order to standardize the ratings, if the sentence was already simple and little to no modifications were made, the score will be low as the model has not undergone any simplification.
- Fluency (F): The fluency of a text refers to its ability to maintain a correct grammatical and syntactic structure, resulting in a natural-sounding piece of writing.
- Meaning Preservation (MP): refers to the level of retention of the content within a text.

The simplification criteria are established based on these three parameters, which are rated on a scale from one to three. The highest rating, three, indicates the best performance in each category. It is crucial to highlight that the ratings assigned to these parameters are independent of each other. Therefore, a sentence can possess a suitable structure but still lack coherence.

In order to conduct a manual evaluation, we will employ a color-coded system. The color green will be utilized to highlight commendable modifications that promote simplification. Conversely, any changes that are unfavorable for simplification or result in a loss of meaning or fluency will be marked in red. Additionally, changes that neither enhance nor diminish the simplicity of the text will be indicated in orange.

Hyperpartisan News Detection

Within this section, we analyze five distinct examples that exhibit diverse outcomes in terms of simplification. Since HN texts are typically lengthy, we will only evaluate a portion of each instance for this manual assessment.

	Original Text	Simplified Text	S	F	MP
Ex1	Planned Parenthood will run ads targeting four incumbent Republican senators in tight races. Yes, the organization that cuts through the faces of babies with still beating hearts to extract the brain has the gall to attack vulnerable Republican senators while asking people to stand up for "Planned Parenthood healthcare."From The Hill: The ads will run in the home states of Sens. Kelly Ayotte (R-N.H.), Rob Portman (R-Ohio), Ron Johnson (R-Wis.) and Pat Toomey (R-Pa.), all of whom face tough reelection races next year.	Planned Parenthood is running ads against four Republican senators in tight races.Yes, the organization that cuts through the faces of babies with still beating hearts to extract the brain has the gall to ask people to stand up for "Planned Parenthood healthcare" while attacking vulnerable Republican senators. From The Hill: The ads will run in the home states of Sens. Kelly Ayotte (R-N.H.), Ron Johnson (R-Wis.), Rob Portman (R-Ohio) and Pat Toomey (R-Pa.), all of whom face re-election next year.	2	2	1
Ex2	Progressively-minded people do not make political stances because they're young and that's what young people do, it's because of an overarching desire to better our nation by providing the social, economic, and political liberties that our current incumbents have repeatedly failed to do. Wanting the right for every single person to go to school regardless of income or the opportunity for individuals to love one another without fear of violence or public humiliation is not a choice that progressives make because they are young, but because they are decent human beings who value people over profit.	Progressively-minded people don't take political positions because they're young, because that's what young people do. They do it because they want to better our country by giving social, economic, and political rights that our current leaders have failed to do. Wanting the right for every person to go to school no matter their income or for people to love each other without fear of violence or public humiliation is not a choice that progressives make because they are young. It is a choice they make because they are good people.	3	3	3
Ex3	The law of attraction: garbled fragment of a lost tradition – Jon Rappoport.There is no way to state the law of attraction with finality, because thousands of people have tinkered with it, and some of them earnestly believe they have the only "true" version. #160;I'll present several of the more popular descriptions first, and then comment.TND Guest Contributor: #160;Jon Rappoport.	The law of attraction: a lost tradition, by Jon Rappoport. There is no way to say the law of attraction is true, because thousands of people have changed it. Some people think they have the only true version, though. #160; I will present some of the most common descriptions first, and then comment. TND Guest Contributor: #160; Jon Rappoport. #160.	2	3	3
Ex4	Today is Africa Day, a day to celebrate the continent's culture, history and future. As part of our Protecting the Next Generation project, the Guttmacher Institute partnered with nine African institutions to speak with 20,000 young people in four Sub-Saharan countries. We asked them about their concerns and their hopes for the future in terms of their health, education and employment.	Today is Africa Day. It is a day to celebrate the continent's culture, history and its potential for the future. As part of our Protecting the Next Generation project, the Guttmacher Institute spoke with 20,000 young people in four Sub-Saharan countries with nine African institutions. We asked them about their concerns and their hopes for the future. We also asked about their health and education	2	3	2
Ex5	Police officers in New Mexico already accused of improperly subjecting one man stopped for a minor traffic offense to 14 hours of anal probes and a colonoscopy because they incorrectly thought he was carrying drugs inside his body now have been accused of doing that to a second victim.According to a report from KOB-TV in New Mexico, "our investigation reveals another chapter.	Police in New Mexico already accused of improperly subjecting a man to 14 hours of anal probes and a colonoscopy because they incorrectly thought he was carrying drugs inside his body have now been accused of doing the same thing to a second person. According to New Mexico's KOB-TV, "our investigation reveals another chapter."	2	3	3

Table 3.1: Examples of simplified text using MUSS on HN attack dataset. Modifications are indicated with color-coding defined.

Table 3.1 presents a representation of five distinct examples selected from the evaluated samples. This representation serves to provide a comprehensive understanding of the simplifier model's performance on the given dataset.

The initial sentences in Ex1 effectively simplify the text by replacing "targeting four incumbent" with "against four." Nonetheless, the second sentence, highlighted in red, fails to convey its intended meaning and does not enhance the simplification process.

Ex2, stands out for its commendable performance. It replaces certain sentences, such as "overarching desire to better our nation by providing" with "to better our country by givin" or "regardless" with "no matter." Additionally, it breaks down lengthy sentences.

Ex3 incorporates several alterations, such as simplifying the phrase "from garbled fragment of a lost traditio" to "a lost tradition," by eliminating complex vocabulary that does not add to the overall comprehension of the sentence. Nonetheless, the remaining sentences undergo only minor modifications.

In Ex4, the sentence "the Guttmacher Institute spoke with 20,000 young people in four Sub-Saharan countries with nine African institution" has been restructured in order to improve clarity. However, this reordering introduces a slight awkwardness that may lead to confusion regarding whether the institutions were spoken with or if they played a role in facilitating the conversation with the young people. Furthermore, the last sentence has been split into two parts to enhance readability. Nonetheless, this division results in the omission of "employment" from the original list, which could be perceived as a loss of crucial information.

In Ex5, challenging words like "subjecting" or "colonoscopy" are neither replaced nor explained, and the sentences are excessively long. While certain words like "colonoscopy" may not have direct replacements, they could be substituted with phrases like "Colon camera exam" or "Tube exam for the intestine."

With the observed examples, we can sense that, in general, simplifications are good. Some simplifications effectively streamline the content, while others may encounter challenges in altering the text in a coherent manner or overlook certain complex elements that require simplification. Despite minor discrepancies, the overall message of the text can still be comprehended effectively.

Rumor Detection

For the RD task, each thread consists of a flat feed of concatenated text fragments, which include the initial post as well as the subsequent responses. For the purpose of convenience, only a selection of comments will be displayed in this manual assessment.

Some of the evaluated examples in HN can be found in table 3.2 and display a different examples on how the model simplification perofrms in this task:

	Original Text	Simplified Text	S	F	MP
Ex1	The Earth without Art is just 'Eh'. #streetart #JeSuisCharlie http://t.co/72XpFM2kOn @charliesaidthat @inthecompanyof Unfortunately, for some murderers, the "earth" without "art" leaves just "He". Which is the problem. @charliesaidthat Your name is all over Twitter. @GaryDayEllison It feels like it. @charliesaidthat I actually saw your status saying #JeSuisCharlie before the news and wondered why you're saying you're Charlie in French... @charliesaidthat You said it. "@charliesaidthat: The Earth without Art is just 'Eh'. #streetart #JeSuisCharlie http://t.co/qvFOam85RJ" love this @charliesaidthat Rightly said! All the art lovers out there would surely agree with it. #BeInteresting #Streetart	The world without art is just 'Eh'. #streetart http://t.co/72XpFM2kOn @charliesaidthat @inthecompanyof Unfortunately, for some people, the "art" without the "earth" leaves just "He." @charliesaidthat His name is everywhere. @GaryDayEllison It looks like it. @charliesaidthat I actually saw your status saying #JeSuisCharlie before the news and I was wondering why you were saying that. @charliesaidthat You said it. "@charliesaidthat: The Earth without Art is just 'Eh'. #streetart #JeSuisCharli" - Charlie Sands @charliesaidthat Well said! All the art fans out there would certainly agree with it. #Streetart #BeInteresting	2	2	1
Ex2	In times of outrage, intolerance becomes contagious. Guard against those who blame all Muslims for Paris attacks. @billybragg Absolutely Billy and this is just the attitude and reaction that MSM will promote. @demindblower. @billybragg you are a true example to follow.	In times of crisis, intolerance can spread. Watch out for those who blame all Muslims for Paris attacks. @billybragg Definitely Billy and this is just the way the media will promote it. @demindblower. @billybragg you are a great example to follow.	3	3	3
EX3	Imagine the thousands of Muslims living in France whose lives will never be the same after today.	Imagine the thousands of Muslims living in France. Their lives will never be the same after this.	3	3	3
EX4	If criticism of Islam results in violence, this is hard evidence that criticism of Islam is desperately needed.	If criticism of Islam leads to violence, this is hard proof that criticism of Islam is needed very much.	2	3	3
EX5	The Prophet PBUH never killed anyone for blasphemy. There is no justification is #Islam for #CharlieHebdo murders. Trend #MuslimsForPeace	The Prophet PBUH never killed anyone for blasphemy. There is no justification is #Islam for #CharlieHebdo murders.	1	3	3

Table 3.2: Examples of simplified text using MUSS on RD attack dataset. Modifications are indicated with color-coding defined.

In Ex1, the text's meaning is changed by altering the order and making some modifications that do not improve the text's comprehensibility. Both "Earth" and "world" are equally clear, and changing "murderer" to "people" weakens the sentence's coherence. Ex2 shows a praiseworthy simplification. Complex terms such as "outrage," "contagious," and "guard" are replaced with more easily understandable alternatives. Additionally, Ex3 demonstrates a good simplification through sentence splitting. In Ex4, "Results in" is replaced with "leads," which may be more difficult to understand. However, some effective word substitutions are made, such as replacing "evidence" with "proof" and "desperately" with "much." Ex5: No changes are made; although the sentence is simple, certain words like "blasphemy" could be replaced for better clarity.

Fact Checking

In the FC task, we perform simplification separately; on the one hand, we simplify the claim, and on the other hand, we simplify the evidence. We have observed that in some instances, the evidence claimed was removed, which undermines the purpose of the task.

Some of the highlighted examples for FC are the following (see table 3.3): Ex1 is highlighted as an exemplary instance where the phrasal verb"passed over" is replaced with its meaning. Ex2 showcases commendable lexical simplification, such as substituting "construct" with "build" or "until" with "down to'.' However, the second part of the instance, used for fact-checking, conveys an incorrect meaning, which undermines the entire purpose of the task. This is an observed scenario in some fact-checking examples. Ex3 serves as a good example, demonstrating effective lexical simplification through word substitution and the division of lengthy sentences. Ex4, on the other hand, loses some meaning due to the omission of certain sentences. Although the initial sentence was already simple, minimal changes were required initially. Ex5 exhibits modified sentences that differ significantly from the original ones. The semantic similarity between them is extremely low. (Refer to the sentences highlighted in red.)

Propaganda detection

Some of the examples highlighted in this task are the following (see table 3.4): Ex1: In this particular instance, we have come across a term called antisemitism, which may not be familiar to everyone in terms of its meaning. Additionally, we can observe a more easily understandable replacement for the word "emerge", which is substituted with "happen".

Ex2: The meaning of the sentences in this example is reversed, as the original sentence suggests that he was given permission to retire, whereas the second

	Original Text	Simplified Text	S	F	MP
Ex1	Agent Raghav ,Äì Crime Branch. The serial received Indian Telly Awards in the category of Best Thriller and Horror Show and was also nominated for Best Weekend Show. Agent Raghav, Äì Crime Branch was passed over by the Indian Telly Awards.	Raghav - Crime Branch. The serial received Indian Telly Awards in the category of Best Thriller and Horror Show and was also nominated for Best Weekend Show but did not win the award for Best Serial.	3	3	3
Ex2	Martin Van Buren. He then moved on Washington where he did more than anyone to construct the modern Democratic Party which dominated American politics down to the American Civil War. Martin Van Buren was exclusively unelected.	Martin Van Buren. He then moved on Washington where he did more than anyone to build the modern Democratic Party that dominated American politics until the American Civil War. Martin Van Buren was elected only once.	3	2	1
Ex3	Port of Spain. Port of Spain is the capital city of the Republic of Trinidad and Tobago and the country 's third-largest municipality, after Chaguanas and San Fernando. Port of Spain is the capital of Trinidad and Tobago.	Port of Spain. Port of Spain is the capital city of Trinidad and Tobago. It is also the third largest city in the country, after Chaguanas and San Fernando. Port of Spain is the country's capital city.	3	3	3
Ex4	Joe Walsh.As a member of the Eagles , Walsh was inducted into the Rock and Roll Hall of Fame in 1998 , and into the Vocal Group Hall of Fame in 2001. Joe Walsh was barely inducted in 2001.	Joe Walsh. He was inducted into the Rock and Roll Hall of Fame in 1998. He was also added to the Vocal Group Hall of Fame in 2001. Joe Walsh was only added in 2001. much.	2	3	2
Ex5	Japan national football team. Japan is one of the most successful teams in Asia, having qualified for the last five consecutive FIFA World Cups with second round advancements in 2002 & 2010 , and having won the AFC Asian Cup a record four times in 1992 , 2000 , 2004 & 2011. Japan national football team had second round advancements in 2010.	Japan national team. Japan national football team is the national football team of Japan and is the most successful team in the world, having qualified for the last five FIFA World Cups, and having won the AFC Asian Cup four times, in 1992, 2000, 2004 and 2011. Japan national football team is the national football team of Japan.	1	2	1

Table 3.3: Examples of simplified text using MUSS on FC attack dataset. Modifications are indicated with color-coding defined.

	Original Text	Simplified Text	S	F	MP
Ex1	Antisemitism at McGill does not emerge from a vacuum.	Antisemitism at McGill does not happen by accident.	2	3	3
Ex2	Why was Cardinal Mahony allowed to retire in good standing?	Why was Cardinal Mahony allowed to stay in office?	3	2	1
EX3	I have no idea how we are going to implement all these ridiculous plans of 'on-going formation'.	I don't know how we're going to implement these ridiculous "on-going formation" plans.	2	3	3
Ex4	Whataburger told Fox News that it is cooperating with the police investigation.	Whataburger told Fox News that it is helping police with their investigation.	3	3	3
Ex5	The Guardian has been vocal in the so-called "fake news" hysteria, decrying the influence of social media, the only place where leftwing dissidents have managed to find a small foothold to promote their politics and counter the corporate media narrative.	The Guardian has been vocal in its opposition to the so-called "fake news" hysteria,decrying the influence of social media. The only place leftwing dissidents have found a small foothold is on social media to promote their politics.	2	2	1

Table 3.4: Examples of simplified text using MUSS on PR2 attack dataset. Modifications are indicated with color-coding defined.

sentence implies that he was given permission to remain in his position.

Ex3: Although the sentence in question has been rephrased, it has not been made simpler.

Ex4: This is a good simplification, as it eliminates complex elements that are not necessary for a general understanding. However, some of the highlighted words could be substituted with alternative terms.

Ex5: Difficult words have not been replaced in this case.

Conclusions on simplification

The utilization of the MUSS tool for simplification highlights various key points. Primarily, the tool effectively improves text comprehensibility by substituting complex words with simpler alternatives, restructuring sentences, breaking down lengthy sentences, and eliminating unnecessary parts that hinder understanding.

Nevertheless, challenges arise during the simplification process with MUSS. At times, the alterations made are insufficient to truly simplify the text, or the modifications made impact the text's intended meaning. This becomes particularly problematic when dealing with already simple sentences, as the tool tends to make unnecessary changes that result in a loss of meaning or unnecessary paraphrasing. These challenges can also have a negative impact on the text's readability and coherence.

Regarding FC instances we have to consider that alterations made to the verification component within the text have the potential to completely change the results of the task. If the interpretation evidence is modified, the classification of the instance becomes inaccurate, resulting in potential errors. Despite these obstacles, the fluency score suggests that, although some results may not be simpler or may alter the meaning, the text generally maintains a strong syntactical structure.

Overall, it is evident that there is room for improvement in the simplification process. While some examples are completely inaccurate, changing the meaning of the sentence or even reversing it, others fall somewhere in between achieving successful simplification. Additionally, since MUSS relies on sentence paraphrasing, some instances may be modified without enhancing simplicity.

3.4 Attack

An attack consists of sending a perturbed piece (adversarial modification) of non-credible content to the victim model. The success scenario corresponds to the one where this piece of text is falsely recognized as credible thanks to the adversarial modification. Therefore, if the classifier changes its decision with respect to the original text, the attack is valid.

To create an attack we need to define two fundamental components:

- A victim model f, predicting a class label $\hat{y}_i$ based on instance features: $\hat{y}_i = f(x_i)$,
- A modification function m, turning x_i into an adversarial example $x_i^* = m(x_i)$.

The victim model is a classifier that returns 0 for content classified as credible and 1 for non-credible content, as well as the likelihood score. The modification function, also called the attacker function, will be, in our case, the simplifier model.

The attack is performed in a grey-box scenario in which we assume that the attacker knows the architecture of the classifier as well as the training, development, and evaluation sets. Without using the complete knowledge of the model, we maintain some resemblance to practical scenarios.

The experiments will be performed in both untargeted scenario, where any change in the victim's predictions is considered a success, and targeted scenario, which seeks to obtain a specific response, in this case changing the classification of non-credible content to be considered credible. Moreover, in the context of the attacking scenario, it is assumed that the attacker has unlimited access to query the victim model. So an infinite number of queries are assumed.

In our research, we leverage BODEGA as the primary framework for evaluating adversarial attacks on text classification models. BODEGA is an advanced system designed to perform adversarial attacks aimed at misinformation detection. It is built on the principles of Openattack methodologies and specifically tailored to enhance the robustness and accuracy of misinformation detection algorithms. BODEGA aims to detect misinformation by simulating various adversarial scenarios, which have been described in detail in previous sections of this paper (3.2).

To evaluate successful attacks, we use the metrics defined by BODEGA framework in refrence [Przybyła et al., 2023]. The measures used are the following:

- Semantic score: It is an average of individual semantic scores in cases with changed decision. The semantic score (Succ_score(x_i, x_i^*)) of individual instances is a measure of meaning preservation based on BLEURT [Sellam et al., 2020]. BLEURT is a measure at sentence level, it requires using LAMBO [Przybyla, 2022] to split sentences and then finds alignment between original and simplified pairs using Levenshtein [Levenshtein, 1965] distance before computing semantic similarities between sentence pairs. The values are from 0 to 1, being 1 identical text.
- Character Score: Computed as an average of the character score over the cases with changed decision. For individual examples, Char_score(x_i, x_i^*):

assesses the similarity between the original and simplified text based on character-level changes. It uses Levenshtein distance to express how different one string of characters is from another. And is computed:

$$\text{Char\_score}(a, b) = 1 - \frac{lev\_dist(a, b)}{max(|a|, |b|)}$$

Char_score also ranges between 0 and 1, with higher values corresponding to larger similarity

- BODEGA score: Defined as a measure of the quality of adversarial modifications in successful attacks and is computed as an average of all individual BODEGA scores in the attack. Each individual score being computed as:

$$\begin{aligned} BODEGA\_score(x_i, x_i^*) = Succ\_score(x_i, x_i^*) \times Sem\_score(x_i, x_i^*) \\ \times Char\_score(x_i, x_i^*) \end{aligned}$$

It ranges from 0 to 1, where a high value indicates that the modification preserves the original meaning, while a low value indicates poor modification.

- Success Score: Also called confusion score, it indicates the proportion of successful attacks.

- Queries per Example: The average number of queries made to the model per example.

3.5 Victim Models

For our experiments we will perform the attacks to two different victim models defined by the BODEGA and implemented using PyTorch: BiLSTM and BERT.

1. BERT: BERT model is based on the "bert-base-uncased" pre-trained model [Devlin et al., 2019], a transformer-based architecture designed for sequence classification tasks. It uses the BERT tokenizer for text preprocessing and is fine-tuned with a linear classifier to output probabilities for two classes. The optimization process employs the AdamW optimizer with linear weight decay starting from 0.00005, and it is trained for 5 epochs.

2. BiLSTM: BiLSTM model employs a bidirectional Long Short-Term Memory (LSTM) architecture [Hochreiter and Schmidhuber, 1997]. It uses the same

"bert-base-uncased" tokenizer as BERT model for consistency. The architecture consists of an embedding layer with vectors of length 32, a bidirectional LSTM layer with a hidden size of 128, a linear layer to map the LSTM outputs to the final output space, and a log softmax layer for computing class probabilities. The model is optimized using the Adam optimizer a learning rate of 0.001 and batches of 32 examples each.

Both models are trained to classify text into two categories.

3.6 Attacker Model

The attacker is defined using OpenAttack guidelines [Documentation, 2024]. We use BODEGA[1] which provides code and pretrained models to reproduce experiments. In our attacks we first apply syntactic tool LAMBO [2] to the input sentences to divide them prior to applying simplification tool MUSS [3] at sentence level.

[1]https://github.com/piotrmp/BODEGA
[2]https://gitlab.clarin-pl.eu/syntactic-tools/lambo
[3]https://github.com/facebookresearch/muss

Chapter 4

RESULTS

Within this chapter, we undertake a comprehensive analysis of the attack conducted on the four misinformation tasks. Both targeted and untargeted scenarios were employed, utilizing the victim models BERT and BiLSTM. Through this analysis, we aim to highlight several noteworthy observations. Each task will be individually examined, with a focus on the distinct behaviors observed and the corresponding numerical values obtained. By thoroughly exploring these findings, we can gain valuable insights into the effectiveness and implications of our attack.

4.1 Attack performance

After executing the attack detailed in Chapter 3, we proceed to evaluate the outcomes obtained from the four different tasks. In this particular section, we will conduct a comprehensive analysis of the numerical results provided by the designated metrics, while also undertaking a manual evaluation of the results to determine the level of success achieved. To facilitate the manual evaluation, we will utilize the criteria defined in previous sections, wherein metrics of similarity (S), fluency (F), and meaning preservation (MP) are assigned ratings ranging from one to three.

4.1.1 Evaluation: Task HN

Upon examining the numerical outcomes derived from performing the experiments on the HN task (see table 4.1), it is noticeable that the BODEGA score is notably low. Evaluation of the other metrics shows that only a small proportion of examples achieved success, roughly 10% in the targeted attack and 20% in the untargeted attack. As illustrated in table 4.1, both the semantic and character scores surpass

HN	Success score	Semantic score	Character score	BODEGA score	Number Queries
Targeted					
BERT	0.139	0.686	0.733	0.070	2.139
BiLSTM	0.155	0.662	0.711	0.073	2.155
Untargeted					
BERT	0.17	0.682	0.723	0.083	2.17
BiLSTM	0.20	0.67	0.71	0.09	2.20

Table 4.1: Results of Adverarial Attack with simplification on HN task

the values of 0.6 and 0.7, respectively. This indicates a substantial preservation of meaning and similarity between the original and simplified pairs.

In general, the examples that manage to confuse the victim model (see in table 4.2) effectively simplify the text, making it more readable and concise. However, not all simplifications are beneficial. Certain modifications in the text result in sentences that do not convey the same meaning as the original ones, while others introduce subtle nuances and altered emphases. For instance, in table 4.2 Ex5, the simplified sentence loses the specific allegation "secret government biowarfare research with ticks and other animals on Plum Island," which is crucial to the original meaning. Similarly, in 4.2 Ex3, the sentence "who flee the comparatively poor conditions of their homeland, made worse with political conflict" is not entirely analogous to "who have been forced to leave their homeland because of political conflict." The former suggests that the conditions of the homeland are the cause, while the latter attributes it solely to political conflict. Although the other showcased examples demonstrate modifications, none of them significantly alter the overall meaning of the text.

Additionally, with regard to the attack, while certain revisions have been made to simplify the text and make it more straightforward (as highlighted in table 4.2), further modifications could be undertaken to make the content even more accessible.

It is important to highlight that, due to the lengthy nature of the texts involved in the HN task, any slight modifications in tone, emphasis, or meaning within certain parts of the text are less likely to affect the overall labeling of the text. This is why semantic and character scores rank highest among all tasks, as minor alterations have a lesser impact on the overall outcome.

4.1.2 Evaluation: Task RD

In the RD task, similar observations are noted as in HN, with even lower BODEGA values. Despite a similar proportion of successful instances, the semantic score is lower. This suggests that the simplification function in RD instances faces challenges in generating effective adversarial attacks. These results are according

id	Original Text	Simplified Text	S	F	MP
Ex1	The Syrian Arab Army (SAA) has reported today that the entirety of East Aleppo is fully back under government control, meaning the city is now **completely liberated**. The SAA has completed the evacuations of **anti-government fighters and civilians** looking **to flee** with these groups as of today. This is a major victory for the Syrian forces in Aleppo **coming after** almost 4 years of fighting in the city. Thousands of people have already taken to the streets to celebrate the last of the terrorists inside the city leaving. The fighters were removed in government-provided buses and have gone to the suburbs **on the southwest outskirts of the provincial capital.**	The Syrian Arab Army (SAA) has reported today that all of East Aleppo is back under government control. This means that the city is now completely **free**. The SAA has completed the evacuation of **civilians and anti-government fighters** looking **to leave** with these groups as of today. This is a major victory for the Syrian army in Aleppo after almost four years of fighting in the city. Thousands of people have taken to the streets to celebrate the last of the terrorists leaving the city. The fighters were removed in government-provided buses and have gone to t**he provincial capital's southwest outskirts.**	2	3	3
Ex2	Walker's deputy knew **it was wrong** scott walker and his **defenders steadfastly** have **evaded** questions about the existence and use of a secret email network to **evade** open records laws . the #walkerdocs destroyed any **ambiguity**.	Walker's deputy knew. Scott Walker and his **supporters have consistently avoided** questions about the existence and use of a secret email network to **avoid** open records laws. the #walkerdocs removed any **doubt**.	3	3	3
Ex3	DR have a shared history of colonialism based on sugar cane, an industry which continues to exist **mostly** on the backs of low-paid Haitian migrant workers **who flee the comparatively poor conditions** of their homeland, **made worse with** political conflict and natural disaster, the most recent of which being the 2010 earthquake	DR have a shared history of colonialism based on sugar cane, an industry which continues to exist mostly on the backs of low-paid Haitian migrant workers, **who have been forced to leave** their homeland **because of political conflict**, natural disaster, and, most recently, the 2010 earthquake.	3	3	1
Ex4	With the prospect of another shutdown looming when the next continuing resolution runs out on Feb. 8, federal employees should also know that contrary to a few initial news reports, the back pay p**rovision as approved by Congress** applies only to last weekend's **shutdown**. It does not guarantee back pay for furloughed feds in the event of another shutdown in fiscal 2018, according to Aaron Fritschner, a spokesman for Rep. Don Beyer, D-Va., who has been a key advocate of back pay for furloughed federal workers. Congress would need to pass separate language **to grant back pay for feds sent home in the event of another lapse in appropriations**. .Erich Wagner **contributed** to this report.	With the prospect of another shutdown looming when the next continuing resolution runs out on Feb. 8, federal employees should also know that, contrary to some initial news reports, the back pay will only apply to last weekend's **closure**. It does not guarantee back pay for furloughed federal workers in the event of another shutdown in fiscal 2018. Aaron Fritschner, a spokesman for Rep. Don Beyer, D-Va., **said the bill does not guarantee back pay.** Congress would need to pass separate language **to give back pay to federal employees sent home if there is another funding lapse**. Erich Wagner **wrote** this report.	1	3	2
Ex5	The Officially **Ignored** Link Between Lyme Disease and Plum IslandBy; Melissa Dykes In this video, Melissa Dykes of Truthstream Media breaks down the origins of Lyme disease **and how it appears to be linked to Plum Island secret government biowarfare research done with ticks and other animals**. Aaron amp; Melissa Dykes are the founders of; TruthstreamMedia.com.	The Official Link Between Plum Island and Lyme Disease. By Melissa Dykes In this video, Melissa Dykes of Truthstream Media breaks down the origins of Lyme disease **and how it appears to be linked to Plum Island and other places** where people say they have the disease. Aaron amp; Melissa Dykes are the founders of; TruthstreamMedia.com.	2	3	1

Table 4.2: Examples of good adversarial modifications that were successful, performed by MUSS against BiLSTM. Changes are highlighted in boldface.

to the results obtained in reference [Przybyła et al., 2023]. The study suggests that bad results in this task may be attributed to the fact that since each rumor consists of numerous posts and each has some indication of the credibility of the news, changing just one or some of them may not be enough to change the attacker's decision. This assertion aligns with our manual examination of certain instances, where we observe the simplifier struggling to make changes to the input text.

The reasons behind the struggles of the model in simplifying the text can be attributed to several factors. One of the main factors is the irregular nature of the input text, which consists of tweet posts and follow-up responses from social media users. These texts often contain informal language, including widely recognized expressions and references. These elements can pose challenges for the simplifier model. For instance, in table 4.4 Ex1, a user utilizes the phrase "kudos to Google," which means "good work!". However, the simplifier model simply omits this expression. While this may not seem significant in this specific case, it can be more critical in other instances. Furthermore, the model encounters additional irregularities, such as responses in different languages, which it is unable to simplify, as demonstrated in Ex5.4 in table 4.4. Additionally, poorly structured sentences, abbreviations, spelling mistakes, and hashtags can also present challenges. All of these factors contribute to the poor performance of the simplification algorithm, resulting in inadequate or no changes made to the original input.

From the manual evaluation of texts, we can derive certain observations in relation to simplification:

In Ex1 several sentences have been altered, with most of them being simplified effectively, while a few introduce subtle nuances to the intended meaning. However, the overall meaning of the text remains unaltered.

In Ex2, the simplification is good, but omitting "Charlie Hebdo" eliminates crucial context from the text, which may mislead algorithms. Also, this instance is short, so small changes have a bigger impact. A similar scenario is observed in Ex4, where the statement "Charlie Hebdo suspects tell police they want to die as martyrs" differs from "Police tell Charlie Hebdo suspects they want to die as martyrs."

Ex3 is a good attack; to enhance comprehension, certain components have been omitted while ensuring the preservation of its intended meaning.

To enhance the understandability of the text, there is room for additional modifications in the majority of the examples provided.

4.1.3 Evaluation: Task FC

In FC, the BODEGA score is relatively high in comparison with previous tasks (see table 4.6). By observing other metrics, we can attribute this to an improved success score, especially with victim BiLSTM where the attacker manages to confuse the

RD	Success score	Semantic score	Character score	BODEGA score	Number Queries
Targeted					
BERT	0.1	0.529	0.731	0.039	2.1
BiLSTM	0.106	0.589	0.732	0.046	2.106
Untargeted					
BERT	0.060	0.559	0.721	0.020	2.057
BiLSTM	0.060	0.611	0.740	0.027	2.057

Table 4.3: Results of Adversarial attack with Simplification on RD

algorithm around 25%. Regarding semantic and character score we are obtaining similar values than in previous tasks between 0.6 and 0.7 in both cases. In the FC task, the attacker's decision is based on the evidence claim at the end of each instance. In this attack, we can observe a higher success score than in the previous task, as the nature of the dataset makes smaller changes to have a bigger impact on the model's decision.

Upon analyzing successful attacks, three main scenarios emerge: instances where the victim model is effectively misled by a well-crafted simplified text that retains the original meaning (table 4.7); cases where the attack is successful, but the simplifier model distorts the meaning of the evidence, thereby facilitating the model's confusion (table 4.8); and finally, situations where the simplifications are poor, resulting in an attack that fails to preserve the original meaning (table 4.9).

4.1.4 Evaluation: Task PR2

In PR2, the outcomes exhibit a comparable trend to those in FC. The rate of successful outcomes in specific instances is around one-third in targeted assaults and 10% in untargeted ones, resulting in higher BODEGA results. Again, as in the previous task, the short nature of the input text makes it more vulnerable to modifications. Nevertheless, in this particular task, we achieve the top score for semantic metric in all scenarios very good results for character score, with around 70% of the text remaining unaltered (refer to table 4.10).

Through a manual analysis of the resulting instances, we can observe that there are some instances where the simplifier model performs well by making accurate modifications (see table 4.11). However, it becomes apparent that there are other cases where there are nuances in the meaning of the simplified sentences. While these nuances generally do not fundamentally change the overall meaning, they do introduce slight modifications to the tone or specificity of the text, which could potentially impact the algorithm's decision, as in the cases shown in table 4.12. For example, in Ex1, the original sentence implies that the information is sent to agents or entities acting on behalf of foreign interests, whereas the simplified version generalizes to other countries. It captures the main idea but uses more general

Id	Original Text	Simplified Text	S	F	MP
Ex1	1. 1 of the cops killed at #charliehebdo was, like most **victims of Islamist militants**, a Muslim. http://t.co/qmnj7r9zja http://t.co/s9apcajcjf 2. Of Arabic descent **indeed**. But it doesn't imply anything about his **faith**. @astroehlein 3. @anapalacio They were two Muslims: Moustapha Ourrad and Ahmed Merabe. @astroehlein @agvicente 4. Ya estamos con el relativismo. Basta de complejos y ponerse de perfil ante éste peligro, antes de que sea tarde. @astroehlein 5. @kdashtgirkhan Don't call them Islamist **militants**, just call militants or terrorists. Islam does not allow this. @astroehlein 6. @tounsiahourra Killed after being wounded these are lower than animals. Animals kill for food, they kill for the lust of blood. @astroehlein 7. @kdastgirkhan Sir, not fair to say Islamist as Islam does not permit terrorism. Will it be ok if we say **rapist** by his religion? @astroehlein 8. Nowadays, Muslims are most likely victims of any militants @astroehlein 9. So what? @astroehlein 10. This proves that it is not Islam but a group of Muslims **have gone stray**. Also abuse in name of free speech should be stopped. @astroehlein 11. Terrorism in all forms must be condemned in strongest terms. Those who committed Paris attack should face justice. #condemnthem 12. @khamoshi36 @astroehlein @kdastgirkhan Can't brush it off that easily. Islam's prophet ordered rape & kill all women/children by his army. 13. @khamoshi36 @astroehlein @kdastgirkhan Mohammed was no pacifist like the Nazarene, and his followers are still as bloodthirsty. 14. @clearseer @astroehlein @kdastgirkhan Absolutely wrong. Even in war, the prophet had said no killings of children, women, old age. Disinformation @astroehlein 15. No. He was an officer on duty, so when he was shot, he was neither Muslim nor Christian **nor atheist**; he was France. @astroehlein 16. @clearseer @astroehlein @kdastgirkhan Islam strictly prohibits killings of innocents **and appreciates for**	1. A police officer killed at #charliehebdo was, like most **Islamist victims**, a Muslim. http://t.co/qmnj7r9zja http://t.co/s9apcajcjf 2. **Of course**, Arabic descent. But that doesn't mean anything about his **religion**. @astroehlein 3. @anapalacio There were two Muslims: Moustapha Ourrad and Ahmed Merabe. 4. Ya estamos con la relativismo. Basta de complejos y ponerse de perfil ante éste peligro, antes de que sea tarde. 5. @kdastgirkhan I don't call them **terrorists**, I call them terrorists. Islam does not allow this. @astroehlein 6. @tounsiahourra Killed after being wounded these are lower than animals. Animals kill for food, they kill for the lust of blood. @astroehlein 7. @kdastgirkhan This is not what Islam is all about. Will it be ok if we call him a **murderer** by his religion? @astroehlein 8. Nowadays, Muslims are most likely victims. 9. What? @astroehlein 10. This shows that it is not Islam, but a small group of Muslims. Also abuse in name of free speech should be stopped. @astroehlein 11. Terrorism in all forms must be condemned in strongest terms. #condemnthem 12. @khamoshi36 @astroehlein @kdastgirkhan You can't brush it off easily. Islam's prophet ordered rape & kill all women/children by his army. 13. Absolutely wrong. Mohammed was no pacifist like the Nazarene, and his followers are still as bloodthirsty. 14. Even in war, the prophet said **no murders** of children, women, old age. @astroehlein 15. No. He was an officer on duty, so when he was shot, he was France. He was neither Muslim nor Christian; he was France. 16. slam strictly does not allow killings of innocents.	2	2	2

Table 4.4: Examples of adversarial modifications that were successful, performed by MUSS against BiLSTM, on RD. Changes are highlighted in boldface.

Id	Original Text	Simplified Text	S	F	MP
Ex2	The latest on the **manhunt** for Charlie Hebdo suspects: http://t.co/F3hv000UQXhttp://t.co/r0TgWbIxsr @WSJ Can't imagine people's fear and lost sense of **safety**.	The latest on the **search** for the suspects: http://t.co/F3hv000UQX, http://t.co/r0TgWbIxsr Can't imagine the fear and loss of **security**.	3	3	2
Ex3	1. **Kudos to Google** for donating €250,000 to help Charlie Hebdo publish next week http://t.co/QnmfLssRtf 2. @mathewi @glynmoody Good work! If there ever was a defender of privacy,free speech and democracy, its Google ! **That's what friends are for** 3. @mathewi If **stupidity** is not to win, those who knows open a fund rising account named to Charlie Hebdo 4. @mathewi Did you read the article? "**Financed** [by a fine] but not managed by Google that money will go to support **the survival** of the weekly"	1. Google has given €250,000 to help Charlie Hebdo publish their next issue http://t.co/QnmfLssRtf 2. @mathewi @glynmoody Good work! If there ever was a defender of freedom of speech and privacy, its Google. 3. @mathewi If **ignorance** is not to win, those who know open a fund named after Charlie Hebdo 4. @mathewi Did you read the article? "**The money, paid** by Google but not managed by Google, will go to support the weekly"	3	3	3
Ex4	Let the hostages go first, but OK. RT jaketapper: Charlie Hebdo suspects tell police they want to die as martyrs http://t.co/cb9zSckTZ2 @JasStanford @WhitneyNeal @jaketapper Only if those Lady Kurd soldiers can do it, w**ith bacon-laced ammunition.** @JasStanford @WhitneyNeal @jaketapper Just saying, Danielle Mitterrand was a huge fan of the Lady Kurds. **Bet they'd volunteer.**	Let the hostages go first, but OK. RT @jaketapper: **Police tell Charlie Hebdo suspects they want to die as martyrs.** @JasStanford @WhitneyNeal @jaketapper Only if those Lady Kurd soldiers can do it, right? @JasStanford @WhitneyNeal @jaketapper Just saying, Danielle Mitterrand was a Lady Kurds fan. **I'd volunteer**.	2	3	1

Table 4.5: Some examples of adversarial modifications that were successful, performed by MUSS against BiLSTM, on RD. Changes are highlighted in boldface.

FC	Success score	Semantic score	Character score	BODEGA score	Number Queries
Targeted					
BERT	0.194	0.652	0.691	0.088	2.194
BiLSTM	0.243	0.671	0.680	0.112	2.243
Untargeted					
BERT	0.180	0.653	0.670	0.079	2.180
BiLSTM	0.244	0.667	0.670	0.110	2.244

Table 4.6: Results of Adversarial attack with Simplification on FC

id	Original Text	Simplified Text	S	F	MP
Ex1	The Mod Squad. The Mod Squad **is** an American crime drama series that ran on ABC from 1968 to 1973. The Mod Squad is a Peruvian television series.	The Mod Squad. The Mod Squad **was** an American television series. It ran from 1968 to 1973. The Mod Squad is a Peruvian television series.	3	3	2
Ex2	Mani Ratnam. Cited by the media as one of India's **influential** filmmakers, Mani Ratnam is widely credited with **revolutionising** the Tamil film industry **and altering the profile of Indian cinema**. Mani Ratnam is **widely credited with altering** the profile of Indian cinema.	Mani Ratnam. Cited by the media as one of India's most **important** filmmakers, Mani Ratnam is widely credited with **changing** the Tamil film industry **for the better**. Mani Ratnam's films have **changed the profile** of Indian cinema.	3	3	3

Table 4.7: Examples of well-crafted adversarial modifications that were successful, performed by MUSS against BiLSTM, on FC. Changes are highlighted in boldface.

id	Original Text	Simplified Text	S	F	MP
Ex3	Sabbir Khan. In 2009 he made his **directorial debut** with the film Kambakkht Ishq (2009) that starred Akshay Kumar and Kareena Kapoor. Sabbir Khan directed **zero** films.	Sabbir Khan. In 2009, he **directed his first movie**, Kambakkht Ishq. The movie starred Akshay Kumar and Kareena Kapoor. Sabbir Khan directed **one** movie.	3	3	1
Ex4	Janet Leigh. Janet Leigh (born Jeanette Helen Morrison; July 6, 1927 – October 3, 2004) was an American actress, singer, dancer and author. Janet Leigh was **Canadian**.	Janet Leigh. Janet Leigh (born Jeanette Helen Morrison; July 6, 1927 â October 3, 2004) was an American movie and television actress. Janet Leigh is **American**	2	3	1

Table 4.8: Examples of adversarial with incorrectly simplified evidence claim, that were successful, performed by MUSS against BiLSTM, on FC. Changes are highlighted in boldface.

id	Original Text	Simplified Text	S	F	MP
Ex5	**Exit** the King. Exit the King (Le Roi se meurt) is an **absurdist drama** by Eugène Ionesco that premiered in 1962. Exit the King the show **premiered** in 1962.	The King **is gone**. Exit the King (French: Le Roi se meurt) is a French **comedy** by Eugène Ionesco. Exit the King The show **started** in 1962.	3	3	1
Ex6	Earl Scruggs. Earl Eugene Scruggs (January 6, 1924 – March 28, 2012) was an American musician **noted for** popularizing a three-finger banjo picking style, now called "Scruggs style," that is a defining characteristic of bluegrass music. Earl Scruggs **was incapable of** being involved in bluegrass music.	Earl Scruggs. Earl Eugene Scruggs (born January 6, 1924 - died March 28, 2012) was an American musician. He is **most famous for** popularizing the three-finger banjo style, which is now known as the Scruggs style. Earl Scruggs **did not want to** get involved in bluegrass music.	2	3	1

Table 4.9: Examples of adversarial modifications that altered the meaning but were successful, performed by MUSS against BiLSTM, on FC. Changes are highlighted in boldface.

PR2	Success score	Semantic score	Character score	BODEGA score	Number Queries
Targeted					
BERT	0.388	0.669	0.605	0.159	3.805
BiLSTM	0.333	0.702	0.692	0.162	2.333
Untargeted					
BERT	0.088	0.719	0.704	0.045	2.088
BiLSTM	0.168	0.725	0.697	0.085	2.168

Table 4.10: Results of Adversarial attack with Simplification on PR2

id	Original Text	Simplified Text	S	F	MP
EX1	Farrell's **graciously offers notorious praise** for Martin's new book, Building a Bridge (to hell, ed.)	Farrell has also given **good reviews** to Martin's new book, Building a Bridge (to hell, ed).	3	3	3
Ex2	Antisemitism at McGill does not emerge **from a vacuum**.	Antisemitism at McGill does not happen **by accident**.	3	3	3

Table 4.11: Examples of good adversarials modifications that were successful, performed by MUSS against BiLSTM, on PR2. Changes are highlighted in boldface.

language, making it less specific and less intense. In terms of simplification, it could be improved by replacing "alleged" with a simpler term. Also in table Ex2, the sentence "using any despicable tactic at hand to derail" is omitted in the simplified version, making it less biased and less emotionally charged. The simplified version still conveys that Senate Democrats are being accused of character assassination, but it does so in a way that is slightly less inflammatory. In Ex3, the simplified version does not specify migrant caravans, which could mislead the meaning of the text.

4.2 Performance Analysis

In the analysis of the performance of the automatic metrics and the manual evaluation of successful results, several conclusions can be drawn. The evaluation of the BODEGA score indicates that the use of the BiLSTM victim model attack generally resulted in the highest BODEGA score and success rate, suggesting that BiLSTM model is more vulnerable to attacks than BERT.

The semantic scores ranged from 0.5 to 0.7, indicating a decent preservation of meaning, with some cases showing particularly good results while others had a larger gap in meaning between the original and simplified sentences. The character score also fell within the same range, suggesting that the texts were quite similar.

From the manual evaluation of successful AEs we can see that some of them are well crafted with a good simplification that satisfactorily fulfills the requirements and obtains good values for the manually evaluated metrics. Other examples fail to preserve their meaning, making it easier for the victim model to misclassify them.

id	Original Text	Simplified Text	S	F	MP
Ex1	House investigators found the House server was being used for **nefarious purposes**: the alleged Muslim spies were **removing** information and sending it for **foreign actors.**	House investigators found that the House server was being used **for bad purposes**. The alleged Muslim spies were **taking** information and sending it to **other countries**.	2	3	2
Ex2	**Using any despicable tactic at hand to derail Judge Brett Kavanaugh's Supreme Court** confirmation less than a week before the Senate Judiciary Committee is scheduled to vote on whether to approve his nomination, Senate Democrats have sunk to their lowest level of character assassination yet.	With less than a week before the Senate Judiciary Committee is scheduled to vote on whether to approve Judge Brett Kavanaugh's Supreme Court confirmation, Senate Democrats have sunk to their lowest level of character assassination yet.	3	3	2
Ex3	While the US still has very little **intel** about the composition of the **migrant** caravans, one way or the other the White House plans to outmatch the number of **potential illegals** and, if not, suppress them with **brutal force**.	While the US still has little **information** about the composition of the caravans, one way or the other, the White House plans to outmatch the number of caravans and, if necessary, **force them back**.	3	3	1

Table 4.12: Examples of adversarial modifications that were successful, performed by MUSS against BiLSTM, on PR2. Changes are highlighted in boldface.

Since we can easily see these changes, we can tell that those AEs are not fulfilling their main goal, which is to remain unnoticeable [Zhang et al., 2020].

We can see that tasks with shorter input texts (FC and RD) have the highest success rate, as smaller changes have a bigger impact on the victim's model decision. While tasks with longer input (HN and RD) have a lower rate of confusion, as for changing the model output, more changes are needed. In terms of the different tasks, it must be noted that PR and FC generally produced better results, as it involved straightforward short texts, which the attacker model had fewer problems simplifying and therefore producing more accurate AEs and the model was more easily confused. On the other hand, Task RD had very few successful attacks, resulting in a low BODEGA score due to the difficulty of the model in simplifying this type and the long nature of these texts.

The investigation carried out by [Przybyła et al., 2023] provides an in-depth analysis of various attack techniques for the tasks previously outlined. The comparison of results is visible in tables 4.14 and 4.15, which display the findings of this study in conjunction with the results obtained through our implementation of MUSS in conducting adversarial attacks. A significant observation is that, despite having low confusion and BODEGA scores, the number of queries required to generate successful examples is substantially reduced with MUSS. For instance, while SCCN attacks typically necessitate around 11 queries, our simplified approach only required an average of 2 to 3 queries. In contrast, the worst case seen is Genetic algorithms, which required up to 2000 on average to achieve successful results, emphasizing the favorable outcomes of our AE.

	HN	**RD**	**FC**	**PR**
	Targeted			
BERT	0.07	0.03	0.08	0.15
BiLSTM	0.07	0.04	0.11	0.16
	Untargeted			
BERT	0.08	0.02	0.07	0.04
BiLSTM	0.09	0.07	0.11	0.08

Table 4.13: Comparison of BODEGA score results using BiLSTM and BERT victim models

	Untargeted					Targeted				
Method	**B.**	**succ**	**sem**	**char**	**Q.**	**B.**	**succ**	**sem**	**char**	**Q.**
HN										
BAE	0.34	0.60	0.58	0.96	606.83	0.18	0.34	0.57	0.95	713.42
BERT-ATTACK	**0.60**	**0.96**	0.64	0.97	648.41	**0.57**	**0.95**	0.62	0.96	753.91
DeepWordBug	0.22	0.29	**0.78**	**1.00**	395.94	0.15	0.20	**0.78**	**1.00**	389.81
Genetic	0.40	0.86	0.47	0.98	2713.80	0.30	0.71	0.44	0.97	4502.51
SememePSO	0.16	0.34	0.50	0.99	341.70	0.05	0.12	0.44	0.99	417.99
PWWS	0.38	0.82	0.47	0.98	2070.78	0.27	0.64	0.44	0.95	2107.02
SCPN	0.00	0.92	0.08	0.02	11.84	0.00	**0.95**	0.09	0.02	11.89
TextFooler	0.39	0.92	0.44	0.94	660.52	0.32	0.85	0.41	0.90	850.79
MUSS	0.08	0.17	0.68	0.72	2.17	0.07	0.13	0.68	0.73	2.13
PR										
BAE	0.11	0.18	0.69	0.94	33.96	0.13	0.20	0.68	0.94	45.68
BERT-ATTACK	0.43	0.70	0.68	0.90	80.16	**0.50**	0.79	0.69	0.92	99.95
DeepWordBug	0.28	0.36	**0.79**	**0.96**	27.43	0.50	0.64	**0.81**	**0.96**	36.04
Genetic	**0.50**	**0.84**	0.65	0.89	962.40	0.49	**0.84**	0.65	0.89	1211.56
SememePSO	0.41	0.68	0.66	0.90	96.17	0.35	0.53	0.71	0.91	173.71
PWWS	0.47	0.75	0.68	0.91	131.92	0.44	0.72	0.68	0.89	179.68
SCPN	0.09	0.47	0.36	0.46	11.47	0.11	0.79	0.32	0.39	11.79
TextFooler	0.43	0.77	0.64	0.87	57.94	0.46	0.77	0.66	0.89	77.81
MUSS	0.04	0.08	0.71	0.70	2.08	0.15	0.38	0.66	0.60	3.80
FC										
BAE	0.34	0.51	0.70	0.96	80.69	0.18	0.27	0.70	0.94	92.47
BERT-ATTACK	**0.53**	0.77	0.73	0.95	146.73	**0.41**	0.62	0.71	0.93	207.23
DeepWordBug	0.44	0.53	**0.84**	**0.98**	54.32	0.22	0.27	**0.85**	**0.98**	52.31
Genetic	0.52	0.79	0.70	0.95	1215.19	0.39	0.63	0.66	0.92	1808.08
SememePSO	0.44	0.64	0.71	0.96	148.20	0.25	0.37	0.70	0.94	230.58
PWWS	0.48	0.69	0.72	0.96	225.27	0.31	0.47	0.70	0.94	226.78
SCPN	0.09	**0.90**	0.29	0.31	11.90	0.09	**0.97**	0.29	0.30	11.97
TextFooler	0.46	0.70	0.70	0.93	106.13	0.29	0.49	0.65	0.88	131.88
MUSS	0.07	0.18	0.65	0.67	2.18	0.08	0.19	0.65	0.69	2.19
RD										
BAE	0.07	0.18	0.41	0.98	313.01	0.18	0.44	0.42	0.98	196.69
BERT-ATTACK	0.18	0.44	0.43	0.96	774.31	0.30	0.69	0.45	0.97	366.14
DeepWordBug	0.16	0.23	**0.70**	**0.99**	232.74	**0.39**	0.56	**0.70**	**0.99**	174.03
Genetic	**0.20**	**0.46**	0.45	0.96	4425.11	0.35	0.79	0.46	0.95	2266.91
SememePSO	0.10	0.21	0.46	0.97	345.89	0.27	0.57	0.49	0.96	233.88
PWWS	0.16	0.38	0.45	0.95	1105.99	0.32	0.75	0.45	0.93	838.83
SCPN	0.01	0.38	0.16	0.10	11.35	0.02	**0.90**	0.15	0.10	11.90
TextFooler	0.16	0.41	0.43	0.91	657.15	0.31	0.70	0.47	0.96	358.37
MUSS	0.02	0.06	0.55	0.72	2.05	0.03	0.1	0.52	0.73	2.10

Table 4.14: The results of adversarial attacks on the **BERT classifier** in four misinformation detection tasks in untargeted and targeted scenario. Evaluation measures include BODEGA score (B.), success score (succ), semantic score (sem), character score (char) and number of queries to the attacked model (Q.). The best score in each task and scenario is in boldface. [Przybyła et al., 2023]

	Untargeted					Targeted				
Method	**B.**	**succ**	**Sem**	**Char**	**Q.**	**B.**	**succ**	**sem**	**char**	**Q.**
HN										
BAE	0.48	0.77	0.64	0.98	489.27	0.45	0.74	0.62	0.98	477.65
BERT-ATTACK	**0.64**	**0.98**	0.66	0.99	487.85	**0.61**	**0.96**	0.65	0.99	565.05
DeepWordBug	0.41	0.53	**0.77**	**1.00**	396.18	0.37	0.47	**0.78**	**1.00**	379.20
Genetic	0.44	0.94	0.48	0.98	2029.31	0.42	0.90	0.47	0.98	2882.19
SememePSO	0.21	0.42	0.50	0.99	313.51	0.14	0.28	0.49	0.99	361.38
PWWS	0.44	0.93	0.48	0.99	2044.96	0.42	0.89	0.48	0.97	1994.95
SCPN	0.00	0.94	0.08	0.02	11.86	0.00	0.95	0.08	0.02	11.83
TextFooler	0.43	0.94	0.47	0.97	543.68	0.41	0.91	0.47	0.96	598.46
MUSS	0.09	0.20	0.67	0.71	2.20	0.07	0.15	0.66	0.71	2.15
PR2										
BAE	0.15	0.23	0.72	0.94	32.94	0.26	0.38	0.71	0.94	38.72
BERT-ATTACK	0.53	0.80	0.72	0.91	61.41	**0.66**	0.94	0.74	0.94	50.14
DeepWordBug	0.29	0.38	**0.79**	**0.96**	27.45	0.56	0.72	**0.81**	**0.96**	35.30
Genetic	**0.54**	**0.88**	0.67	0.89	782.15	0.62	0.94	0.71	0.93	802.20
SememePSO	0.47	0.76	0.68	0.89	85.34	0.60	0.92	0.71	0.92	69.62
PWWS	0.53	0.84	0.69	0.90	130.85	0.63	0.92	0.73	0.94	168.60
SCPN	0.12	0.55	0.39	0.50	11.55	0.20	**0.98**	0.37	0.48	11.98
TextFooler	0.51	0.85	0.67	0.88	52.59	0.63	0.94	0.72	0.92	54.62
MUSS	0.16	0.33	0.70	0.69	2.16	0.08	0.16	0.72	0.69	2.16
FC										
BAE	0.36	0.55	0.69	0.96	77.76	0.32	0.48	0.69	0.96	73.43
BERT-ATTACK	0.60	0.86	0.73	0.95	132.80	**0.59**	0.85	0.73	0.96	123.24
DeepWordBug	0.48	0.58	**0.85**	**0.98**	54.36	0.54	0.64	**0.85**	**0.98**	50.72
Genetic	**0.61**	**0.90**	0.71	0.95	840.99	0.57	0.88	0.69	0.94	1015.44
SememePSO	0.53	0.76	0.72	0.96	112.84	0.46	0.67	0.72	0.96	132.28
PWWS	0.57	0.82	0.73	0.96	221.60	0.50	0.73	0.71	0.95	211.05
SCPN	0.08	0.75	0.29	0.32	11.75	0.11	**1.00**	0.30	0.35	12.00
TextFooler	0.55	0.82	0.71	0.94	98.31	0.50	0.75	0.70	0.94	99.98
MUSS	0.11	0.24	0.66	0.67	2.24	0.11	0.24	0.67	0.68	2.24
RD										
BAE	0.09	0.21	0.43	0.98	312.77	0.27	0.64	0.43	0.98	123.16
BERT-ATTACK	0.29	**0.79**	0.41	0.89	985.52	0.43	0.95	0.46	0.97	130.64
DeepWordBug	0.16	0.24	**0.68**	**0.99**	232.75	**0.62**	0.91	**0.69**	**0.99**	153.61
Genetic	**0.32**	0.71	0.47	0.96	3150.24	0.44	**0.96**	0.48	0.95	1355.52
SememePSO	0.15	0.31	0.48	0.97	314.63	0.32	0.67	0.50	0.97	185.47
PWWS	0.29	0.64	0.47	0.97	1059.07	0.44	0.95	0.48	0.95	742.12
SCPN	0.01	0.55	0.17	0.09	11.53	0.02	0.84	0.15	0.12	11.84
TextFooler	0.24	0.64	0.41	0.87	639.97	0.44	**0.96**	0.48	0.96	184.97
MUSS	0.02	0.06	0.61	0.74	2.05	0.04	0.10	0.58	0.73	2.10

Table 4.15: The results of adversarial attacks on the **BiLSTM** classifier in four misinformation detection tasks in untargeted and targeted scenario. Evaluation measures include BODEGA score (B.), success score (succ), semantic score (sem), character score (char) and number of queries to the attacked model (Q.). The best score in each task and scenario is in boldface. [Przybyła et al., 2023]

Chapter 5

CONCLUSION AND FUTURE WORK

5.1 Summary

The objective of this study was to evaluate the effectiveness of Text Simplification in generating adversarial attacks in the domain of misinformation. The methodology involved selecting a suitable technique, analyzing its simplification capabilities, implementing the technique, and evaluating the attack outcomes. MUSS was chosen for its efficacy given the resource constraints. Despite its effectiveness, not all instances were successful, with some cases lacking simplicity and others experiencing changes that altered the text's connotation, indicating the need for significant improvement.

The attack targeted two victim models, BERT and BiLSTM, within a grey-box scenario with an assumption of unlimited queries. Key observations from the experiments include:

- Some findings from [Przybyła et al., 2023] are consistent with this attack, such as BERT being more robust than BiLSTM. Additionally, these findings also indicate that shorter sentences are more susceptible to such attacks this trend can also be observed in our experiments.

- The most favorable outcomes were obtained through untargeted attacks to BiLSTM model on PR2. Having the best BODEGA score due to high success rate as well as high values for semantic and character score. The brevity facilitated a better simplification as well as an easier model confusion.

- The number of queries required for successful attacks using simplification was notably lower than that required by alternative methodologies.

5.2 Discussion

One major limitation observed with MUSS is the repetitive output for identical inputs. Sending to the model one input text in the same environment conditions returned the same output result. Therefore, we faced limitations in experimenting with the model, as we could not retry to simplify the text until the output was successful. This restricts the diversity of results, limiting the system's success, as it does not allow for exploring different simplifications until satisfactory outcomes are achieved.

Moreover, the inaccuracies in the simplifications frequently result in the omission of vital information or misrepresentation, so the label of the original text might not apply. This issue is particularly pronounced in FC tasks, where the label heavily relies on evidence sentences. If the meaning is not preserved or reversed, then the initial label might not apply and also defeat the main purpose of adversarial examples, which is to create modified inputs where the differences are barely noticeable.

5.3 Future Work

For future research, it is essential to explore the use of Large Language Models (LLMs) for Text Simplification. LLMs, such as GPT-4, have demonstrated superior accuracy and a more nuanced understanding of language compared to earlier models. Implementing these advanced models could potentially enhance the results of our adversarial attacks by generating more varied and contextually accurate simplifications.

Furthermore, LLMs can offer greater flexibility and creativity in generating adversarial examples, which could improve the robustness testing of text classification algorithms.

Moreover, expanding the scope of tasks beyond misinformation detection to include other domains like sentiment analysis, spam detection, and plagiarism identification could offer broader insights into the effectiveness of Text Simplification as an adversarial attack method. Conducting extensive empirical studies across diverse datasets and classification tasks will be crucial in validating the generalizability of the proposed techniques.

By addressing these areas, future research can significantly advance the field of adversarial attacks in NLP, contributing to the development of more robust and resilient text classification systems.

5.4 Conclusions

In conclusion, our research indicates that Text Simplification is a promising method for generating adversarial attacks in the domain of misinformation detection. The use of simplification techniques, such as MUSS, has demonstrated potential for challenging the robustness of text classification algorithms. However, our study also highlights the need for further enhancements and adjustments to improve the overall effectiveness of these adversarial attacks.

While MUSS provided a foundation for our experiments, it exhibited certain limitations, such as repetitive outputs and inaccuracies in preserving the original meaning of the text. These issues underscore the necessity for more advanced simplification models that can generate diverse and contextually accurate adversarial examples.

In summary, while our research confirms the potential of Text Simplification as a method for generating adversarial attacks with few queries needed, it also underscores the need for ongoing development and innovation in this field. By addressing the identified limitations and exploring new methodologies, future research can contribute to the creation of more robust and resilient text classification systems, ultimately enhancing the security and reliability of online platforms.

Bibliography

[Ahmed et al., 2022] Ahmed, S., Hinkelmann, K., and Corradini, F. (2022). Development of fake news model using machine learning through natural language processing. *ArXiv*, abs/2201.07489.

[Akers et al., 2018] Akers, J., Bansal, G., Cadamuro, G., Chen, C., Chen, Q. Z., Lin, L. H., Mulcaire, P., Nandakumar, R., Rockett, M., Simko, L., Toman, J., Wu, T., Zeng, E., Zorn, B., and Roesner, F. (2018). Technology-enabled disinformation: Summary, lessons, and recommendations.

[Al-Thanyyan and Azmi, 2021] Al-Thanyyan, S. and Azmi, A. M. (2021). Automated text simplification. *ACM Computing Surveys (CSUR)*, 54:1 – 36.

[Alva-Manchego et al., 2020] Alva-Manchego, F., Martin, L., Bordes, A., Scarton, C., Sagot, B., and Specia, L. (2020). ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

[Barzilay and McKeown, 2001] Barzilay, R. and McKeown, K. R. (2001). Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 50–57, Toulouse, France. Association for Computational Linguistics.

[Bott et al., 2012] Bott, S., Rello, L., Drndarevic, B., and Saggion, H. (2012). Can Spanish be simpler? LexSiS: Lexical simplification for Spanish. In Kay, M. and Boitet, C., editors, *Proceedings of COLING 2012*, pages 357–374, Mumbai, India. The COLING 2012 Organizing Committee.

[Boucher et al., 2021] Boucher, N. P., Shumailov, I., Anderson, R., and Papernot, N. (2021). Bad characters: Imperceptible nlp attacks. *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1987–2004.

[Carter et al., 2021] Carter, M., Tsikerdekis, M., and Zeadally, S. (2021). Approaches for fake content detection: Strengths and weaknesses to adversarial attacks. *IEEE Internet Computing*, 25(2):73–83.

[Chi et al., 2023] Chi, A., Chen, L.-K., Chang, Y.-C., Lee, S.-H., and Chang, J. S. (2023). Learning to Paraphrase Sentences to Different Complexity Levels. *Transactions of the Association for Computational Linguistics*, 11:1332–1354.

[Da San Martino et al., 2020] Da San Martino, G., Barrón-Cedeño, A., Wachsmuth, H., Petrov, R., and Nakov, P. (2020). SemEval-2020 task 11: Detection of propaganda techniques in news articles. In Herbelot, A., Zhu, X., Palmer, A., Schneider, N., May, J., and Shutova, E., editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

[Documentation, 2024] Documentation, O. (2024). Example 3. Accessed: 2024-06-07.

[Dwivedi and Dwivedi, 2022] Dwivedi, A. K. and Dwivedi, M. (2022). A study on the role of machine learning in natural language processing. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*.

[Feng et al., 2023] Feng, Y., Qiang, J., Li, Y., Yuan, Y., and Zhu, Y. (2023). Sentence simplification via large language models. *ArXiv*, abs/2302.11957.

[Ferrés et al., 2016] Ferrés, D., Marimon, M., Saggion, H., and AbuRa'ed, A. G. T. (2016). Yats: Yet another text simplifier. In *International Conference on Applications of Natural Language to Data Bases*.

[Flores and Hao, 2022] Flores, L. J. Y. and Hao, S. (2022). An adversarial benchmark for fake news detection models. *ArXiv*, abs/2201.00912.

[Gong et al., 2018] Gong, Z., Wang, W., Li, B., Song, D. X., and Ku, W.-S. (2018). Adversarial texts with gradient methods. *ArXiv*, abs/1801.07175.

[Guo et al., 2021] Guo, C., Sablayrolles, A., J'egou, H., and Kiela, D. (2021). Gradient-based adversarial attacks against text transformers. In *Conference on Empirical Methods in Natural Language Processing*.

[Han et al., 2019] Han, S., Gao, J., and Ciravegna, F. (2019). Neural language model based training data augmentation for weakly supervised early rumor detection. *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 105–112.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

[Jiang et al., 2020] Jiang, C., Maddela, M., Lan, W., Zhong, Y., and Xu, W. (2020). Neural CRF model for sentence alignment in text simplification. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Online. Association for Computational Linguistics.

[Kew et al., 2023] Kew, T., Chi, A., Vásquez-Rodríguez, L., Agrawal, S., Aumiller, D., Alva-Manchego, F., and Shardlow, M. (2023). BLESS: Benchmarking large language models on sentence simplification. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309, Singapore. Association for Computational Linguistics.

[Kiesel et al., 2019] Kiesel, J., Mestre, M., Shukla, R., Vincent, E., Adineh, P., Corney, D., Stein, B., and Potthast, M. (2019). SemEval-2019 task 4: Hyperpartisan news detection. In May, J., Shutova, E., Herbelot, A., Zhu, X., Apidianaki, M., and Mohammad, S. M., editors, *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

[Levenshtein, 1965] Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710.

[Li et al., 2020] Li, L., Ma, R., Guo, Q., Xue, X., and Qiu, X. (2020). BERT-ATTACK: Adversarial attack against BERT using BERT. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics.

[Liang et al., 2017] Liang, B., Li, H., Su, M., Bian, P., Li, X., and Shi, W. (2017). Deep text classification can be fooled. *ArXiv*, abs/1704.08006.

[Lu et al., 2021] Lu, X., Qiang, J., Li, Y., Yuan, Y., and Zhu, Y. (2021). An unsupervised method for building sentence simplification corpora in multiple languages. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors,

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 227–237, Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Martin et al., 2021] Martin, L., Fan, A., de la Clergerie, É., Bordes, A., and Sagot, B. (2021). Muss: Multilingual unsupervised sentence simplification by mining paraphrases. *arXiv preprint arXiv:2005.00352*.

[Naskar et al., 2019] Naskar, S., Saha, S., and Mukherjee, S. (2019). Text embellishment using attention based encoder-decoder model. In Burtenshaw, B. and Manjavacas, E., editors, *Proceedings of the 4th Workshop on Computational Creativity in Language Generation*, pages 28–38, Tokyo, Japan. Association for Computational Linguistics.

[Nishihara et al., 2019] Nishihara, D., Kajiwara, T., and Arase, Y. (2019). Controllable text simplification with lexical constraint loss. In Alva-Manchego, F., Choi, E., and Khashabi, D., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 260–266, Florence, Italy. Association for Computational Linguistics.

[Oseni et al., 2021] Oseni, A., Moustafa, N., Janicke, H., Liu, P., Tari, Z., and Vasilakos, A. V. (2021). Security and privacy for artificial intelligence: Opportunities and challenges. *ArXiv*, abs/2102.04661.

[Paetzold and Specia, 2016] Paetzold, G. and Specia, L. (2016). SemEval 2016 task 11: Complex word identification. In Bethard, S., Carpuat, M., Cer, D., Jurgens, D., Nakov, P., and Zesch, T., editors, *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

[Przybyla, 2022] Przybyla, P. (2022). Lambo: Layered approach to multi-level boundary identification.

[Przybyła et al., 2023] Przybyła, P., Shvets, A. V., and Saggion, H. (2023). Verifying the robustness of automatic credibility assessment.

[Ren et al., 2019] Ren, S., Deng, Y., He, K., and Che, W. (2019). Generating natural language adversarial examples through probability weighted word saliency. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy. Association for Computational Linguistics.

[Saggion, 2017] Saggion, H. (2017). Book review: Automatic text simplification by horacio saggion. *Computational Linguistics*, 44:659–661.

[Sellam et al., 2020] Sellam, T., Das, D., and Parikh, A. (2020). BLEURT: Learning robust metrics for text generation. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

[Tajner, 2021] Tajner, S. (2021). Automatic text simplification for social good: Progress and challenges. In *Findings*.

[Thorne et al., 2018] Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., and Mittal, A. (2018). The fact extraction and VERification (FEVER) shared task. In Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., and Mittal, A., editors, *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

[Tucker et al., 2018] Tucker, J. A., Guess, A. M., Barberá, P., Vaccari, C., Siegel, A. A., Sanovich, S., Stukal, D. K., and Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature.

[Wubben et al., 2012] Wubben, S., van den Bosch, A., and Krahmer, E. (2012). Sentence simplification by monolingual machine translation. In Li, H., Lin, C.-Y., Osborne, M., Lee, G. G., and Park, J. C., editors, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024, Jeju Island, Korea. Association for Computational Linguistics.

[Xu et al., 2015] Xu, W., Callison-Burch, C., and Napoles, C. (2015). Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

[Zhang et al., 2020] Zhang, W. E., Sheng, Q. Z., Alhazmi, A., and 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 and Lapata, 2017] Zhang, X. and Lapata, M. (2017). Sentence simplification with deep reinforcement learning. In *Conference on Empirical Methods in Natural Language Processing*.

[Zhang et al., 2023] Zhang, Y., Li, Y., Li, Y., and Guo, Z. (2023). A review of adversarial attacks in computer vision. *arXiv preprint arXiv:2308.07673*.

[Zhao et al., 2018] Zhao, S., Meng, R., He, D., Saptono, A., and Parmanto, B. (2018). Integrating transformer and paraphrase rules for sentence simplification. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3164–3173, Brussels, Belgium. Association for Computational Linguistics.

[Zhu et al., 2010] Zhu, Z., Bernhard, D., and Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. In Huang, C.-R. and Jurafsky, D., editors, *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China. Coling 2010 Organizing Committee.