UKRAINIAN CATHOLIC UNIVERSITY

MASTER THESIS

Computational Detection and Analysis of Manipulation Techniques in News Channels on Telegram in Ukraine

Author:
Nataliia VOLKOVA

Supervisor:
Oleksii IGNATENKO

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2024

Declaration of Authorship

I, Nataliia VOLKOVA, declare that this thesis titled, “Computational Detection and Analysis of Manipulation Techniques in News Channels on Telegram in Ukraine ” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date: 17.05.2024

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Master of Science

Computational Detection and Analysis of Manipulation Techniques in News Channels on Telegram in Ukraine

by Nataliia VOLKOVA

Abstract

In this research, we aim to identify specific instances of manipulation techniques such as doubts, black-and-white fallacy, appeal to fear, and loaded language and their granularity in news text from social media. To achieve this, we developed our own manually annotated corpus with 1,877 posts from Ukrainian news channels on Telegram, all in the Ukrainian language, consisting of 3,472 manipulation techniques. Each annotation includes the fragment or span where a manipulation technique is detected, along with the corresponding technique from a set of selected techniques. We then trained a pre-trained BERT model to recognize these spans and their associated manipulation techniques. Additionally, we generated syntactic data to enhance the model's performance.

Acknowledgements

I appreciate the Mantis Analytics team for their support and guidance. I would like to express my gratitude to Martino et al. for providing a well-developed methodological approach to detecting fine-grained manipulation in texts. I am thankful to my mentor, professors, fellow students, and husband for their unwavering support throughout my studies. I appreciate the volunteers who dedicated their time to annotating the texts. My thanks also extend to the Ukrainian and international NLP community; their contributions, such as transformer models on Hugging Face for the Ukrainian language and tools like Grammarly and ChatGPT, have significantly improved my research and report. I'm grateful to LabelBox for their open position on supporting young scientists.

Lastly, I am deeply grateful to the Armed Forces of Ukraine for making this research possible.

Contents

Declaration of Authorship ii

Abstract iii

Acknowledgements iv

1 Introduction and Motivation 1

2 Definition and related work 3
2.1 Definition and Foundations 3
2.2 Related Computational Work 5

3 Research Gap and Problem 7
3.1 Research Gap 7
3.2 The Problem Formulation 7

4 Research Setting and Approach to Solution 10
4.1 Approach to Solution 10
4.2 Raw data collection 10
4.3 Preparing and sampling data for annotation 11
4.4 Data Annotation 12
4.5 Corpus statistics 13
4.6 Tasks 14
4.7 Baselines 15
4.8 Evaluation Measures 15

5 Experiments and Discussion 16
5.1 Model 16
5.2 Synthetic text generation for upsampling 17
5.3 Experiments and Results 18
5.3.1 Span Identification task 18
Error Analysis 19
5.3.2 Technique Classification task 19
Error Analysis 22

6 Limitations and Ethical Implications 24

7 Conclusion 25

A Report on Volunteer Engagement 26

List of Figures

3.1 Selected manipulation techniques for the work. 8

4.1 The distribution of the labeled posts. 13
4.2 The examples of posts with manipulation techniques (shown in blue). The left column shows the character spans and the name of the manipulation technique(s) belonging to that span. 15
4.3 The example of span merging. 15

5.1 The frequency of the manipulation techniques in train and test datasets. 16
5.2 Splitting texts to sequences with binary tags for SI task. 17
5.3 Splitting texts to sequences with labels tags for TC task. 17
5.4 Distribution of span lengths. 19
5.5 The confusion matrices for both models. 22
5.6 Example of prediction only one class for the same span. 23
5.7 Example of prediction of the same class for partially overlapping spans. 23

List of Tables

4.1 Raw data and data for annotation statistics. 11
4.2 Corpus statistics: each manipulation technique's frequency, average length in terms of words and characters. 14

5.1 Statistics of the training and test subsets of the dataset. Lengths are calculated in words and characters. 17
5.2 The evaluation of the results for the Span Identification task. 18
5.3 The evaluation of the results for the Technique Classification task for 10 classes. 20
5.4 The classification reports for BERT model (initial data in test sub-set). . 21
5.5 The classification report for BERT model with upsampling in train and test sub-sets with no weights 23

Chapter 1

Introduction and Motivation

The widespread dissemination of fake news poses a significant threat to individuals and society, disrupting the authenticity of the news ecosystem. This global impact has been evident during critical events, such as the 2016 U.S. Presidential election and the U.K. Brexit referendum [1], the COVID-19 pandemic [2], the Crimea annexation [1], and the fully escalated Russian war against Ukraine.

Fake news, intentionally false information, is crafted to potentially mislead readers by distorting facts, taking information out of context, or presenting partial truths. While traditionally associated with entirely false information, recent research has shown that fake news is rarely devoid of truth, and even factual information can be manipulated to convey a misleading narrative, emphasizing the complexity of evaluating news authenticity.

The Russian war against Ukraine underscores the urgent need for an effective tool to detect fake news in the social media landscape. Social media platforms, especially Telegram, the most popular among others, witnessed a proliferation of numerous channels disseminating partial truths, pro-Russian narratives, and outright fabrications to the Ukrainian audience. The contemporary strategy of Russia's "information warfare" deliberately seeks to sow confusion, polarize opinions, instill distrust, and construct a fundamentally distorted worldview [3]. These tactics serve Russia's overarching goals, allowing the state to address its political agenda directly to users as an obvious objective solution.

Automatically detecting fake news on social media in Ukraine is an urgent and essential problem. Therefore, we developed our own approach to a solution considering the unique language features, Ukrainian context, and purposefully designed "information warfare" against Ukraine by Russia. The best direction for our purpose was style-based approaches in fake news detection, focusing on identifying manipulative techniques embedded in the writing style of news content. We took four manipulation techniques and their granularity used in news writing that could be spotted immediately and do not require supporting information from external resources. They suit the best for Russia's information operations: doubts, black-and-white fallacy, appeal to fear, and loaded language. Either deliberately or unintentionally, the news having these manipulation techniques leads to the fragmentation of Ukrainian society and helps Russia to achieve its invasive ambitions. The computation detection and analysis of manipulative news in Ukrainian social media could lead to increased public resistance against Russian propaganda. It could empower fact-checkers, journalists, public figures, and bloggers to catch manipulated news early, preventing its dissemination. Telegram users can better understand the underlying motives and strategies behind the stylistic aspects of manipulation texts, enabling them to make informed decisions and be better equipped to navigate the challenges posed by propaganda and disinformation campaigns.

This research identifies specific instances of propaganda techniques in news text from social media. To achieve this, we developed our own corpus, consisting of manually annotated posts from Ukrainian news channels on Telegram, all in the Ukrainian language. Each annotation includes the fragment or span where a manipulation technique is detected, along with the corresponding technique from a set of 10 selected techniques. We then trained a pre-trained BERT model to recognize these spans and their associated manipulation techniques. Additionally, we generated syntactic data to enhance the model's performance.

The report's structure is as follows: Section 2 explores the foundations and definitions of fake news from social, psychological, and political perspectives. Then, we examine the relevant works. Section 3 highlights the research gaps in these studies and formulates the research problem for our work. In Section 4, we describe the research setting and our approach to a solution. Section 5 presents the experiments and discusses the results. Finally, Section 6 addresses limitations and ethical considerations.

Chapter 2

Definition and related work

The identification of fake news in the media text is a pressing subject explored across various disciplines. To encompass the collective efforts of the multidisciplinary research community and enhance our comprehension of the concept, framework, and propaganda techniques, we drew upon literature from both Computer Science and Social Science fields, including Sociology, Psychology, Linguistics, and Politics.

The approach to gathering pivotal scientific papers included searching widely recognized scholarly platforms such as Semantic Scholar and Scopus by key phrases and using the ResearchRabbit recommendation platform to enhance our collection of articles, prioritizing those with a higher number of citations. Among others, our selection criteria encompassed review and research articles from 2014 onwards.

2.1 Definition and Foundations

Before the examination of the literature analysis on "fake news" detection to ensure clarity and coherence, we rely on the following definition for this term: fake news is a news article that is intentionally and verifiably false by fact or perception, designed to potentially mislead readers [5].

It is the narrower category of disinformation focused on the media landscape. This may involve distorting facts, taking information out of context, or presenting partial truths to manipulate or mislead the audience. The objective is to construct a false narrative, often sensationalized or biased, with the aim of influencing opinions or actions.

Considering the concept of fake news through the psychological, social, and political perspectives, as well as studying the features inherent in social networks that facilitate the creation and spread of fake news, the following foundations can be identified:

Psychological Foundations of Fake News. The discernment of whether news is true extends beyond rational considerations, as it is also shaped by various psychological factors that could be hidden or underestimated by humans: (1) Social credibility: individuals are inclined to view a source as credible if others deem it credible [5]. (2) Confirmation Bias: individuals prefer to receive information that confirms their preexisting beliefs [5]. (3) Frequency heuristic: individuals may naturally trust information they come across frequently, even if it happens to be fake news [5].

Furthermore, correcting fake news is exceptionally challenging once it has taken root. Psychological studies suggest that fact-checkers efforts to debunk false information do not fully negate the impact of the initial exposure to the news. Additionally, certain ideological groups might perceive such corrective actions as an attempt to conceal the truth rather than an earnest effort to provide accurate information.

Social Foundations of Fake News. Social dynamics, encompassing interactions within groups and adherence to social norms, play a crucial role in the spreading of fake news. Two key social concepts contributing to this understanding are (1) Social Identity Theory, which suggests individuals categorize themselves and others based on shared characteristics, and (2) Normative Influence Theory, exploring how individuals conform to gain approval or avoid disapproval [5]. Joining certain groups, reactions to posts, and the number of favorable or unfavorable reviews on social media can affect the perception of the credibility of the news received. The power of fake news lies in its ability to manipulate these social tendencies and capitalize on the human need for acceptance and approval within social groups.

Political Foundation of Fake News. When fake news is intentionally generated and disseminated with a political agenda to influence public opinion or manipulate perceptions, it can be considered a form of propaganda. This type of politically motivated fake news can be identified by its persuasive function, the use of faulty reasoning, emotional appeals, and blend of truths and falsehoods to represent a specific agenda [6]. Analyzing Russia's strategy of information propaganda, C. Paul and M. Matthews [3] came to the conclusion that it contradicts the previously existing strategy of authoritarian regimes that focused on the ideological-oriented view of the situation and news consistency, avoiding contradictions. Russia's contemporary "information warfare" deliberately seeks to sow confusion, polarize opinions, instill distrust, and construct a fundamentally distorted worldview. They abuse the weaknesses of social networks for propaganda, such as the rapid dissemination of posts, the ability to boost reactions to posts that signal the popularity of the news, and the anonymity of channel authors, which makes it possible to camouflage channels as "local" news ones [1], and etc. As the pro-Russian propagandist channels operate without the need for fact-checking, claim verification, or following the consistency in the storyline, it enables them to be highly responsive, often being the first to broadcast "news" about events, thereby shaping initial impressions. These channels frequently propagate disinformation, rumors, and conspiracy theories and launch attacks on Ukrainian mainstream media, the Ukrainian government, and the Ukrainian army to evoke distrust, confusion, and fragmentation in society.

Psychological and social foundations lead to the hypothesis that checking and revealing the true facts in fake news is not as effective, as fact-checking is time-consuming, and the dissemination of research on truth disclosure may not occur as frequently as the circulation of fake news. Besides, readers may align fake news with their preexisting beliefs, trust their friends' opinions, or the first emotional impact remains unconscious, reinforcing their confidence in fake news.

Therefore, in addressing fake news, our aim is not to focus on verifying the truthfulness of the news but rather detecting the linguistic tactics deliberately employed to distort the reader's perception and mislead them. Moreover, Russia's contemporary "information warfare" strategy, as we have discussed above, does not always present false news or one ideological idea but rather uses a specific writing style that raises doubts and prejudices in the audience. The style-based approach may be challenging as it demands a deep understanding of the language used in news articles, the intent of the authors, and familiarity with a wide range of manipulation techniques.

2.2 Related Computational Work

The significance of identifying fake news or propaganda within news content has received increased attention, prompting research from various perspectives. Some of the existing research has focused on the document level fake news detection only [6] [7], while some are on the span level, which was first presented by Da San Martino et al. [8].

Rashkin et al. [6] developed the TSHP-17 corpus, which uses document-level annotation with four classes: trusted, satire, hoax, and propaganda. Their approach involved a linguistic analysis employing stylistic lexicons. The researchers trained various models, including Maximum Entropy (MaxEnt), Naive Bayes, and the LSTM model, commonly used for text categorization at that time. The LSTM model demonstrated the most promising results in their study.

Barrón -Cedeno et al. [7] developed the QProp corpus for binary propaganda detection task (propaganda vs. non-propaganda). It includes news from traditional English outlets and annotations on the article level. Their approach considered various representations, ranging from the writing style and readability level to the presence of specific keywords. Results from their experiments demonstrated that representations emphasizing writing style and text complexity outperformed traditional word n-grams, which primarily focus on topics, in effectively identifying propaganda.

Da San Martino et al. [8] introduced a fine-grained propaganda analysis, creating a PTC corpus of news articles from traditional English outlets annotated on the span level with 18 propaganda techniques that do not require supporting information from external resources. This corpus was used in two shared tasks: one at SemEval-2020 [9] and another at NLP4IF-2019 [10]. Since manipulation is conveyed through several techniques, such detection allows for deeper analysis at the paragraph and the sentence level that goes beyond a single document-level judgment described above. Besides, it gives transparency to the decision.

They are focused on two sub-tasks: (i) binary classification given a sentence in an article, predict whether any of the 18 techniques have been used in it (Span Identification); (ii) multi-label multi-class classification and span detection task – given a raw text, identify both the specific text fragments where a manipulation technique is being used as well as the type of technique (Technique Classification).

Best approaches based on [9], [10] that were successfully applied for both tasks: transformers (BERT, Grover, RoBERTa, GPT-2, XLM), embeddings (ELMo, GloVe), RNN (Bi-LSTM), adding additional context, feature engineering, the ensemble with statistical ML models (Logistical Regression, XGBoost), and unsupervised tuning. Besides that, considering multiple features of readability, sentiment, and emotions was beneficial [11]. For example, tackled the Load Language manipulation technique with emotion features (sadness, joy, fear, disgust, and anger) using IBM Watson NLU API.

The research work by Seunghak Yu et al. [12] offered interpretability of the data, which may lead to better-performing results of the existing solution. They experimented with semantic and structural information related to propaganda techniques like the relative position of the sentence, topic similarity, and stance with respect to the title, the sentiment of the sentence, etc.

The fine-grained propaganda detection approach with 18 classes was also applied to the social media data, particularly on Twitter comments in English by P. Vijayaraghavan and S.Vosoughi [13]. They weakly annotate data on the span level and propose additionally extracting different aspects of the input text, including the context, entities, their relationships, and external knowledge, and use that as additional layers for classification tasks.

Various approaches of fine-grained propaganda techniques to fake news detection exhibit distinct advantages and drawbacks concerning computational efficiency, suitability for processing lengthy text, contextual comprehension, and interpretability. Each approach exhibited varying degrees of effectiveness in detecting all 18 manipulation techniques, with some methods proving more adept than others at identifying specific forms of manipulation. Currently, to the best of our knowledge, the SOTA for both tasks is shown by bert-based transformer models.

Chapter 3

Research Gap and Problem

3.1 Research Gap

A promising avenue for style-based fake news detection is identifying fine-grained propaganda techniques. This approach aligns with the contemporary fake news strategy of blending true and false statements or manipulating attitudes through writing style. Besides, manipulation techniques, well-documented in the literature, serve as the foundation for fake news, making them a logical focus for the research project. However, the existing approaches of fine-grained manipulation techniques detection have certain limitations:

(1) Having the ground truth data is essential for this task. Previous research projects based on corpora in English [8] and Arabic [14] are biased toward the US and Arab culture and political landscape due to the origin of the data. To our knowledge, no annotated dataset is fully ready for training for such a task in Ukrainian.

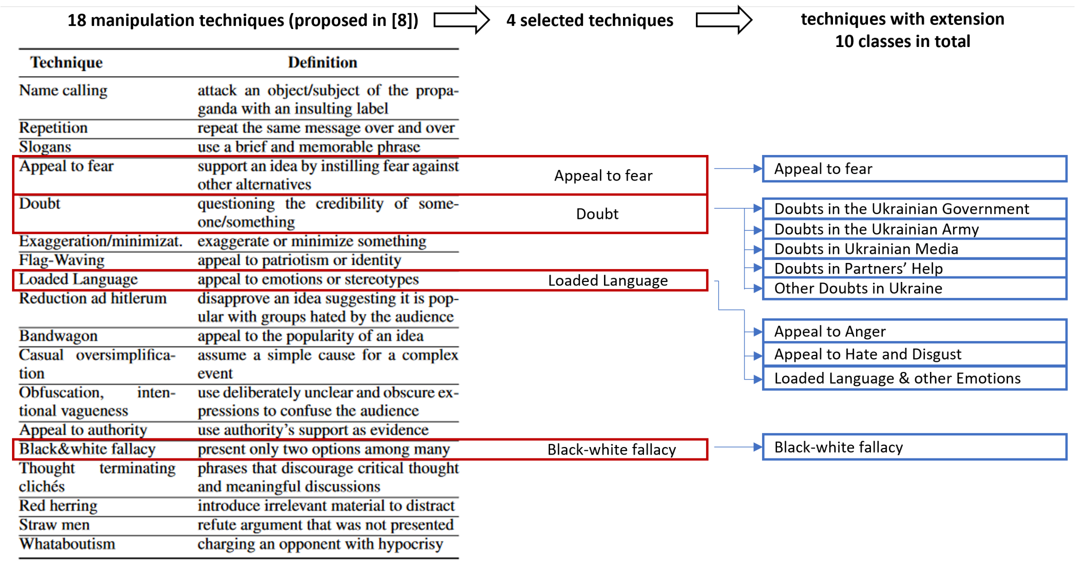
(2) Most existing research projects relied on the corpora containing news text from newspapers, which are structured, without mistakes, new words, slang, dialect, or a mix of different languages. Additional efforts may be required to address challenges from social media data.

(3) The task of identifying fine-grained propaganda techniques has been tackled in English and Arabic but not in Ukrainian. The multi-lingual content, hidden patterns, and distinct features of Ukrainian texts necessitate additional efforts.

(4) Although detecting manipulation techniques at the fragment level provides explainability and transparency in decision-making, it falls short of explaining the intent aspect of fake news. Considering Russia's "information warfare" against Ukraine, understanding the aim of manipulation techniques and their targets can be valuable features for manipulation detection models.

3.2 The Problem Formulation

To fill these gaps, we thoroughly studied some manipulation techniques that are being used in Ukrainian news on social media. Based on the research by [1] and [3], who have studied Russian propaganda strategy and tactics, we have selected four manipulation techniques (Appeal to Fear, Doubts, Loaded Language, and White-Black fallacies) out of 18 that instill distrust, confusion, apathy, and polarize opinion in society - all of which help Russia better implement its ideological narratives. Taking into account the Ukrainian context and the Russo-Ukrainian war, we decided that it was not enough for us to work with only 4 manipulative techniques because they are very general. Furthermore, for greater transparency, we have also extended some techniques like Doubts and Loaded Language to understand the purpose and tactics behind the manipulation, which resulted in 10 classes in total. Figure 3.1 demonstrates the process of the techniques' selection.


18 manipulation techniques (proposed in [8])
4 selected techniques
techniques with extension
10 classes in total
Appeal to fear
Doubt
Loaded Language
Black-white fallacy
Appeal to fear
Doubts in the Ukrainian Government
Doubts in the Ukrainian Army
Doubts in Ukrainian Media
Doubts in Partners' Help
Other Doubts in Ukraine
Appeal to Anger
Appeal to Hate and Disgust
Loaded Language & other Emotions
Black-white fallacy

Technique	Definition
Name calling	attack an object/subject of the propaganda with an insulting label
Repetition	repeat the same message over and over
Slogans	use a brief and memorable phrase
Appeal to fear	support an idea by instilling fear against other alternatives
Doubt	questioning the credibility of someone/something
Exaggeration/minimizat.	exaggerate or minimize something
Flag-Waving	appeal to patriotism or identity
Loaded Language	appeal to emotions or stereotypes
Reduction ad hitlerum	disapprove an idea suggesting it is popular with groups hated by the audience
Bandwagon	appeal to the popularity of an idea
Casual oversimplification	assume a simple cause for a complex event
Obfuscation, intentional vagueness	use deliberately unclear and obscure expressions to confuse the audience
Appeal to authority	use authority's support as evidence
Black&white fallacy	present only two options among many
Thought terminating clichés	phrases that discourage critical thought and meaningful discussions
Red herring	introduce irrelevant material to distract
Straw men	refute argument that was not presented
Whataboutism	charging an opponent with hypocrisy

FIGURE 3.1: Selected manipulation techniques for the work.

The definitions of the selected manipulation techniques are following (the manipulation technique is highlighted in blue color in examples):

(1) **Doubts in the Ukrainian Government**: questioning the credibility and efficiency of Ukrainian government. For example: "У розтраті майже 4,5 млн бюджетних коштів підозрюють керівника Ніжинського КП. Скільки ж можна так нагло грітися на всьому?"

(2) **Doubts in the Ukrainian Army**: questioning the credibility anf efficiency of Ukrainian army. For example: "Керівництво армії намагається заспокоїти громадськість заявами про боротьбу з корупцією, але реальність показує, що це залишається лише на словах."

(3) **Doubts in Ukrainian National/Credible Media**: questioning the credibility and efficiency of Ukrainian media. For example: "Мешканці обурені незаконною забудовою у своєму районі, але чи можна вірити тому, що ЗМІ не приховують правду про те, як це сталося та хто з цього скористався?

(4) **Doubts in Partners' Help**: question the effectiveness of actions and honesty of partners who help Ukraine. For example: "Ігнорування кризи: Захід знову відмовляється взяти на себе відповідальність. Відмова від підтримки викликає занепокоєння та сумніви у вірність партнера."

(5) **Other Doubts in Ukraine**: other doubts about the existence of Ukraine, the actions of Ukrainians as a people, or the actions of a large group of Ukrainians, refugees abroad, etc. For example: "Українські біженці знову отримують дотації в країні перебування, і, як завжди, на них подорожують по Європі, беззастережно продовжуючи своє життя".

(6) **Black-and-white fallacy**: presenting two alternative options as the only possibilities, when in fact more possibilities exist. For example: "Або ми припинимо толерувати булінг та впровадимо нульову терпимість до цього виду насильства, або ми засудимо себе до безкінечного циклу страждань та психологічних травм!"

(7) **Appeal to Fear**: seeking to build support for an idea by instilling anxiety and/or panic in the population towards an alternative, possibly based on preconceived judgments [8]. For example: "ПОТУЖНІ магнітні бурі накриють Землю у жовтні".

(8) **Appeal to Anger**: causing anger by emphasizing the negative aspects of a situation in order to evoke an emotional reaction. For example: "Скільки нирок треба продати, щоб купити житло в Одесі?"

(9) **Appeal to Hate and Disgust**: causing hate, intense dislike, or disgust by using strong negative emotions, stereotypical phrases, or words that humiliate a person or group of people or a certain idea they support. For example: "На чоловіка, який побив транс-підора військового у Львові, наклали штраф у 17 000 гривень. Добре, що хоч не посадили, бо від про-ліваційських копів зрадників можна чекати що завгодно."

(10) **Loaded Language and Appeal to other Emotions**: other emotional appeals and phrases with strong connotations to persuade another by evoking feelings rather than providing arguments and evidence. For example: "У Львові затримали неадекватну жінку, в якої завершився термін дії посвідки на законне перебування в Україні. Щоб її не депортували, вона намагалась укласти шлюб. План-капка не спрацював і росіянку видворили за межі України з забороною на 8 років повертатись назад."

In our work, we have developed an approach to detecting fragments in social media news text with manipulative techniques and defining these techniques, taking into account all issues related to NLP in the Ukrainian language and data collection. Taking into account the Ukrainian context, we worked with 10 manipulation techniques, which include extensions of Doubts and Loaded Language techniques for better interpretation.

Chapter 4

Research Setting and Approach to Solution

4.1 Approach to Solution

We framed our task to detect selected manipulation techniques in the news posts. We have two sub-tasks: to identify the span of the manipulation technique and, in a given span, classify the manipulation technique, considering the approach provided by Martino et al [8].

To implement the solution, we prepared the dataset following three steps: 1) collecting raw data; 2) preparing and sampling data for annotation; and 3) the annotation process.

4.2 Raw data collection

As our objective is to analyze Ukrainian news text from social media to detect manipulation techniques, we require a collection of news from the primary social media source of news consumption by Ukrainians. According to a USAID-Internews research study conducted in November 2023[1], 76% of Ukrainians consider social networks as one of the sources for news consumption. When examining individual platforms, 72% of them obtain news from Telegram, followed by Facebook at 19%, Viber at 15%, and Instagram at 10%. Therefore, we selected Telegram as the data source for our research project.

Channels on Telegram can be divided by country of registration and specific categories such as News, Technology, Politics, Business, etc. (Telegram does not disclose the method of such division). Consequently, we collected data from "News and media" public channels on Telegram in Ukraine (except for the air alert and radar monitor channels) with more than 10,000 subscribers over the last three months of 2023. The list of channels was scraped on the TGStat[2] website. The data was gathered with the help of Mantis Analytics[3].

We cleaned up the raw data: we unified the names of the channels that had been renamed and removed posts about missile attacks, advertising, or those where the channel name was unknown. We also removed the posts with less than 4 words. After cleaning, our dataset comprises 965 channels with around 1,5 million posts. Among them, there are 60 pro-Russian channels with around 129 thousand posts. The list of pro-Russian channels was compiled based on research of propaganda

[1] https://internews.in.ua/wp-content/uploads/2023/10/USAID-Internews-Media-Survey-2023-EN.pdf

[2] uk.tgstat.com

[3] mantisanalytics.com

Telegram channels by the fact-checking resource Chesno[4], Center for Countering Disinformation[5], and the online media Detector.media[6].

72% of the data is in Ukrainian language, 27% is in Russian, and 1% in other languages. The statistics are presented in Table 4.1.

In our work, we chose Ukrainian as the language for the study, among other things, because of the prevalence of this language in our raw dataset. Therefore, we only worked with posts that the *langdetect* library identified as written in Ukrainian. We used this sub-set to select the texts for labeling. In Table 4.1 are presented statistics on the number of all channels, pro-Russian channels, posts, and language in the Ukrainian dataset, and the selected data for annotation.

	N channels	**N posts**	**N posts in pro-Russian channels**	**N posts in pro-Russian channels**	**Language**
Raw data	965	1,58 mil	60	128,722	72% Ukrainian, 27% Russian, 1% Other
Raw data in Ukrainian	884	1,15 mil	37	45,899	Ukrainian
Data for annotation in Ukrainian	313	7,394	34	6,394	Ukrainian

TABLE 4.1: Raw data and data for annotation statistics.

In the raw dataset with Ukrainian posts only, there are around 1,1 million posts from 884 channels. Among them, 37 pro-Russian with around 46 thousand posts. The distribution of posts across channels in the subset is uneven. The channels “Бабуся Світуся Новини 24/7”, “СВІДОК НОВИНИ УКРАЇНИ ТА СВІТУ”, “1NEWS”, “УКРАЇНА НАЖИВО" and “ВІЧЕ НСН 2014-2024” have the highest number of messages, exceeding 10,000 each. The distribution of subscribers per channel is right-skewed, with a long tail. Only five channels have over a million subscribers: “Труха Україна” (2,616,144 sub.), “Николаевский Ванёк” (1,943,429 sub.), “Реальна війна Україна” (1,346,813 sub.), “Всевидящее ОКО Украина Новости” (1,134,514 sub.), “Реальний Київ Україна” (1,122,447 sub.). A detailed analysis is presented here [4].

4.3 Preparing and sampling data for annotation

We annotated 7,394 news posts with 10 manipulation techniques (techniques with the extension for Doubts and Loaded Language techniques) on the span level to obtain ground-truth data.

Posts for annotation were selected from the raw data in Ukrainian only with a predominant emphasis on posts from pro-Russian channels: 90% of posts are from pro-Russian channels and 10% from other channels. This choice ensured a higher likelihood of containing manipulation fragments, enhancing the dataset’s relevance

[4] chesno.org

[5] cpd.gov.ua

[6] detector.media

for detecting manipulation techniques. Additionally, focusing on pro-Russian channels gave a greater variety of specific propaganda words, phrases, or sentence structures.

Posts were randomly selected in proportion to the number of posts in the channel in sub-sets pro-Russian or otherwise from posts that were not full duplicates. The duplicates were removed by the function *pandas.DataFrame.duplicated*.

In the result, the annotation dataset contains 7,394 posts from 313 channels, among them 6,394 posts from 34 pro-Russian channels (Table 4.1).

4.4 Data Annotation

To ensure accurate and consistent annotation, detailed instructions were written covering specific requirements for each manipulation technique. These instructions outline the definition of each technique, inclusive and exclusive criteria, commonly employed narratives and associated lexicons. Moreover, it is important to highlight the framework of the manipulation technique used in our work, and some rules and exceptions we made for certain types of posts, phrases, or symbols.

(1) We proceeded from the assumption that news posts should be objective and contain facts and arguments, but not emotional appeals. We do not verify the truthfulness of the news in our work. We evaluate only the text of the news (words and style of writing) that is provided.

(2) Our work is aimed at identifying manipulations that poison Ukrainian society. Therefore, most manipulative techniques reveal manipulations only in relation to Ukrainian society and its groups.

(3) News aimed at glorifying and supporting patriotism, words such as "Слава Героям! Слава Україні! Все буде Україна. Разом до Перемоги!" and condolences were not considered to be a manipulation with appealing to emotions.

(4) Quotes from people in news texts were also annotated if they contained manipulation, although this is a subjective opinion, and the news may contain a contradiction.

(5) Advertisements and messages about the missile threat were marked as not containing manipulation during the annotation process if they were presented.

(6) In this work, we did not work with emojis. First of all, they are not always a sign of manipulation. Emojis can indicate an emotion or attitude to the news itself and not be related to manipulation. For example, "НАБО спіймало суддю Апеляційного суду Києва Глиняного на хабарі $35 тис (angryface)". This news evokes an emotion of anger towards a corrupt judge, and the emoji is appropriate here, but lexically the news does not contain a manipulative technique, so the emoji is not a sign of manipulation in this context. Secondly, even when an emoji is next to a manipulative phrase, it can refer to different manipulative techniques that appeal to the emotions that we distinguish.

The instructions for data annotation are available here [4].

For the annotation process, the LabelBox* program is utilized. To ensure consistency, 20% of the posts undergo overlap for consensus analysis.

To annotate the data, volunteers were invited. Eligibility criteria for volunteers included Ukrainian citizenship and proficiency in Ukrainian language. Volunteers did not require special skills and knowledge. The initial step with volunteers involved conducting interviews to assess their suitability and grasp of motivation. Subsequently, annotators were provided with instructions, and their questions were addressed. Following this, each annotator's first 80 annotated posts were reviewed, and feedback was provided within a shared Telegram group. Additionally, 20 random texts per annotator were examined weekly.

Before reviewing by supervisors, among 1600 texts annotators had 66% of agreements. Considering the highly complex nature of the task, we believe this level of agreement is reasonably high. The most difficult issue that annotators encountered was consistency in the annotation span. An unselected phrase, word, or punctuation mark caused the percentage of agreement to be decreased. Also, the annotators disagreed about the manipulation technique. For example, they confused the technique Loaded Language and Appeal to other Emotions with Appeal to Anger and Appeal to Fear or even with Doubts in the Ukrainian Government or Doubts in the Ukrainian Army. Sometimes, volunteers identified only one manipulative technique in a span, even though there were two. Moreover, it was very difficult for some annotators to separate emotions from the news itself (even if it is true) with manipulations that appeal to emotions (deliberate intensification of emotions).

Supervisors reviewed all posts that included six manipulative techniques such as Doubts in the Ukrainian Government, Doubts in the Ukrainian Army, Doubts in Ukrainian National/Credible Media, Doubts in Partners' Help, Other Doubts in Ukraine, and Black-and-white fallacy. Supervisors did not double-check posts that concluded only manipulation techniques from these four: Appeal to Fear, Appeal to Anger, Appeal to Hate and Disgust, and Loaded Language and Appeal to other Emotions. If those techniques were in the post with checked techniques, they also were checked. Also, supervisors did not double-check the posts, which were marked by annotators as posts that did not have any manipulations. After revision by supervisors, the agreement between annotators reached 90%.

The analysis of the annotation process, its peculiarities, and its complexity are described in Appendix A.

4.5 Corpus statistics

The corpus has 1,877 posts with at least one manipulation technique. 92% of them are from the pro-Russian channel. Figure 4.1 provides the distribution of the labeled data.

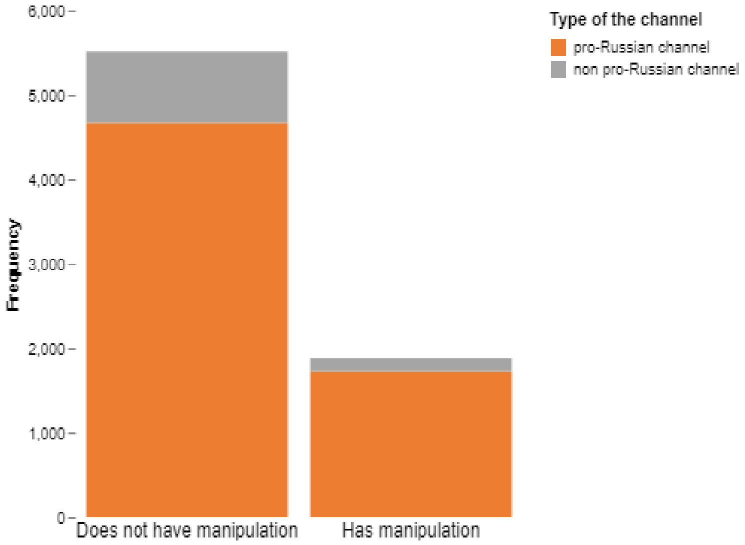


FIGURE 4.1: The distribution of the labeled posts.

After the supervisors' revision, the total number of manipulation technique instances found in the posts is 3,472. Table 4.2 reports some statistics about the manipulation techniques in the corpus. The characters include punctuation if it is presented in a manipulation fragment, but they do omit emojis.

Technique	Freq.	Avr length words	Avr length char
Loaded Language & other Emotions	1119	2	13
Appeal to Fear	633	5	34
Doubts in Government	544	7	42
Appeal to Anger	501	6	37
Appeal to Hate and Disgust	380	3	19
Black-and-white fallacy	106	11	67
Doubts in Army	81	8	53
Doubts in Media	61	7	45
Doubts in Partners' help	35	7	52
Other Doubts in Ukraine	12	8	51
Total	**3472**	-	-
Average	-	**4**	**29**

TABLE 4.2: Corpus statistics: each manipulation technique's frequency, average length in terms of words and characters.

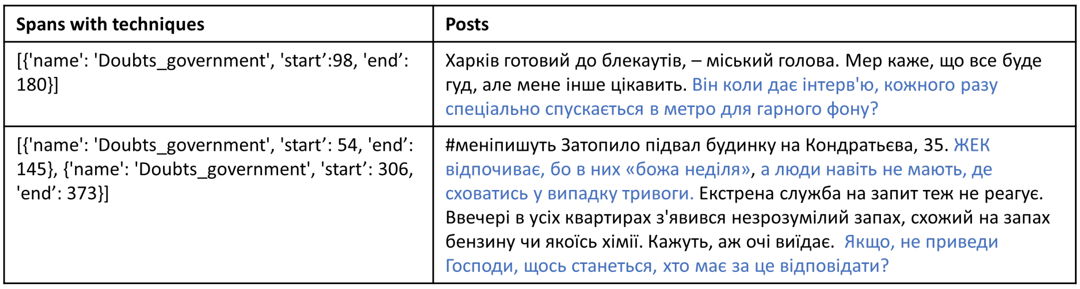
The classes are imbalanced. The most common manipulation techniques are Loaded Language and other Emotions, Appeal to Fear, Doubts in Government, and Appeal to Anger. The least common techniques are Other Doubts in Ukraine, Doubts in Partners' help, Doubts in Media, and Doubts in Army. The average length of the manipulation fragment is 4 words or 29 characters. On average, the longest technique is the Black-and-white fallacy because it often includes complete phrases with two options. While the shortest is the Loaded Language and other Emotions technique. This class includes, among other things, swearing and exclamations, which are often short.

4.6 Tasks

Span Identification is a binary task to detect manipulation span, which means identifying the specific span or text fragments containing any manipulation technique in a given plain-text document.

A span is a fragment of a text that belongs to a specific category, in our case, a manipulation technique. A text can consist of multiple spans, each with the same or different labels. Spans are similar to Named Entity Recognition (NER). However, spans with different techniques can overlap: a word or phrase may simultaneously be part of two different spans. The example of spans are shown in Figure 4.2 below.

Technique Classification is a multi-label, multi-class classification task, to classify what manipulation technique is used in a given span in the context of a full document. We addressed the classification task with ten manipulation techniques. In our work, we do not identify the manipulation span for this task.



Spans with techniques	Posts
[{'name': 'Doubts_government', 'start':98, 'end': 180}]	Харків готовий до блекаутів, – міський голова. Мер каже, що все буде гуд, але мене інше цікавить. Він коли дає інтерв'ю, кожного разу спеціально спускається в метро для гарного фону?
[{'name': 'Doubts_government', 'start': 54, 'end': 145}, {'name': 'Doubts_government', 'start': 306, 'end': 373}]	#менiпишуть Затопило підвал будинку на Кондратьєва, 35. ЖЕК відпочиває, бо в них «божа неділя», а люди навіть не мають, де сховатись у випадку тривоги. Екстрена служба на запит теж не реагує. Ввечері в усіх квартирах з'явився незрозумілий запах, схожий на запах бензину чи якоїсь хімії. Кажуть, аж очі виїдає. Якщо, не приведи Господи, щось станеться, хто має за це відповідати?

FIGURE 4.2: The examples of posts with manipulation techniques (shown in blue). The left column shows the character spans and the name of the manipulation technique(s) belonging to that span.

4.7 Baselines

The baselines for the Span Identification task are a very simple logistic regression classifier with default parameters (Baseline: LR) and human-level performance (Baseline: Human performance). For human-level performance, 500 randomly selected posts from the dataset for annotation were used for separate manual labeling. Seven volunteers were provided with only the names of the manipulation techniques and brief definitions without further instructions or consultations.

The baseline model for the Technique Classification task is a random selection of one of the 10 techniques (Baseline: Random Selection) and consistently choosing the most commonly represented class, in our case, Loaded Language and other Emotions class for 10 classes detection (Baseline: Majority Class).

4.8 Evaluation Measures

We used standard evaluation metrics for Classification task, including precision, recall, and weighted-average F1-score, which provides a single score that balances precision and recall.

Span Identification is a binary task, so all spans with manipulations are considered equal. Therefore, all overlapping annotations, regardless of their techniques, are first merged. In the example in Figure 4.3, after merging two spans with different techniques, they became one span.

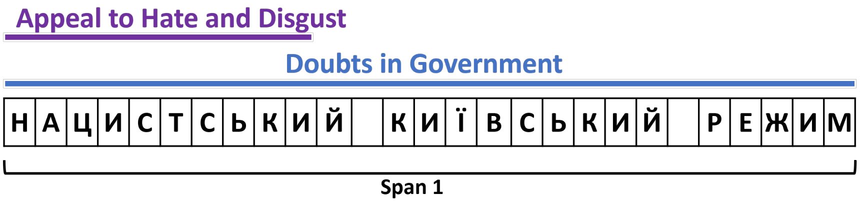

Appeal to Hate and Disgust
Doubts in Government
Н А Ц И С Т С Ь К И Й К И Ї В С Ь К И Й Р Е Ж И М
Span 1

FIGURE 4.3: The example of span merging.

For this task, we used the character-based average score. Using the average score for a binary task offers a balanced and easily interpretable view of both precision and recall, ensuring fair assessment in contexts of class imbalance. Character-based means that in the text "Нацисти з Києва" we count "нацисти" not as 1 True Positive, but as 7 True Positives, that is, one for each character of the token "нацисти".

Chapter 5

Experiments and Discussion

5.1 Model

We split the corpus of posts into train (80%) and test (20%) subsets. While splitting the data into test and train sub-sets, we did not apply stratification based on technique names because a single text often contains multiple manipulation techniques, and ensuring an even distribution of each technique across subsets would have been complex. We divided the dataset based on the ID of the posts but afterward ensured that all classes were presented in both sub-sets. The distribution of the classes is presented in Figure 5.4.

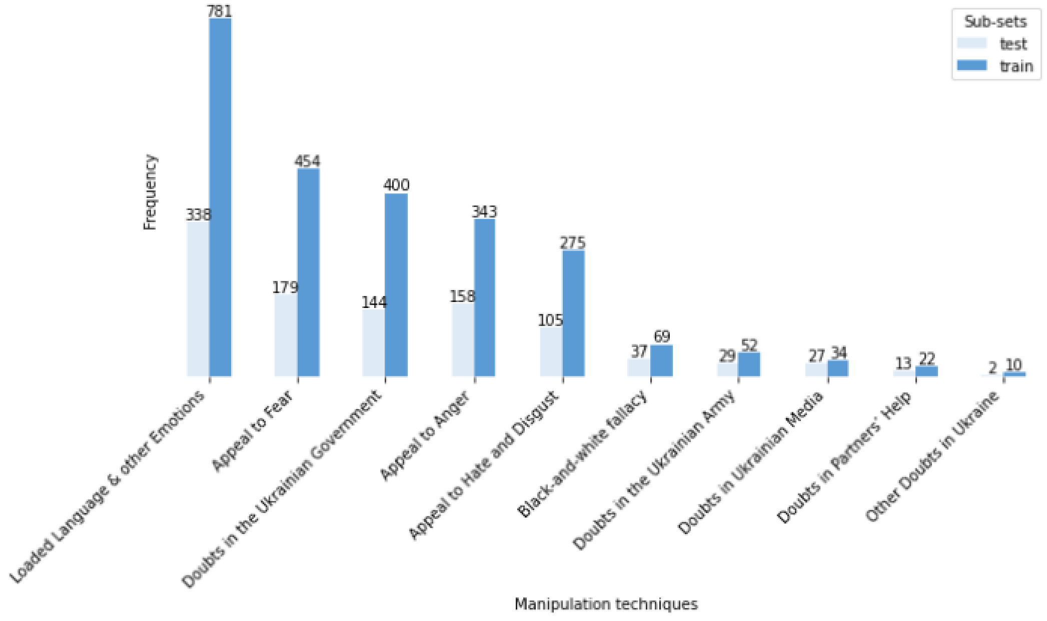


FIGURE 5.1: The frequency of the manipulation techniques in train and test datasets.

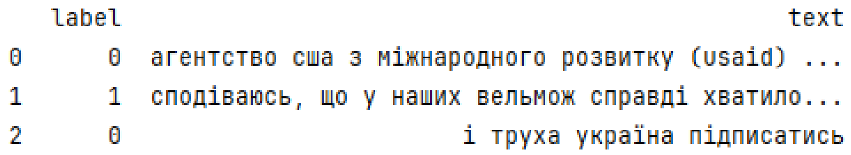
The statistics of the subsets are shown in Table 5.1 below.

We used the PyTorch framework and the pre-trained BERT model, specifically google-bert/bert-base-multilingual-uncased, which supports Ukrainian. We trained all models using the following hyper-parameters: batch size of 8 for train and evaluation, and sequence length of 128. The models for the Span Identification task were trained in 1 epoch, and the models for Technique Classification were trained in 5 epochs. We used AdamW with a learning rate of 4e-5 and a warmup proportion of 0.06 for optimization.

	Train	Test
Number of posts	1,308	561
Avg. post length in words	62	59
Avg. post length in char	429	410
Avg. span length in words	4	5
Avg. span length in char	29	30
Number of techniques	2,440	1,032
Avg. num. of techniques	1.4	1.3

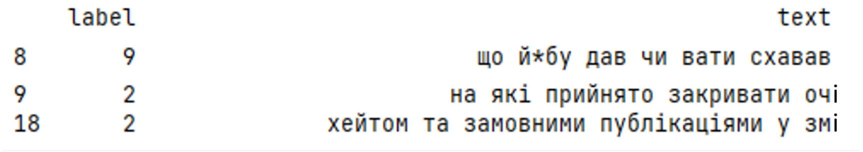
TABLE 5.1: Statistics of the training and test subsets of the dataset. Lengths are calculated in words and characters.

The provided texts to the model were preprocessed: text cleaning to remove newline characters, multiple spaces, URLs, emojis, and converting all text to lowercase. Afterward, the texts were split into sequences with binary tags for the Span Identification task 5.2 and with labels for the Technique Classification task 5.3.



```
   label                                            text
0      0  агентство сша з міжнародного розвитку (usaid) ...
1      1  сподіваюсь, що у наших вельмож справді хватило...
2      0                          і труха україна підписатись
```

FIGURE 5.2: Splitting texts to sequences with binary tags for SI task.



```
    label                                   text
8       9              що й*бу дав чи вати схавав
9       2             на які прийнято закривати очі
18      2     хейтом та замовними публікаціями у змі
```

FIGURE 5.3: Splitting texts to sequences with labels tags for TC task.

Then, to improve the model's span recognition, including the consideration of established phrases and expressions, words spelled with errors, or camouflaging certain words, for example, as swear words, we created all possible continuous subsequences of each sequence. For example, the sentence "це жахливі кадри." is split into tokens ['це', 'жахливі', 'кадри', '.'], and we consider the following continuous sub-sequences: ['це'], ['це', 'жахливі'], ['це', 'жахливі', 'кадри'], ['це, 'жахливі', 'кадри', '.'], ['жахливі'], ['жахливі', 'кадри'], ['жахливі', 'кадри', '.'], ['кадри'], ['кадри', '.'], ['.'].

In our dataset, we encountered class imbalance issues, particularly with the representation of certain classes. To address this, we employed two approaches: the weighted method, which assigns more weight to the less represented classes, and upsampling, which involves increasing the number of instances in the minority classes to achieve a more balanced dataset.

5.2 Synthetic text generation for upsampling

For the less common manipulation techniques, namely Other Doubts in Ukraine, Doubts in Partners' help, Doubts in Media, Doubts in Army, and Black-and-white fallacy, we utilized a language model, specifically LLM GPT-3.5 Turbo, to syntactically generate around 400 texts for each class. This was achieved by crafting prompts tailored to each class and providing instructions for the model to follow.

Each prompt had a structured format, consisting of elements such as who, what is doing, methods to use for manipulation, and in some cases, topic. These elements guided the model in generating text relevant to the specified manipulation technique.

For example, generated text for Doubts in the Media technique with a prompt: Write a news item that national media doesn't tell the truth with doubts about deforestation, using examples and instruction. Output: "Deforestation leads to the destruction of natural habitats and the extinction of numerous species of plants and animals. [START] However, the independent media often avoid this problem by trying to simplify the situation and not paying attention [END] to the measures taken to reduce deforestation".

For example, generated text for the Black-and-white fallacy technique with a prompt: "Write a news item with black-and-white narratives use judgment in provocative question about deforestation, using examples and instruction. Output: [START] The choice between preserving and destroying nature is a choice between the future and death![END]"

In the results, the generated texts often repeated similar phrases or words, sometimes contained manipulative techniques other than the one specified, and did not always correctly mark the beginning and end of the manipulative span. The instructions for each class are provided here [4].

We split the upsampled data on train and test. Firstly, we applied synthetically generated data only on train sub-set. This allowed us to evaluate the model with different parameters on the same test data to measure its performance consistently. Then, we applied the upsampling also on the test sub-set to further analyze the model's behavior with the presence of more balanced data.

5.3 Experiments and Results

Applying our approach described above, we trained models for two sub-tasks: Span identification and Technique Classification. The files with codes are available in the GitHub repository [4].

5.3.1 Span Identification task

Table 5.2 contains the results of span prediction with any manipulative technique with initial data and upsampling, which are compared with the two baselines: simple logistic regression and human-level performance.

Model	Accuracy
Baseline: LR	10.42
Baseline: Human performance	6.09
BERT-base-multilingual	25.90
BERT-base-multilingual (train&test upsampling)	**56.66**

TABLE 5.2: The evaluation of the results for the Span Identification task.

Due to the complex task and the wide range of possible mistakes, both baselines show low accuracy which indicated which demonstrates the relevance of the chosen task for the study.

The model with initial data got around 26% overlapping accuracy on the character level and overperformed baselines. However, on average, it predicted much longer spans compared with the truth dataset (309 characters vs 30 characters) and generated mostly one span with manipulation per post, while in the truth sub-set, on average, there are close to two spans (1.16 vs 1.77).

The fine-tuned BERT model with upsampling data predicted spans much better. The overlapping accuracy is around 57%. It generates an average of 1.27 spans per post, and in the truth subset, the average is similarly close at 1.37 spans per post. However, it also predicted much longer spans than the truth dataset (198 characters vs 63 characters).

Error Analysis

The model with upsampling performed well in identifying spans with manipulation techniques compared to baselines, and the model was trained on the initial data. However, applied models have several drawbacks that need further attention:

1) The model predicts spans that are much longer than they should be.

Figure 5.4 presents the distribution of span lengths for the results of the model with the initial dataset. The model captured some of the short spans but also generated long ones.

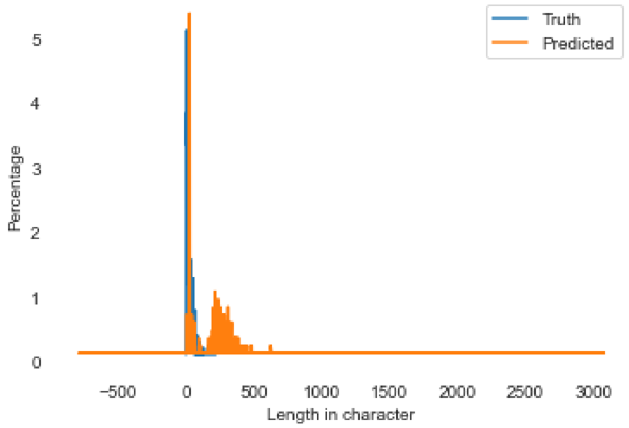


FIGURE 5.4: Distribution of span lengths.

In the upcoming research phase, we aim to address this issue by exploring the integration of both part-of-speech (PoS) and named entity (NE) embeddings.

2) The model predicts only one span per text, whereas often, multiple spans exist in reality. This problem requires further investigation, one of the hypotheses is that the model predicts a single but long span that combines several spans.

5.3.2 Technique Classification task

In the Technique Classification task, we fine-tuned a pre-trained BERT model with initial data and with upsampling in the train set only and in the train and test sub-sets and compared models to two baselines for 10 manipulation techniques, which included extension sub-classes for some manipulation techniques like Doubts and Lauded Language. Table 5.3 shows the results.

Model	Precision	Recall	weighted F1
Baseline: Random Selection	18	10	12
Baseline: Majority Class	11	33	16
BERT-base-multilingual (weights)	42	41	41
BERT-base-multilingual (train upsampling, no weights)	60	63	59
BERT-base-multilingual (train upsampling, weights)	64	60	62
BERT-base-multilingual (train&test upsampling, no weights)	63	64	**63**

TABLE 5.3: The evaluation of the results for the Technique Classification task for 10 classes.

The fine-tuned BERT model, trained on the initial dataset with class weighting, achieved a weighted average F1 score of 41%. It overperformed both baselines. However, the model failed to recognize the two classes with the fewest instances: Other Doubts in Ukraine (2 instances) and Doubts in partners' help (27 instances). They both were confused with Doubts in Government class mostly. Conversely, the classes Doubts in Media and Black-and-White Fallacy, despite their low representation, were effectively identified by the model, achieving F1 scores of 43% and 46%, respectively. The distinct patterns associated with the Black-and-White Fallacy technique likely contributed to its higher recognition rate. The classification report is presented in Table 5.4.

To address data imbalance, upsampling was applied. Firstly, we presented result for models with different parameters with upsampling train sub-set only and then, both train and test sub-sets.

Using the fine-tuned BERT model with synthetically generated data for training sub-set only significantly improved performance without weights and with weights, achieving a micro-weighted F1 score of 59% with no weights and 62% with weights.

Considering the model with no weights, some classes that were upsampled demonstrated there better performance, with F1 scores exceeding 60%, like Black-and-White Fallacy (75% vs 46%), Doubts in Media (90% vs 43%), and Doubts in partners' help (64% vs 0%) techniques, the last class previously had poor performance with the only initial data. Doubts in the Army class has poor Recall, which influences the F1 score (19% vs 26%). The model confuses this class with the Doubts in Government class. Besides, the performance of the Doubts in Government class got much better (62% vs 33%). The class Other Doubts also was not recognized by the model by its poor representation in the test dataset (2 instances). The Appeal to Anger class showed improved recognition with a 39% F1 score, up from 20%. In contrast, the F1 score for the Appeal to Hate and Disgust class significantly dropped (5% compared to 23%), indicating overprediction in that technique. The model frequently misclassified instances with the Loaded Language and Other Emotions class. The Loaded Language and Other Emotions class itself showed better performance. The F1 score for this class grew from 59% to 80%. The results is presented in Table 5.4.

To address the issue of the model not recognizing a specific class as Other Doubts in Ukraine, we added weights. This adjustment resulted in a 3 percentage point increase in the weighted F1 score, bringing it to 62%. The result for the class Other Doubts in Ukraine is more prominent, the F1 score increased to 67% from zero. The

	BERT (weights)	BERT (up+no weights)	BERT (up+weights)	BERT (weights)	BERT (up+no weights)	BERT (up+weights)	BERT (weights)	BERT (up+no weights)	BERT (up+weights)	
Manipulation Technique	**Precision**			**Recall**			**F1-score**			**Support**
Loaded Language & other Emotions	68	74	77	52	88	66	59	**80**	71	338
Appeal to Fear	38	55	74	58	78	60	46	65	**66**	179
Doubts in Government	29	57	58	38	68	60	33	**62**	59	144
Appeal to Anger	25	50	48	17	32	63	20	39	**54**	158
Appeal to Hate and Disgust	21	23	24	23	3	31	22	5	**27**	105
Black-and-white fallacy	43	63	85	49	92	76	46	75	**80**	37
Doubts in Army	33	100	74	21	10	48	26	19	**58**	29
Doubts in Media	46	100	76	41	81	81	43	**90**	79	27
Doubts in Partners' help	0	67	62	0	62	77	0	64	**69**	13
Other Doubts in Ukraine	0	0	100	0	0	50	0	0	**67**	2
accuracy							41	63	60	1032
macro avg	30	59	68	30	51	61	29	50	63	1032
weighted avg	42	60	**64**	41	**63**	60	41	59	**62**	1032

TABLE 5.4: The classification reports for BERT model (initial data in test sub-set).

model correctly identified this class once but misclassified it as Doubts in Government in another case. Additionally, performance for the classes Doubts in Army and Appeal to Hate and Disgust saw significant improvements, with F1 scores increasing from 19% to 58% and from 5% to 27%, respectively. The model also exhibited reduced misclassification of the Appeal to Hate and Disgust class as Appeal to Fear and Loaded Language & Other Emotions. Overall, this model shows good result in reducing the misclassification and paying more attention to classes with less cases. Table 5.4 shows the classification report.

When comparing the confusion matrices for models with initial and upsampling in train sub-set with weights in Figure 5.5, we observe that most confusion occurs among the emotion-appealing techniques (the lower left square in both heatmaps), which is logical because the same words can evoke different emotions depending on the context. Besides, volunteers often mixed these techniques, and supervisors did not fully correct these errors (see Chapter 4.4). Additionally, the technique Doubts in Government is frequently confused with emotion-appealing techniques due to their overlapping spans. And Doubts in the Army are predicted as Doubts in Government class. However, all these problems are better handled by the model trained with additional data.

To address the problem of underrepresented cases in the test sub-set, we added the generated data to the test sub-set as well and evaluated model results on this augmented dataset. We did not use weights as the classes were already balanced. The results of this evaluation are presented in Table 5.5. On average, the weighted

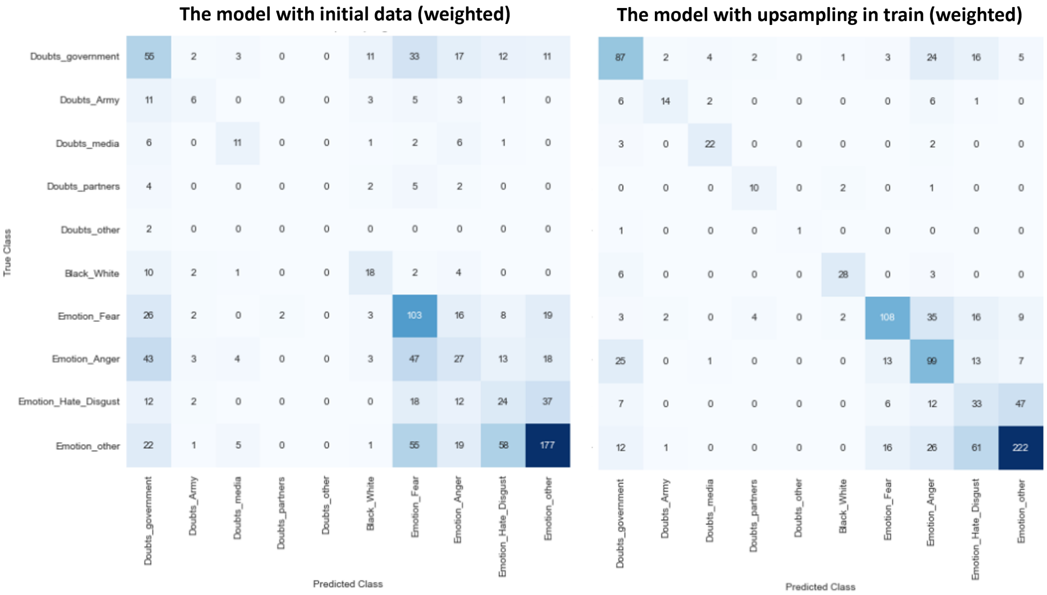


FIGURE 5.5: The confusion matrices for both models.

F1-score increased by 1 percentage point to 63%. For the upsampling classes, the performance improved significantly compared to the best F1 scores shown by previous models: Black-and-White Fallacy (80% vs 87%), Doubts in Army (58% 91 86%), Doubts in Media (90% vs 94%), Doubts in partners' help (69% vs 86%), Other Doubts in Ukraine (from 67% to 94%). However, this improvement might be influenced by the repetitive patterns in the synthetically generated sentences.

Additionally, the model performed worse for all classes related to emotional appeals. The results were worse compared to the previous best outcomes, with the model confusing these classes and favoring the largest one, which is Loaded Language & Other Emotions. The most challenging class for the model was Appeal to Hate and Disgust. To address this, further research is needed to identify the distinctions between the Appeal to Hate and Disgust and Loaded Language & Other Emotions classes, and to train the model to differentiate them accordingly.

Error Analysis

All models possess distinct advantages. They are adept at handling unstructured data from social media in Ukrainian, which often contain swear words (with letters obscured by symbols), incomplete or illogical sentences, and mistakes. However, both models have the same repeated problems that need further research:

1) The model tends to confuse emotion-related classes with each other. To address this issue, the next step involves utilizing the IBM Watson NLU API to separately identify and classify emotions like sadness, joy, fear, disgust, and anger and use that as an additional layer for classification task. Additionally, supervisors will conduct a review of annotations for emotion-related classes to ensure the elimination of any annotation errors.

2) The model encounters difficulties when manipulation techniques overlap within the same span fully or partially, predicting the same technique twice, as in the example in Figures 5.6 and 5.7. This problem requires further exploration.

	Precision	**Recall**	**F1-score**	**Support**
Loaded Language & other Emotions	55	75	64	207
Appeal to Fear	43	51	47	138
Doubts in Government	39	36	37	115
Appeal to Anger	31	29	30	100
Appeal to Hate and Disgust	19	3	6	87
Black-and-white fallacy	86	88	87	108
Doubts in Army	98	86	91	99
Doubts in Media	94	93	94	106
Doubts in Partners' help	87	85	86	81
Other Doubts in Ukraine	95	93	94	83
accuracy			64	1124
macro avg	65	64	64	1124
weighted avg	**63**	**64**	**63**	1124

TABLE 5.5: The classification report for BERT model with upsampling in train and test sub-sets with no weights

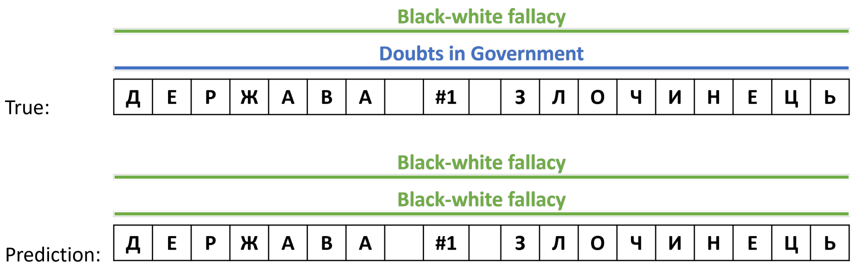


FIGURE 5.6: Example of prediction only one class for the same span.

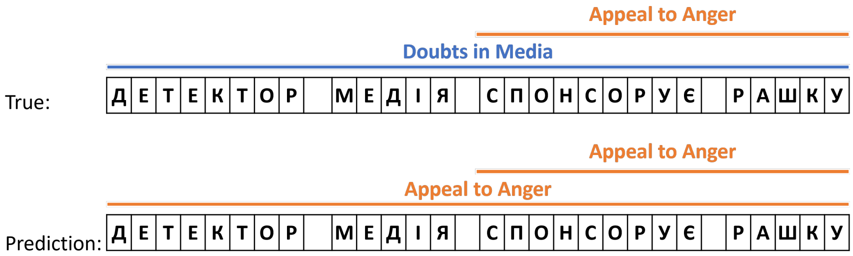

Appeal to Anger
Doubts in Media
True:
Д Е Т Е К Т О Р М Е Д І Я С П О Н С О Р У Є Р А Ш К У
Appeal to Anger
Appeal to Anger
Prediction:
Д Е Т Е К Т О Р М Е Д І Я С П О Н С О Р У Є Р А Ш К У

FIGURE 5.7: Example of prediction of the same class for partially overlapping spans.

Chapter 6

Limitations and Ethical Implications

There are limitations in our work, which is important to mention.

(1) We’ve chosen to focus on just four out of the total 18 manipulation techniques. However, exploring the remaining techniques could offer valuable insights.

(2) We worked only with texts in Ukrainian; working with texts in Russian could provide new manipulative patterns in the lexicon.

(3) We aim to analyze only text data, but multimodal data, such as text with pictures or videos, might reveal more insights into using the manipulation techniques. It might be a task for further research.

(4) In our research project, we consider only data from Telegram, but news data from other social networks, such as Facebook or Instagram could have their own pattern or insights of using the manipulation techniques. We chose the list of the Telegram channels in the “News and media” category in Ukraine tagged by the Russian Telegram channel statistics website TGStat, which may not be accurate and complete due to the incorrect display of channels in the occupied territories of Ukraine as Russian. There is no alternative non-Russian resource with Telegram statistics.

We do not intend to draw causal inferences that all manipulations in the news channels on Telegram in Ukraine are spread by Russia, as some news channels or people behind them may be spreading the manipulated news unconsciously or have their own political agenda. But either way, it helps Russia in their "information warfare" against Ukraine.

It is important to acknowledge that the authors of this research project are Ukrainians residing in Ukraine during a full-scale active Russo-Ukrainian war, and their perspectives may be shaped by their personal experiences. Despite this, we are sure that these unique circumstances provide an advantage to this research project in developing a precise approach to identify some of the manipulation techniques employed by Russia. This effort can prove beneficial not only for Ukrainian society but also for all countries where Russia has its interests and unfulfilled ambitions.

Chapter 7

Conclusion

Automatically detecting fake news that manipulates perceptions and opinions on social media is crucial for Ukraine, particularly during the ongoing Russo-Ukrainian war. To tackle this issue, we developed our own manually annotated corpus, consisting of posts from Ukrainian news channels on Telegram, all in the Ukrainian language. Each annotation includes the fragment or span where a manipulation technique is detected, along with the specific technique from a set of selected manipulation techniques. We then trained a pre-trained BERT model to recognize these spans and their associated manipulation techniques, enhancing the model's performance through syntactically generated text. The models successfully handled unstructured data from social media in Ukrainian. However, both models tend to confuse classes within the meta classes, such as Doubts and Loaded Language. Additionally, they often predict much longer spans and typically identify only one span. A significant area for further research is also addressing overlapping spans, enabling the models to accurately predict multiple spans and assign different classes to them.

In conclusion, our work represents an attempt to analyze manipulation techniques in social media news data, offering valuable insights for the research community in the field of fake news detection.

Appendix A

Report on Volunteer Engagement

Volunteers were enlisted to annotate texts across Facebook, Telegram, and Instagram. Eligibility criteria for volunteers included Ukrainian citizenship and proficiency in Ukrainian language. Volunteers did not require special skills and knowledge. The average age of the volunteers was 38, the youngest was 22, and the oldest was 64.

The first phase with volunteers includes conducting interviews to evaluate their suitability and understanding of motivation. Next, annotators receive instructions and have their questions answered. Then, each annotator's initial 80 annotated posts are reviewed, and feedback is given via a shared Telegram group.

During the process, there were technical, social, and cultural challenges, which are described below:

(1) Most volunteers found it difficult to focus for more than an hour a day, after which they said they started to get confused, make mistakes, double-check themselves, or see one type of manipulation everywhere.

(2) Most volunteers found it difficult to be involved in the work every day for a week or two and to fulfill the acceptable work hours (they forgot, other plans came up, equipment did not work, etc.)

(3) Of the 18 volunteers who passed the interview, only 12 joined the annotation process, and 9 annotated texts for over 3 hours. Therefore, recruitment and training of volunteers should be ongoing to achieve results.

(4) Volunteers who said at the interview that they were interested in the topic of propaganda were more involved in annotation than those who did not mind devoting time to a social project, either because it was not difficult or because they would also like to write a diploma thesis on this topic, etc.

(6) Depending on their experience and the information field, it was easy for each volunteer to master some manipulation techniques and difficult to find or overlook others. These were the difficulties that everyone faced:

- labeling the news with manipulation of the emotion of anger because the facts of the news cause anger, although the lexical content of the news does not contain manipulation or intensification of emotions;
- difficulties in distinguishing between manipulation of appealing to anger and appealing to fear;
- difficulty distinguishing manipulation with black-and-white narratives.

As a result, after the annotation process, the following conclusions were formed: recruiting and working with volunteers requires precise planning, writing clear instructions and testing them on several people before the full start, time allocation for interviews and training, identifying patterns of mistakes and additional training, constant recruitment of volunteers and training supervisors to check the quality of work.

Bibliography

[1] DiResta, R., Shaffer, K., Ruppel, B., Sullivan, D., Matney, R.C., Fox, R., Albright, J., Johnson, B. (2018). The tactics tropes of the Internet Research Agency.

[2] Jaiswal, J., LoSchiavo, C.E., Perlman, D.C., Perlman, D.C. (2020). Disinformation, Misinformation and Inequality-Driven Mistrust in the Time of COVID-19: Lessons Unlearned from AIDS Denialism. AIDS and Behavior, 24, 2776 - 2780.

[3] Paul, C., Matthews, M. (2016). The Russian "Firehose of Falsehood" Propaganda Model: Why It Might Work and Options to Counter It.

[4] Volkova N. (May 2024). Github repository. Computational Detection and Analysis of Manipulation Techniques in News Channels on Telegram in Ukraine. https://github.com/NataVolkova/manipulation-techniques-detection-in-news-uk

[5] Shu, K., Sliva, A.L., Wang, S., Tang, J., Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. ArXiv, abs/1708.01967.

[6] Rashkin, H., Choi, E., Jang, J.Y., Volkova, S., Choi, Y. (2017). Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. Conference on Empirical Methods in Natural Language Processing.

[7] Barrón-Cedeño, A., Jaradat, I., Da San Martino, G., Nakov, P. (2019). Proppy: Organizing the news based on their propagandistic content. Inf. Process. Manag.

[8] Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., Nakov, P. (2019). Fine-Grained Analysis of Propaganda in News Article. Conference on Empirical Methods in Natural Language Processing.

[9] Da San Martino, G., Barr'on-Cedeno, A., Wachsmuth, H., Petrov, R., Nakov, P. (2020). SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. International Workshop on Semantic Evaluation.

[10] Da San Martino, G., Barrón-Cedeño, A., Nakov, P. (2019). Findings of the NLP4IF-2019 Shared Task on Fine-Grained Propaganda Detection. ArXiv, abs/1910.09982.

[11] Gupta, P., Saxena, K., Yaseen, U., Runkler, T.A., Schütze, H. (2019). Neural Architectures for Fine-Grained Propaganda Detection in News. ArXiv, abs/1909.06162.

[12] Yu, S., Da San Martino, G., Mohtarami, M., Glass, J.R., Nakov, P. (2021). Interpretable Propaganda Detection in News Articles. Recent Advances in Natural Language Processing.

[13] Vijayaraghavan, P., Vosoughi, S. (2022). TWEETSPIN: Fine-grained Propaganda Detection in Social Media Using Multi-View Representations. North American Chapter of the Association for Computational Linguistics.

[14] Hasanain, M., Ahmed, F., Alam, F. (2024). Can GPT-4 Identify Propaganda? Annotation and Detection of Propaganda Spans in News Articles.