UKRAINIAN CATHOLIC UNIVERSITY

MASTER THESIS

What are your triggers? Context-Dependent Detection of Emotional Triggers in Influence Campaigns

Author:
Anastasia HOLOVENKO

Supervisor:
Andriy KUSYY

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, Anastasia HOLOVENKO, declare that this thesis titled, “What are your triggers? Context-Dependent Detection of Emotional Triggers in Influence Campaigns” 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:

“A samurai has no goal, only a path.”

Bushido

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Master of Science

What are your triggers? Context-Dependent Detection of Emotional Triggers in Influence Campaigns

by Anastasia HOLOVENKO

Abstract

Manipulating with emotions is a known technique used in influence campaigns to shape public opinions. Nevertheless, it requires a deep understanding of such emotional "triggers" within the populations, demographic groups, or individuals being targeted. While humans may intuitively identify triggers, we demonstrated that fine-tuning an LLM model can help automatically detect emotional triggers for a selected population segment by utilizing the context of a representative persona. This approach has shown significant improvement in the compound tasks of triggered emotion detection, emotion sentiment classification, trigger intensity evaluation, and identification of text spans that indicate the cause of the trigger. As a result, we propose an active learning system that enables continuous improvement of the LLM with minimal resources.

Acknowledgements

I would like to express my heartfelt gratitude to everyone who supported me in completing my Master's Degree and writing my thesis:

- Grammarly: Thank you for providing me with the scholarship to study at the Ukrainian Catholic University.
- Andriy Kusyy and LetsData Inc: Your support, patience, and time and resources were invaluable in fulfilling my research interests.
- Rodica Pîrgari, Vida Nedova, and the WatchDog.MD Community: Your assistance in organizing the data annotation for this project was crucial.
- All my loved ones: Your unwavering support throughout this journey meant a lot to me.

And, of course, I would like to thank the Armed Forces of Ukraine, without whom I would never have been able to complete this work.

Contents

List of Figures

List of Tables

List of Abbreviations

EDA	**E**xploratory **D**ata **A**nalysis
LLM	**L**arge **L**anguage **M**odel
ZSL	**Z**ero-**S**hot **L**earning
FSL	**F**ew-**S**hot **L**earning
FT	**F**ine-**T**uning
JSON	**J**avaScript **O**bject **N**otation
API	**A**plication **P**rograming **I**nterface

Dedicated to the Armed Forces of Ukraine and their families

Chapter 1

Introduction

Emotional appeals play a significant role in several known propaganda techniques (Da San Martino et al., 2019). The analysis of readers' emotional responses to news and social media posts has previously shown promising results in tasks such as fake news detection (Vosoughi, Roy, and Aral, 2018; Ghanem, Rosso, and Rangel, 2020; Kolev, Weiss, and Spanakis, 2022; Luceri, Boniardi, and Ferrara, 2023) and influence campaign detection (Bhaumik et al., 2023; Tseng et al., 2024). However, this has highlighted a significant limitation: emotions are highly subjective.

Extracting emotions from text and identifying their causes are influenced by the biases of the annotators who create the datasets, thereby introducing this bias into the models built on top of them. Consequently, manipulation detection is also biased towards the subjective emotional triggers of this group.

In contrast, influence campaigns often target specific topics and groups of people (Tucker et al., 2018). This requires malign actors to have a thorough understanding of their targets.

Our research aims to incorporate a degree of subjectivity to enhance the detection of emotions and emotional triggers. We utilize the context that an individual or group of individuals possesses, particularly focusing on how demographic segmentation can assist in this task.

For selected demographic clusters, we generate a representative Persona context and use it for LLM inference. Furthermore, we assess different approaches for such predictions, based on both in-context learning and fine-tuning. Finally, we design an active learning system that incorporates continuous improvement for the task of context-dependent emotional trigger prediction.

We have chosen Moldova as our country of interest. In October 2024, a referendum will be held to decide whether Moldova will join the EU. This is a crucial decision for the population of Moldova, given its strategic location, as well as its proximity to the Russia-Ukraine war.

To assess the emotional triggers present in the subset of Moldovan population, we selected a sample of manipulative and neutral Moldovan Telegram channels that are potentially, malign. The notion of malign aligns with the methodology presented by the European External Action Service ((EEAS), 2021; (EEAS), 2023; (EEAS), 2024), and also provides a framework for analyzing and responding to risks posed by such malign actors. In addition, this framework has not been previously utilized in related research.

This Master's thesis is structured as follows: Chapter 2 discusses the motivation for the selected task and the methodology for detecting emotional triggers. Chapter 3 outlines related research and explores the identified research gap. Chapter 4 describes the data-related steps, including dataset selection, filtering, and annotation, and presents a basic exploratory data analysis of the final dataset as well as a single output structure. Chapter 5 reports on the experimental setup, approaches, and results. Finally, Chapter 6 discusses the research’s limitations and suggests future work directions.

Chapter 2

Motivation and Methodology

2.1 The Era of Social Media and Interpretations

Social media has become a primary source of information and news. While journalistic standards assume neutrality in presenting information, people often seek interpretations from state actors, experts, bloggers, influencers, and the social media channels or groups they follow. This shift has made social media a critical platform for shaping public opinion.

This trend poses a significant threat to critical thinking, trapping many individuals in information bubbles shaped by their online consumption. These bubbles create isolated realities where people are exposed primarily to information that aligns with their preexisting beliefs, reinforcing their viewpoints and reducing exposure to diverse perspectives.

As stated in Eady et al., 2019, following the study on the data of Twitter users in the US, "More than a third of respondents do not follow any media sources, but among those who do, we find a substantial amount of overlap (51%) in the ideological distributions of accounts followed by users on opposite ends of the political spectrum".

The creation of controlled (social) media networks has become an appealing strategy for information intelligence operations. Modern authoritarian regimes invest enormous amounts of money to create and utilize "interpreters" to shape public opinion and promote the interests of the authority both in their own country, and to promote disinformation and destabilize foreign countries.

According to the European External Action Service (EEAS), 2021, Foreign Information Manipulation and Interference or

> *"FIMI is a pattern of behaviour that threatens or has the potential to negatively impact values, procedures and political processes. Such activity is manipulative in character, conducted in an intentional and coordinated manner. Actors of such activity can be state or non-state actors, including their proxies inside and outside of their own territory".*

In this paper, we refer to actors involved in FIMI as *malign*. We believe that analyzing social media content and understanding the emotional triggers that malign actors might exploit will help detect the presence of influence campaigns. This is crucial for developing strategies to mitigate their impact and promote a more informed and critically engaged public.

The detection of emotional triggers can fit within the framework for propaganda classification (analyzing specific types of propaganda that exploit emotional triggers) and be considered FIMI incidents (actions with malign objectives and intents), according to the EEAS methodology ((EEAS), 2023).

Moreover, the latter has also presented the Response Framework with examples of strategies that can be used to mitigate such threats, particularly in election campaigns ((EEAS), 2024).

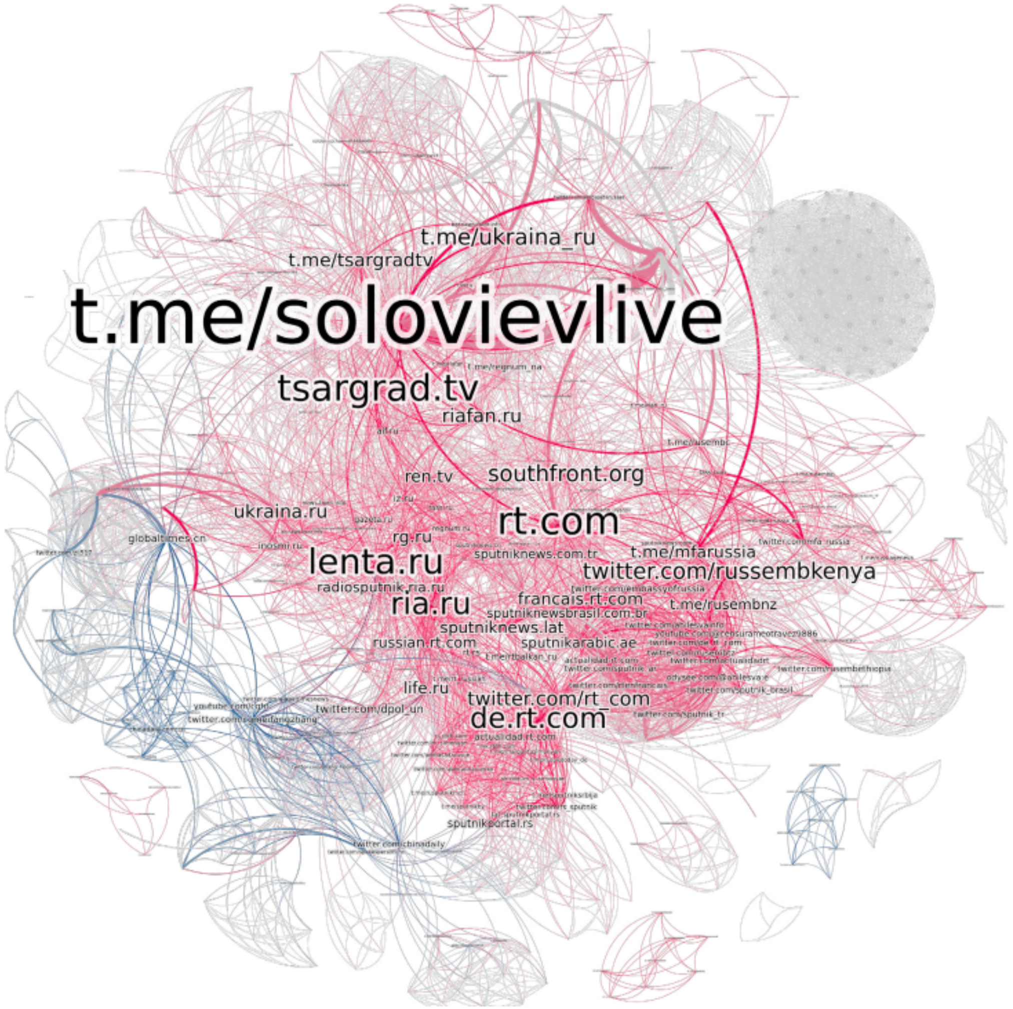


FIGURE 2.1: Example Network of Malign Actors (taken from (EEAS), 2023

2.2 Methodology

In our research, we aim to build a system that will help identify emotional triggers in social media posts.

The main question we focus on is the subjectivity present in each individual's perspective. We refer to this subjectivity as "context," which each individual possesses. Based on a person's context, they may or may not have an emotional response to the text they interact with.

To address this context-missing problem, we aim to use this notion of context description for LLM models to impersonate the target group whose subjectivity we want to explore. And to adopt a data-driven approach for selecting the contexts to assess, we use demographic clusters obtained from sociological research and questionnaires.

We then assess different LLM learning mechanisms to incorporate context and run inferences on the selected social media domain. Our aim is to identify the best approach for designing a learning system that continuously incorporates this context.

Influence campaigns are selected as our area of interest for the task of emotional triggers detection, but the task of detecting or classifying such campaigns is out of the scope of this current work.

2.3 Analysis Model

We define a model for emotional triggers that includes several components (Figure 2.2).

Firstly, it identifies the specific emotion being triggered. We select 6 primary emotion classes described in Shaver et al., 1987 and Parrott, 2001: anger, fear, sadness, surprise, love, and joy.

We then classify emotions as either negative or positive. We do not assign a default classification to the selected emotions to avoid incorporating bias into these classifications. For example, surprise can be both positive or negative, while sadness is typically seen as negative. However, nostalgic sadness can also be a positive feeling. Moreover, according to Parrott, 2014, anger can also be a positive feeling and act as a great motivational force.

Additionally, we introduce a trigger level, which, although highly subjective, can help us determine where the most significant emotional response is generated. The level of the emotional response is crucial because the strongest reactions can have the most substantial impact on a person's opinion and behavior.

Furthermore, our model includes the analysis of text causes for such triggers. This aspect is particularly interesting as it provides insights into the specific sensitivities that were targeted. By examining how the content was framed, we can explore whether the emotional triggers were used deliberately to influence the audience.

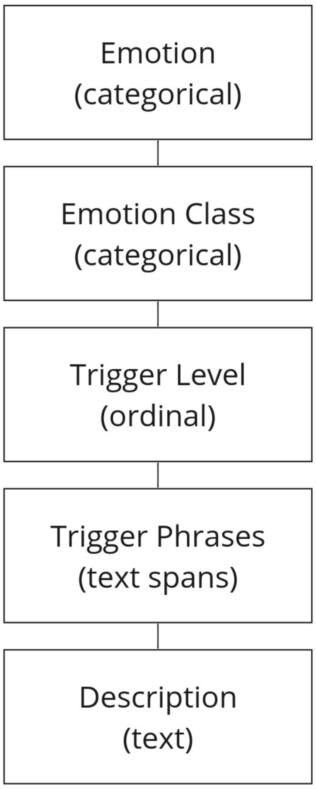


FIGURE 2.2: Proposed Components for Emotional Triggers Analysis Model

Finally, our model includes a general description of the elicited emotions. This allows us to provide explanations for individuals' reactions and incorporate these explanations into the training of the models and, later, model evaluation.

It is important to note that, by design, this model assumes that some emotion should be elicited after reading a social media publication; no emotion or neutral emotion is not allowed. Conversely, a trigger level can be used as a control variable in such cases. For example, what one may first consider as "neutral" could be presented as an emotion of joy but with the lowest trigger level.

Overall, these factors combined allow us to represent the emotional triggers as a complex construct.

2.4 Domain Selection

According to Office of the High Commissioner for Human Rights (OHCHR, 2024), "2024 has been called the biggest election year in history, with more than 60 countries, representing nearly half the world's population, holding elections". And all these elections are "testing democracy's health."

The Republic of Moldova is one of the countries that will also make a major decision for its future. Specifically, the referendum on the European Union membership will be held. Citizens will be expressing their desire to join or not join the European Union.

This decision is crucial because of Moldova's strategic location and its proximity to the Russia-Ukraine war, which poses a threat to Moldova itself if it were to expand. Additionally, part of Moldovan territory, Transnistria—a self-proclaimed state—already has Russian forces present and remains under significant Russian influence.

According to EEAS Report, Moldova, Moldovan government and Maia Sandu, in particular, were among those targeted by malign actors in the presented sample of FIMI incidents in 2022 (see Figure 2.3).

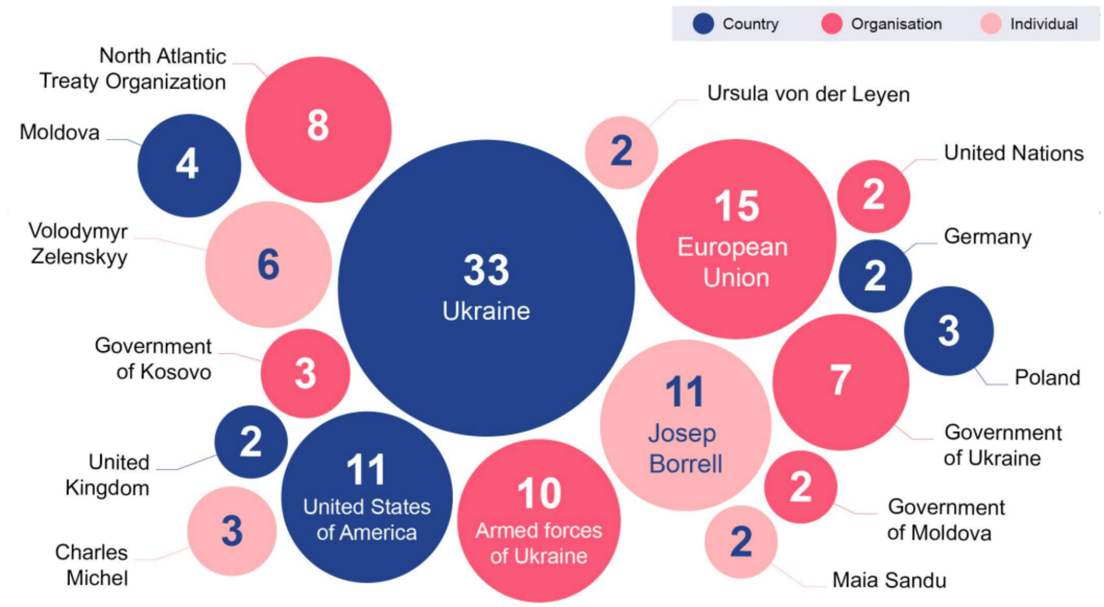


FIGURE 2.3: Sample of Targets of Malign Actors (taken from (EEAS), 2023

Finally, Telegram was selected as the social media platform for this research. It gained significant popularity following Russia's full-scale invasion of Ukraine in 2022, becoming the most popular mobile app in the Republic of Moldova (Nae, 2022, SimilarWeb, 2024). This highlights the importance of monitoring the channels that people use and detecting any potentially malign actors and narratives that could be present in the Moldovan segment.

2.5 Personas Context

To select demographic clusters and receiving characteristics to generate Persona context, we used proprietary sociological research conducted in Moldova in 2023. This research was based on a questionnaire that included basic demographic factors as well as questions on political and personal views, among other topics.

The presented approach and methodology can be applied to any selection of demographical clusters. For our work, we selected two demographical groups: with strong pro-EU views and strong anti-EU views, which accounted for 15% and 5% of the population, respectively.

The main objective for selecting these two clusters was to evaluate the similarities and differences between the emotional triggers possessed by representative Personas from these clusters, in addition to mastering the emotional triggers detection task.

For each of the selected clusters, we used their main characteristics to generate the context for a representative Personas from that cluster (see Table 2.1) using a prompt for GPT-4 (see Table 2.2).

Persona	**Context**
Persona 1	You are Mihai, a 35-year-old political analyst from Moldova, fluent in Romanian and passionately engaged with the latest news. You pride yourself on your pro-European stance, viewing your identity as intertwined with European values and Moldova's potential EU integration. With a keen interest in current affairs, you're a regular on social media platforms, where you consume and discuss news, especially concerning Moldova's political landscape and its relationship with the West. Disillusioned with certain political outcomes, you still hold strong pro-Western views, advocating for EU support in fighting corruption and economic development.
Persona 2	You are Alexei, a 60-year-old retired teacher living in Gagauzia. Fluent in Russian, you spend your days following the news through Russian media outlets, finding solace in their familiar perspective. Skeptical of Western influences, you proudly prioritize your Moldovan identity over linguistic ties, yet feel a strong affinity for Russia. Politically, you lean towards pro-Russian parties, disenchanted with the post-communist evolution and wary of the pace of societal change. Yet you remain deeply distrustful of the presidency and political elites, feeling underrepresented and threatened. Despite the global condemnation, you view Russia's actions in Ukraine with understanding, though not without reservations.

TABLE 2.1: Generated Personas Context

Prompt depends on single input variable , which is in free format:

- {CLUSTER_DESCRIPTION} - sociological cluster description.

Role	Content
system	As a sophisticated AI system designed specifically to support sociologists by turning complex sociological cluster research into engaging and insightful persona profiles, your role is to distill and interpret detailed sociological data. By crafting narratives that bring to life the diverse experiences and characteristics of people within specific sociological clusters, you will provide sociologists with a deeper, more empathetic understanding of different segments of the population. ###Instruction:### Carefully review the input sociological cluster research data to identify significant demographic details, psychographic characteristics, social behaviors, and any other relevant information. Transform this data into rich, narrative-driven persona profiles. Each narrative should paint a vivid picture of the persona, making them feel real and relatable, and should highlight their background, lifestyle, motivations, and the challenges they face. ###Output Format Description:### Your output should be a cohesive narrative paragraph for outlined persona, including the following elements woven into the story: - A fictional name to personify the cluster. - Age range, occupation, and education level to give context to their life stage and social positioning. - Key characteristics, including demographic and psychographic traits, interests, values, and lifestyle choices, to add depth to their personality. - Social behaviors, illustrating how they interact with others, their media consumption habits, and their community involvement. - Political views and mindset characteristics. ###Guidelines:### - Paragraph should be AT MOST 100 words long - Ensure that your answer is unbiased and does not rely on stereotypes - Give particular examples to describe the story of a persona ###Input Text Indicator:### "Input Sociological Cluster Data:" ###Output Indicator:### "Generated Persona Narrative: You are"
user	Input Sociological Cluster Data: “‘{CLUSTER_DESCRIPTION}“‘

TABLE 2.2: Representative Personas Generation

Chapter 3

Related Research

3.1 Emotion Classification

Emotion classification from text is a well-established task in Natural Language Processing. It has evolved from the task of sentiment analysis (Devika, Sunitha, and Ganesh, 2016; Aggarwal, 2018) which focused on analyzing the tone of a given text within three categories: neutral, positive, and negative. Emotion classification introduces a more fine-grained approach to emotion analysis of the given text. It can be based on different emotion models that include not only categorical but also dimensional emotion models (Nandwani and Verma, 2021).

The categorical model defines a finite number of classes and states of an emotion, while the dimensional model presents an emotion in a two-dimensional space,

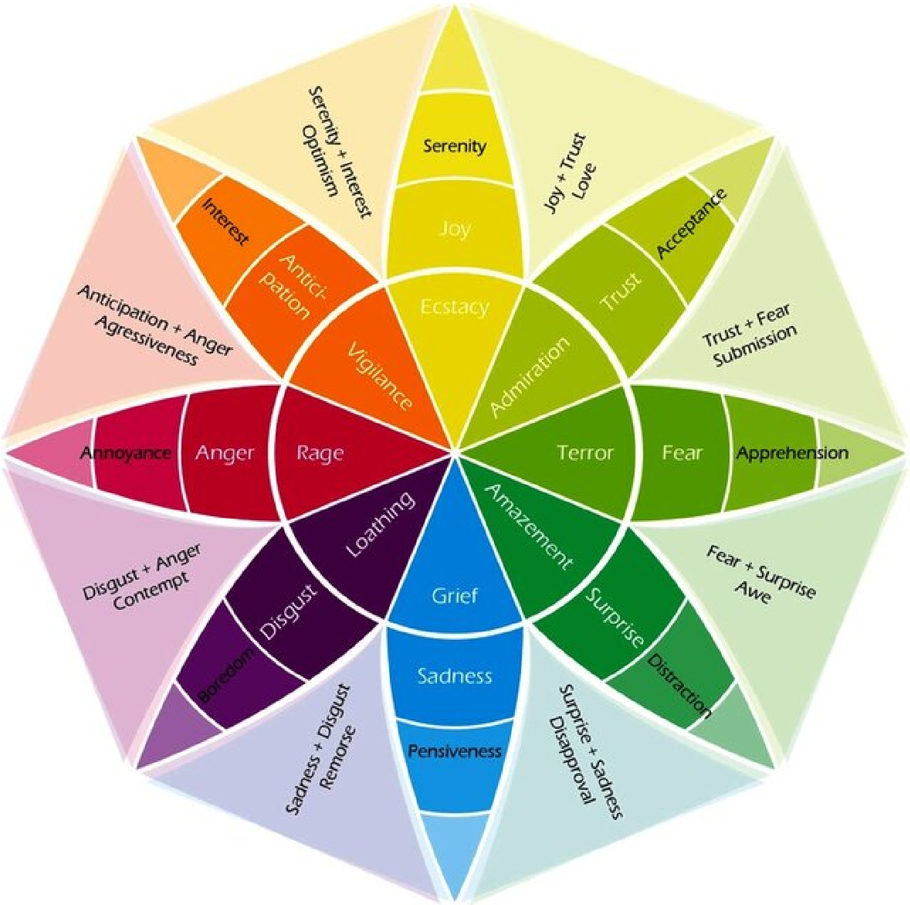


FIGURE 3.1: Plutchik's Emotion Model (Wheel of Emotions) (taken from Mondal and Gokhale, 2020)

composed of arousal (emotion polarity) and power (emotion intensity). While the categorical model is more commonly used in practice, the number of classes depends on the framework utilized in the research. Among the most popular emotion models are those proposed by Shaver et al., 1987, Ekman, 1992, Parrott, 2001 and Plutchik, 2003.

The number of classes can also depends on the research objectives and the data collection process. For example, among the most popular datasets, the number of emotion classes ranges from 6 for the CARER dataset Saravia et al., 2018 and the ROCStories dataset Mostafazadeh et al., 2016 to 27 for the GoEmotions dataset Demszky et al., 2020.

Approaches for tackling emotion classification tasks are diverse and include lexicons (Mohammad and Turney, 2013; Rabeya et al., 2017; Li et al., 2021), supervised algorithms (Jain, Kumar, and Fernandes, 2017; Becker, Moreira, and Santos, 2017, Hasan, Rundensteiner, and Agu, 2019; Asghar et al., 2019), and neural networks (Batbaatar, Li, and Ryu, 2019; Chatterjee et al., 2019; Singh, Jakhar, and Pandey, 2021).

3.2 Emotion Cause Extraction Field

Another dimension that has gained much interest recently is the task is to identify the causes of emotions.

Emotion Cause Extraction (ECE) involves finding text that can be identified as the cause of a given emotional class and was introduced in Lee et al., 2010. The task later evolved into Emotion Cause Pair Extraction (ECPE) Xia and Ding, 2019, which aims to extract both the emotion and its cause. The development of deep learning Su et al., 2023 and transformer models Acheampong, Nunoo-Mensah, and Chen, 2021 has also been popular for addressing emotion-cause-related tasks. Specifically, Zhan et al., 2022 learned to generate summaries to evaluate triggered emotions from text. The latest tasks have focused more on conversations rather than simple text, as highlighted by Singh, Caragea, and Li, 2023.

Recently, Li et al., 2023 introduced the task of Emotion-Cause Pair Extraction in Conversations (ECPEC). Additionally, for the summer of 2024, a shared task titled SemEval-2024 Task 3: Multimodal Emotion Cause Analysis in Conversations was proposed Wang et al., 2024.

3.3 Manipulation Detection based on Emotion

Kühne and Schemer, 2015 discusses how public opinion can be framed through emotional content, further emphasizing the power of emotions in shaping perceptions. Emotional appeals play a significant role in propaganda, often utilizing techniques such as loaded language, name-calling or labeling, appeals to fear, and slogans, as classified and highlighted by Da San Martino et al., 2019.

Emotional features have gotten much attention in the task of fake news analysis (Ruffo et al., 2023; Bakir and McStay, 2017).

Most works utilize emotional features to improve classification on fake news. Among most popular solution are statistical methods (Vosoughi, Roy, and Aral, 2018), LSTM-based models(Ghanem, Rosso, and Rangel, 2020; Ghanem, Rosso, and Rangel, 2020; Giachanou, Rosso, and Crestani, 2021; Hamed, Ab Aziz, and Yaakub, 2023; Zhang et al., 2021) and transformer-based models (Ghanem et al., 2021; Zhang et al., 2021; Kolev, Weiss, and Spanakis, 2022).

Most datasets are based on Twitter (Vosoughi, Roy, and Aral, 2018; Ghanem, Rosso, and Rangel, 2020; Ghanem et al., 2021), but some also include media (Ghanem, Rosso, and Rangel, 2020; Ghanem et al., 2021; Kolev, Weiss, and Spanakis, 2022) Facebook (Giachanou, Rosso, and Crestani, 2021) and Reddit (Hamed, Ab Aziz, and Yaakub, 2023).

The more recent research has increasingly focused on influence campaigns rather than fake news.

Luceri, Boniardi, and Ferrara, 2023 and Bhaumik et al., 2023 used LLM, Llama and RoBERTa, respectively, to analyze influence campaigns on Twitter. They experimented with zero-shot, few-shot, and model fine-tuning approaches.

Tseng et al., 2024 used the SOR psychological framework to analyze news headlines, examining the emotional effects these headlines have on readers and listeners. They selected BERT as the model for their research.

Lastly, a very recent review by Liu et al., 2024 provides a comprehensive overview of misinformation detection methods that incorporate emotional detection.

3.4 LLM Impersonation

With the widespread availability of large language models (LLMs), impersonation by these models has become highly exploited in practice. Researchers have sought both quantitative and qualitative evidence to support the effectiveness of this approach. For example, Zheng, Pei, and Jurgens, 2023 analyzed 126 social roles in LLM prompting to explore how prompting with model social roles could improve performance. However, no evidence for "the most efficient" role was detected in this research. Moreover, Salewski et al., 2023 demonstrated that in-context impersonation by LLMs "can change their performance and reveal their biases."

A very recent paper by Hu and Collier, 2024 aims to provide quantitative support for the success of persona simulation through prompting for various NLP tasks. Specifically, they considered tasks such as toxic content and irony detection, offensiveness, and politeness rating. The models evaluated included GPT-4, LLaMA-2, and Tulu-2. They reported that "persona prompting introduced modest and inconsistent improvement."

3.5 Research Gap

Based on the undertaken literature review, several conclusions can be drawn:

- Most research has focused on analyzing Twitter posts or news headlines/articles.
- Emotion classification and cause identification have evolved from basic classifiers to LLM-based approaches.
- Recent research has increasingly focused on cause extraction, considering readers' emotional responses or effects.
- Emotional features have been used for fake news classification and, more recently, for influence campaign detection and analysis, as the emotional response could significantly influence public opinion formation and propaganda techniques.

- While LLM impersonation is a newer field, some evidence suggests that role descriptions can improve performance for certain tasks.

In our research, we focus on social media posts on Telegram. We aim to improve emotion and emotion trigger detection through LLM impersonation, in-context learning approaches, and fine-tuning. By exploiting demographic segmentation, we aim to enhance context-dependent trigger detection. Specifically, we seek to improve emotion detection to address influence campaign detection tasks in future research.

Chapter 4

Dataset

4.1 Data Collection and Selection

The list of Moldovan Telegram channels was created by LetsData[1] analysts team and included channels with neutral content and channels with a history of manipulative content posting (potentially, malign).

Publications were also collected with the help of LetsData Technology and were accessed through the API.

The time period for the collected publications was of one month: from 13th Match to 13th April 2024. The total number of posts was 86779.

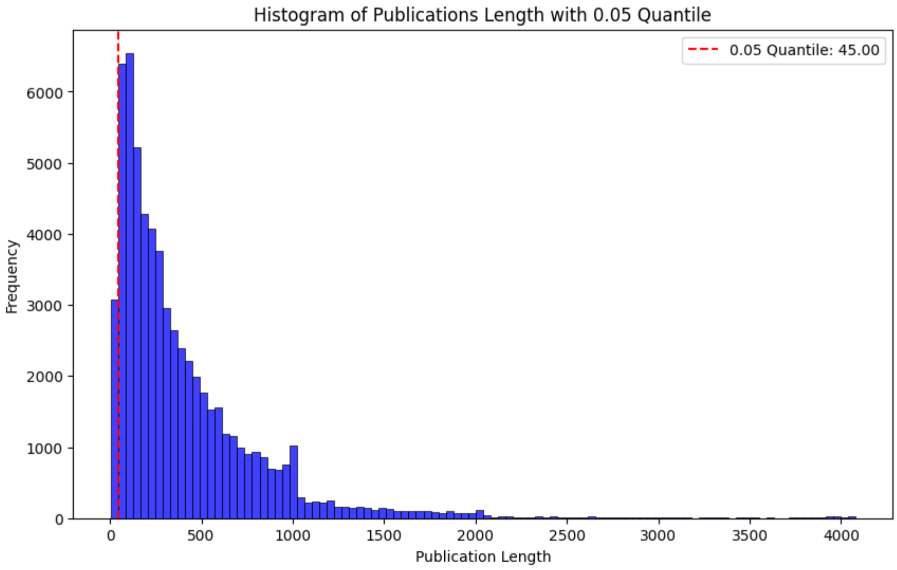


FIGURE 4.1: Posts' Text Length Distribution

Due to the limited resources for annotation, the following filtering steps were then taken:

1. Filtered out posts without the text.

2. Filtered out posts with the text that is too short. Since we focus on text analysis, we wanted to get longer posts that could contain additional interpretations or opinions. Too short was defined based on the histogram of the length of the text and is equal to the 0.05-quantile (45 chars). See figure 4.1.

3. Then, a random sample of 500 posts was selected.

[1] https://letsdata.net/

4. Finally, sample posts were manually assessed based on the question of how stand-alone the post's content was: does it require visual or video attachments to understand what the post was about?

All in all, the dataset consists of 236 publications across 68 channels. Sampled posts were both in Russian and Romanian languages.

"После слов Владимира Воронина о том, что «они (власти) лишили нас нашего родного языка — молдавского языка. Я не говорю по-румынски. Я буду говорить по-русски не только здесь, но и в парламенте, и по другим поводам». В соцсетях стартовала акция «Румынский? Нет! Говорю по-русски!». Участники акции считают, что в качестве ответной меры на решение властей Молдовы («отменить» молдавский язык и заменить его на румынский) надо массово переходить на общение на русском языке. Поддержим акцию! Сделаем перепост! #румынскийнет #нетрумынский"

TABLE 4.1: Sample Telegram Post
Translated to English: "After the words of Vladimir Voronin that "they (the authorities) deprived us of our native language - the Moldovan language. I don't speak Romanian. "I will speak Russian not only here but also in parliament and on other occasions." The campaign "Romanian? No! I speak Russian!". Participants in the action believe that, as a response to the decision of the Moldovan authorities to ("abolish" the Moldovan language and replace it with Romanian), it is necessary to massively switch to communication in Russian. Let's support the action! Let's repost! #Romanianno #non-Romanian"

4.2 Data Annotation

4.2.1 Annotators

Annotators were found through the Moldovan Community Watchdog.MD, "a think-tank based in the Republic of Moldova which builds public resilience to disinformation and manipulative narratives or stories" WatchDog.md, 2024. They helped us to find two people from the chosen demographical clusters, pro- and against-EU, who volunteered to provide us with the *ground truth* labels for our research. All the personal data was excluded from a final public dataset, and each annotator was anonymized and assigned a unique ID.

4.2.2 Labeling Setting and Instructions

For the process of annotations collection, the Labelbox Annotation tool was used *Labelbox* 2024. All of the selected publications were uploaded to the platform, and each of the annotators was asked to assess each publication on 5 separate sub-tasks:

- **Identify Emotion**: Select the emotion that best fits the main emotion that is being triggered from that text. Options include anger, fear, joy, love, sadness, surprise, and undefined.

- **Identify Emotion Class**: Select the emotion class for selected emotion. Options are positive or negative.

- **Assess Intensity**: Rate the intensity of the emotional trigger from 1 (weak) to 5 (strong).

- **Highlight Words/Phrases**: Highlight specific words or phrases or parts of the sentences that trigger the chosen emotion.
- **Explain Choice**: Provide a brief explanation (1-2 sentences) for why the selected phrases elicited the identified emotion.

The average total annotation time for a sample dataset equals 5h 19m, with an average of 1m 21s dedicated to each publication assessment.

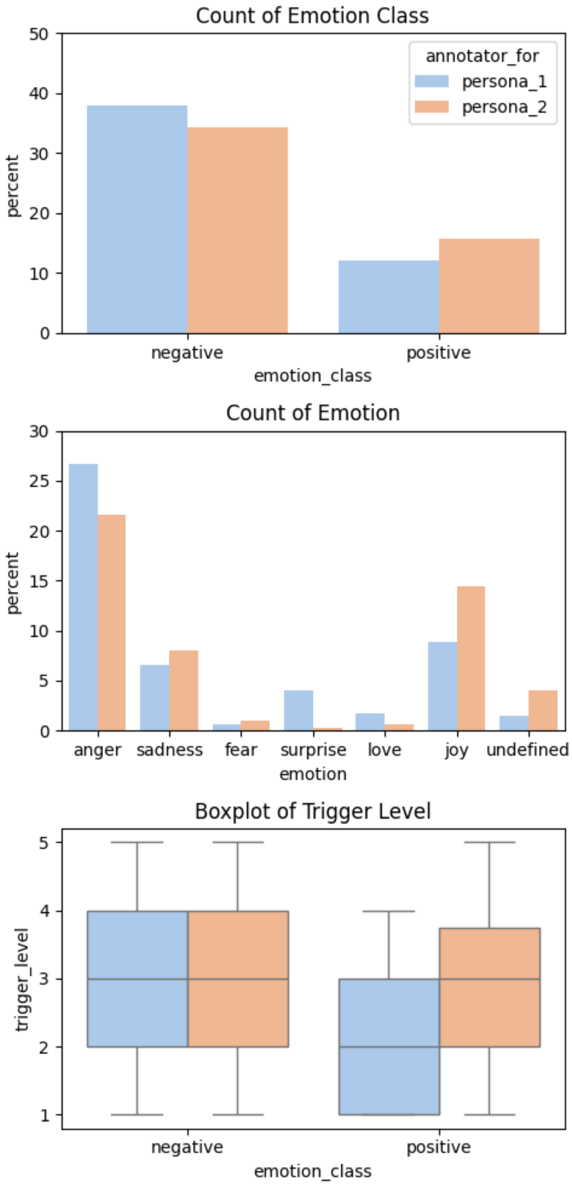


FIGURE 4.2: EDA for Annotated Data

4.2.3 EDA

The annotation agreement for the defined sub-tasks was small. Cohen's Kappa was equal to 0.02, -0.8, and 0.12 for emotion, class, and trigger level, respectively. The negative emotion class prevailed for both annotators, with anger as the dominant emotion. Trigger level was normally distributed for negative emotion for both annotators, while there was a smaller positive emotion trigger for one of the annotators (see 4.2).

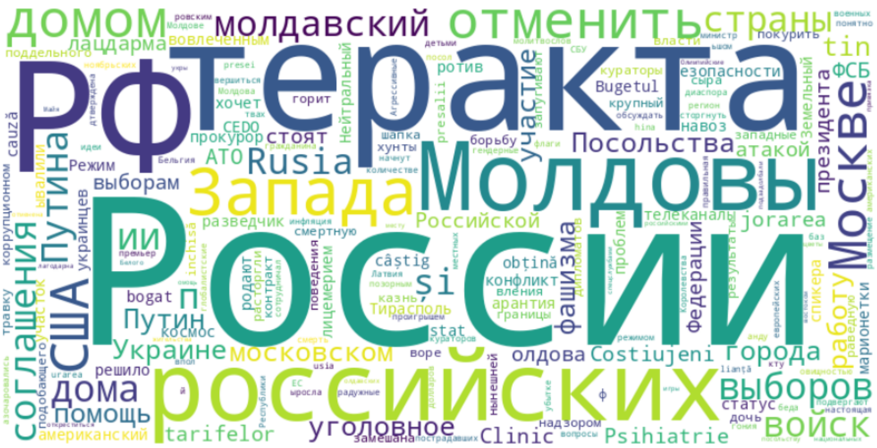


FIGURE 4.3: Trigger Spans Overlap

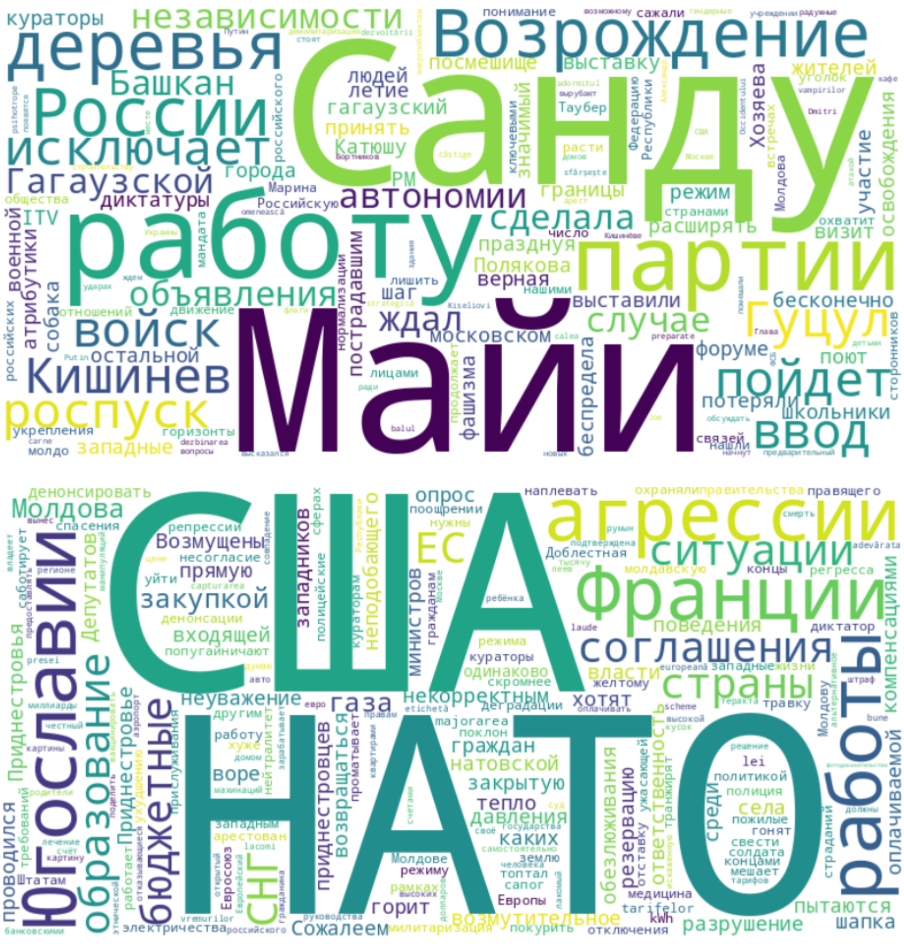


FIGURE 4.4: Biggest Triggering Spans by Persona. Top - Persona 1, Bottom - Persona 2

For triggering text spans, the Jaccard index based on the overlap of annotations was equal to 0.14. This also shows that annotators identified different text parts, causing the labeled emotion. Moreover, there was a significant difference in the length of the highlighted spans, with 49 chars for annotator for Persona 1 and 20 chars for Persona 2 highlighted per item.

Figure 4.3 shows the overlapping text spans in the original languages, Russian and Romanian. Among the top triggers are the words: *Russia/Russian*, *terrorist attack*, *Moldova*, *West*, *USA*, *cancel*, *elections*.

Figure 4.4 shows the biggest negative triggers by Personas. We defined the biggest triggers as those classified within the negative class and having a trigger level equal to 5. For Persona 1, the trigger phrases include *Maia Sandu*, *work*, *party*, and *Renaissance (party)*; for Persona 2, the phrases are *USA*, *NATO*, *Yugoslavia*, *France*, and *aggression*. In both cases, named entities constitute a significant part of the biggest triggers.

The outline annotator (dis)agreement shows that the research hypothesis about the difference in elicited emotional triggers is significant.

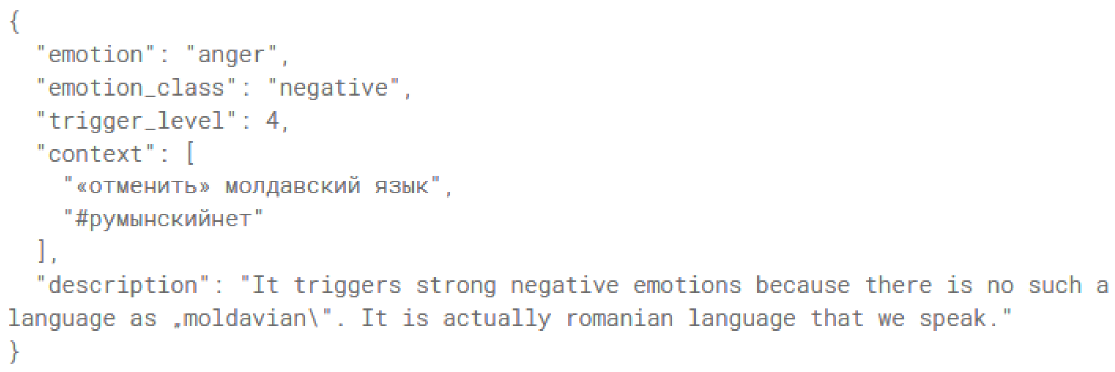
4.3 Final Dataset

4.3.1 Train-Validation Split

A sample dataset was divided into 2 main groups: a subset for training of the models and a subset for validation of models, 30% and 70% of the data, respectively. W Moreover, the split was done based on the published date of the post to minimize data leakage for the same event and/or post topic in training and validation datasets.

4.3.2 Output Data Format

Following the described methodology and the format of the data collected, we defined an output structure for an analysis of a single post.



```
{
  "emotion": "anger",
  "emotion_class": "negative",
  "trigger_level": 4,
  "context": [
    "«отменить» молдавский язык",
    "#румынскийнет"
  ],
  "description": "It triggers strong negative emotions because there is no such a
language as „moldavian\". It is actually romanian language that we speak."
}
```

FIGURE 4.5: Example Output JSON

Chapter 5

Experiments and Results

5.1 Experiments Setting

We selected several approaches for our task:

1. No context (NC)[1] - running LLM inference without s context defined in the prompt. This can be used as a *baseline* for the assessment of the LLM in our task to compare with -based approaches;
2. Zero-shot Learning (ZSH) - running LLM inference with s context (defined in Table **??**) defined in the prompt;
3. Few-shot learning (FSL) - running LLM inference with s context defined in the prompt and few shot examples from ground truth annotations for a set of publications. Two sets of examples were selected, 5 and 10, to evaluate if the number of examples could affect results;
4. Fine tuning (FT) - running LLM inference with s context defined in the prompt and fine-tuning the model based on the ground truth annotations for a set of publications.

5.2 LLM Setting

5.2.1 Selection

The base model for our experiments was GPT-3-turbo from OpenAI. At the time of our research, the latest model, GPT-4, was not available for fine-tuning. Consequently, we could not use it for comparing the results of different approaches. Nevertheless, utilizing OpenAI models offered several advantages for implementing our methodologies.

Firstly, OpenAI provides a fine-tuning API that allows fine-tuning on their cloud with minimal resource requirements. Secondly, this process demands relatively little data, which was ideal given our limited dataset. Additionally, we selected GPT-3-turbo-1106, a model optimized for working with JSON output data and available for fine-tuning.

The low cost of chat completions with GPT-3 was also beneficial for our few-shot learning examples, as increasing the prompt size did not significantly impact costs.

[1] "You are a helpful AI assistant" - context used in prompt instead of the context.

5.2.2 Learning

Two sets of examples were selected for few-shot learning: sizes 5 and 10. Both sets included a variety of emotional trigger examples to enhance the in-context learning process and avoid bias towards particular emotions, trigger levels, etc.

For fine-tuning, we used 50 samples for training and 20 for validation, running the process for 4 epochs with a batch size of 1. The hyperparameters were based on general recommendations from the OpenAI documentation (OpenAI, 2024). In the final inference, the model checkpoint from the 3rd epoch was selected due to the assessment of the learning curves and the balance between training and validation loss. Learning curves for the fine-tuning process can be found in Appendix A.1 - A.4.

5.2.3 Inference

Final prompt that was used for inference is illustrated in Table 5.1. It depends on two input variables:

- {CONTEXT} - to be used for inference (one of **??**).
- {PUBLICATION} - publication text to be assessed for emotional triggers.

Role	**Content**
system	{CONTEXT} ###Instruction:### After reading a provided publication, evaluate the emotional impact it has on you. Your assessment should be structured as a JSON response encompassing the following aspects: - **emotion:** Specify the exact one main emotion elicited by the publication. Options: anger, fear, joy, love, sadness, and surprise. - **emotion_class:** Categorize the sentiment of the emotion as either negative or positive. - **trigger_level:** Rate the intensity of the emotional trigger on a scale from 1 (weak) to 5 (strong). - **context:** Highlight specific words or phrases from the text that support your identified emotion. Return a list of strings. - **description:** Offer a brief explanation of why the publication triggered this emotion in you, referencing your unique perspective of your identity. Explain in at most 100 words. In cases where you're uncertain about the emotion evoked, the emotion key should contain "undefined". Note, that publications can be in Russian or Romanian languages, but output should be in English. ###Input text indicator:### <PUBLICATION> ### Output indicator:### JSON response with keys: emotion, emotion_class, trigger_level, context, description.
user	<PUBLICATION>: “‘{PUBLICATION}“‘

TABLE 5.1: Prompt for Emotional Triggers Detection

The length of the final prompt containing the 's context did not exceed 400 tokens in both cases.

Inference was run for multiple random seed parameters, and a temperature of 0.2 was used across the run pipeline to ensure more focus and determinism in the output.

5.3 Evaluation Metrics

For evaluation metrics, precision, recall, and F1-score were used to evaluate emotion, emotion class, and trigger level.

In particular, for emotion and trigger level weighted versions of these metrics were used (see Equations 5.1-5.3):

$$\text{Recall (weighted)} = \frac{\sum_{i=1}^{C} N_i \cdot R_i}{N} \tag{5.1}$$

$$\text{Precision (weighted)} = \frac{\sum_{i=1}^{C} N_i \cdot P_i}{N} \tag{5.2}$$

$$\text{F1-score (weighted)} = \frac{\sum_{i=1}^{C} N_i \cdot F1_i}{N} \tag{5.3}$$

For triggering phrase detection, two types of metrics were selected based on the recent SemEval-2024 Task (Wang et al., 2024). These are:

- strict - exact match of annotated and predicted spans;
- proportional - overlap proportion between the predicted and annotated spans, see Equations 5.4-5.3).

$$\text{Recall} = \frac{\sum \sum_i overlap_i}{\sum \sum_i len(as_i)}, \tag{5.4}$$

$$\text{Precision} = \frac{\sum \sum_i overlap_i}{\sum \sum_i len(as_i)}, \tag{5.5}$$

$$\text{F1-score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5.6}$$

5.4 Results

The summary of the results is presented in Table 5.2. Metrics were averaged across two s and three random seeds selected for the inference runs.

It is evident that for all the emotion trigger variables, the fine-tuned approach to LLM learning outperformed the baseline, which had no context, showing significant improvement.

Moreover, the fine-tuned approach demonstrated the lowest failure rate among all methods (see Table 5.3). An inference was considered a failure if it produced a corrupted JSON or generated values that did not fall within the range of defined options in the analysis model and written prompt.

In addition, the descriptions generated by the fine-tuned model were manually assessed, revealing that the reasoning behind the predicted emotional triggers for two individuals differed according to the context of each . See the example in Table 5.4.

Summary results for each are provided in Appendix: Table A.2 and Table A.3.

Variable	Approach	Precision	Recall	F1-Score
emotion	NC	0.513 ± 0.06	0.438 ± 0.02	0.464 ± 0.02
	ZSL	0.565 ± 0.22	0.513 ± 0.02	0.508 ± 0.05
	FSL-5	**0.575 ± 0.22**	0.484 ± 0.06	0.509 ± 0.1
	FSL-10	0.535 ± 0.15	0.52 ± 0.06	0.505 ± 0.06
	FT	0.52 ± 0.16	**0.579 ± 0.04**	**0.538 ± 0.1**
emotion_class	NC	0.724 ± 0.14	0.703 ± 0.11	0.713 ± 0.12
	ZSL	0.796 ± 0.07	0.773 ± 0.02	0.764 ± 0.01
	FSL-5	0.768 ± 0.07	0.709 ± 0.05	0.736 ± 0.03
	FSL-10	0.783 ± 0.06	0.766 ± 0.05	0.771 ± 0.05
	FT	**0.838 ± 0.08**	**0.831 ± 0.04**	**0.826 ± 0.07**
trigger_level	NC	0.217 ± 0.22	0.225 ± 0.05	0.129 ± 0.03
	ZSL	0.191 ± 0.28	0.218 ± 0.02	0.115 ± 0.05
	FSL-5	**0.298 ± 0.19**	0.228 ± 0.2	0.199 ± 0.21
	FSL-10	0.228 ± 0.19	0.193 ± 0.04	0.178 ± 0.07
	FT	0.229 ± 0.21	**0.29 ± 0.06**	**0.233 ± 0.1**
context (strict)	NC	0.001 ± 0.0	0.077 ± 0.03	0.002 ± 0.0
	ZSL	0.001 ± 0.0	0.084 ± 0.04	0.002 ± 0.0
	FSL-5	0.001 ± 0.0	0.091 ± 0.02	0.002 ± 0.0
	FSL-10	0.001 ± 0.0	0.094 ± 0.04	0.002 ± 0.0
	FT	**0.002 ± 0.0**	**0.122 ± 0.1**	**0.003 ± 0.0**
context (proportional)	NC	0.095 ± 0.03	0.321 ± 0.15	0.145 ± 0.03
	ZSL	0.179 ± 0.04	0.405 ± 0.19	0.246 ± 0.02
	FSL-5	0.238 ± 0.12	0.394 ± 0.15	0.292 ± 0.05
	FSL-10	0.253 ± 0.13	0.414 ± 0.14	0.309 ± 0.07
	FT	**0.255 ± 0.05**	**0.446 ± 0.08**	**0.324 ± 0.06**

TABLE 5.2: Results Summary

Approach	Failure Rate
NC	0.076 ± 0.02
ZSL	0.065 ± 0.08
FSL-5	0.083 ± 0.13
FSL-10	0.033 ± 0.02
FT	**0.014 ± 0.04**

TABLE 5.3: Comparison of Failure Rates

	Example Description	Emotion	Trigger Level
1	The Russian Federation is trying to manipulate the international community by accusing NATO of aggression against Yugoslavia in 1999. The Russian Federation is trying to divert attention from its own actions and to present itself as a victim, which is outrageous.	anger	3
2	The publication triggered a strong feeling of anger in me. The disrespectful and unprofessional behavior of the French delegation towards the commemoration of the NATO aggression against Yugoslavia, as well as their disregard for the Serbian representative, deeply upset me. As someone who values historical respect and protocol, I find such actions intolerable and disrespectful.	anger	4

TABLE 5.4: Examples of Generated Descriptions

5.5 Evaluation

Some evaluation points of the presented results should be considered.

Firstly, although the F1-score was used as the main evaluation criterion, there are some limitations that should be acknowledged. We assume that for the F1-score, precision and recall are of equal importance. However, when analyzing incidents or manipulations, precision might be more critical due to the significant reputation risks associated with mislabeling. Notably, precision in predicting emotion and trigger levels showed better results with few-shot learning using five examples.

Next, a weighted version of the F1-score is beneficial in datasets with imbalance, as is the case with our data. However, this method weights all classes equally. In our model, there is an "undefined" class based more on the subjective inability of the annotator to define an emotion, which might not be the best approach to weight equally with other classes.

Furthermore, the detection of context phrases performed poorly under the strict evaluation metric. As seen in the exploratory data analysis of the dataset, the spans highlighted by different s varied significantly in text length. This discrepancy makes it challenging for the model to define these spans accurately, which might be addressed by implementing additional guidelines specific to this case.

Moreover, while trigger levels were defined as an ordinal variable, incorporating a continuous metric might be more suitable for LLM inference and evaluation against the true labels.

Finally, each cluster depended on annotations from a single persona, assessing the model's ability to learn that particular persona's triggers. Therefore, employing several annotators for each Persona could help define the true labels better.

5.6 Proposed System Description

Based on the research undertaken, we have defined an active-learning system for detecting context-based emotional triggers for a set of targeted Personas.

System prerequisites:

1. Continuous data collection from the targeted domain (in our case, Moldovan Telegram channels);
2. Selection of demographic clusters of interest;
3. Selection of annotators for a defined set of demographic clusters;
4. Generation of representative Persona contexts for selected clusters.

System iteration *i*:

1. Sampling a representative sample of publications from the most recent time period, *N*;
2. Collection of labels for emotional triggers - train and validation;
3. LLM fine-tuning based on the collected labels;
4. LLM evaluation on the validation set. In case of poor performance, collect more labels and repeat steps 2 and 3;
5. LLM inference on the continuous data for the time period *K*;
6. Inference analysis and interventions.

In our research, we partially implemented the first iteration of the outlined system (steps 1-4). For future development, we will require continuous collaboration with the annotators to collect the ground truth labels for the most recent publications.

Chapter 6

Conclusions

6.1 Conclusions

The research has shown that the incorporation of Personas' simulation can help to achieve improvement in the task of detection of context-dependent emotional triggers when compared to the popular exploitation of the role of a "helpful assistant." An approach for fine-tuning LLM has given the best F1-score results in all of the components of the set task.

6.2 Limitations and Future Work

The main limitation of the outlined research is the size of the dataset and the number of annotators that labeled the sample of publications. Getting a bigger sample of annotations will help to improve the quality of ground truth labels for fine-tuning and the general assessment of the approaches. In particular, this would help to define better target labels for triggering span detection. We have shown that there has already been a great difference in the length of spans annotated between different annotators, which would create confusion during the fine-tuning of the models with training examples for single Persona that are produced by different annotators.

Also, future work should include the assessment of the training data and LLM "knowledge" around the targeted demographics of Moldova. This could be done through a more thorough assessment of the descriptions the LLM generates to the predicted emotion and trigger spans.

What is more, we want to expand the number of Personas and, in this way, demographic coverage of different population groups. For the particular use-case of possible manipulations in upcoming Moldovan elections, it would be better to get a representative Personas for a more moderate category: a segment of those who are undecided and could be affected by campaigns of influence. In contrast, the assessed Personas are already defined as pro- and against-EU and their opinion might be much harder to influence by any campaigns.

In research, we only focus on the text emotion triggers detection while images and videos are often more common to be shared in social media. A system for a multimodal approach can bring more insights about the present triggers.

Finally, we aim to expand this approach to more countries and contexts to compare what would work worse or better in that dimension.

6.3 Responsible AI Statement

The research paper aims to detect emotional triggers for specific demographic groups to help analyze and understand their sensitive topics. This approach is intended to identify and prevent potential manipulations by malign actors. We condemn any misuse of this approach to exploit these triggers.

Additionally, we acknowledge that some characteristics of the Personas generated may appear stereotypical or biased toward the demographic cluster descriptions.

Lastly, we ensure that all of the disclosed as a part of this research will not contain any personal information.

Appendix A

Appendices

A.1 Code

All the relevant code can be found at https://github.com/LetsData-net/emotional-triggers

A.2 List of Telegram Channels in Publications Subset

TABLE A.1: List of Telegram Channels in Publications Subset

@Accent_TV	@CIS91	@KpMoldova
@MOLDOVA_L	@Marina_Tauber	@MoldovaPolitics
@Moldova_20	@MonitoringMD	@Noi_md
@PMRlesnik	@Partidul_Politic_SOR	@PointNews
@PravdaGagauzia	@Primaria_Chisinau	@RepublicaUnirii
@RusEmbMd	@agoramd	@aifmd
@antimaidanmd	@bananatomato	@covid19_moldova
@dimoglonina	@dvijeniemd	@enewsmd
@eurasiamoldova	@europaliberamd	@gabrielcalin
@gagauznewsmd	@gardatinaraRM	@grajdaninmoldovi
@gubernator_pmr	@ilanshor	@indexMD
@ivanovnamd	@jukov_online	@kotletipmr
@latebuimistru	@md_krot	@mdsputnikmd
@moldavskii_piston	@moldova_acum	@moldovaelections
@moldovalibera	@moldovarcnk	@moldovatelegraph
@multumesc_moldova	@newsmd24	@np_inform
@onlinemd24	@pandorapmr	@partidul_renastere
@pridnestrovec	@racumd	@ro_newsmakerlive
@romania_ru	@rupor_md	@rusputnikmd
@salutmd	@sibmd	@smuglianka
@tirdea	@tokanamd	@tricolorthednestr
@triunghiulbasarabean	@tsvtiraspol	@tv8md
@wtfmoldova	@ybp_mdru	

A.3 Learning Curves for Fine-Tuning

The visualizations show the dynamics of the learning process of GPT-3.5-turbo-1106 by two PERSONs models. It also shows the checkpoint of 150 steps or 3 epochs that was used for the final models. The checkpoint was selected based on the evaluation of the results of training and validation together.

- ftjob-RAfwPojEG6os9QLGdRXt0d28 - model for PERSONA 1;
- ftjob-QOnkjWy33XYknnmyUHPNCbA1 - model for PERSONA 2.

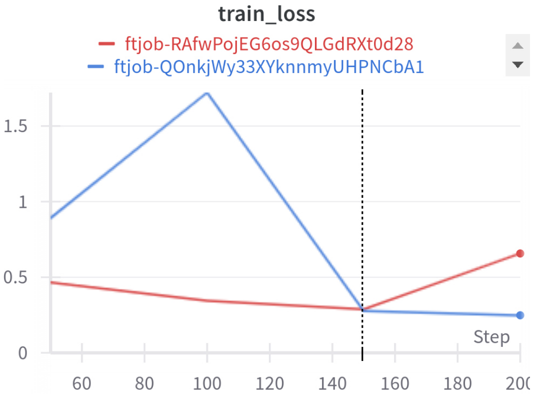


FIGURE A.1: Train Loss

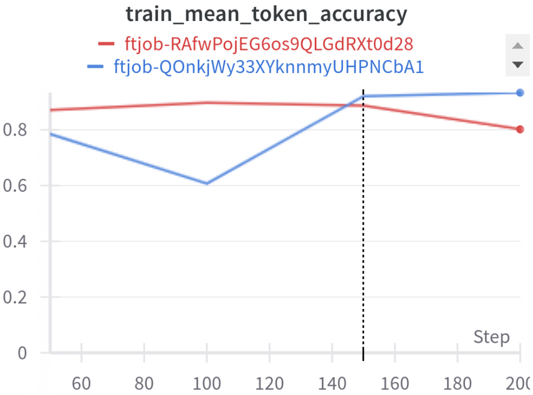


FIGURE A.2: Train Mean Token Accuracy

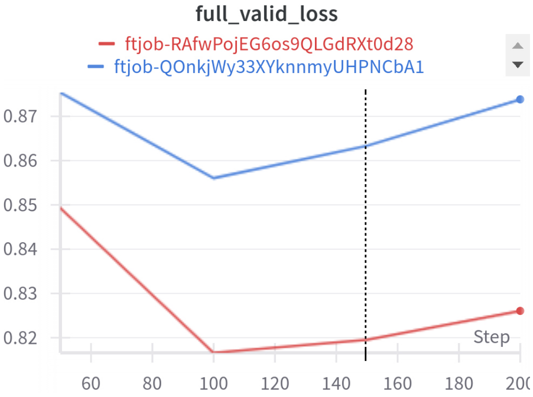


FIGURE A.3: Full Validation Loss

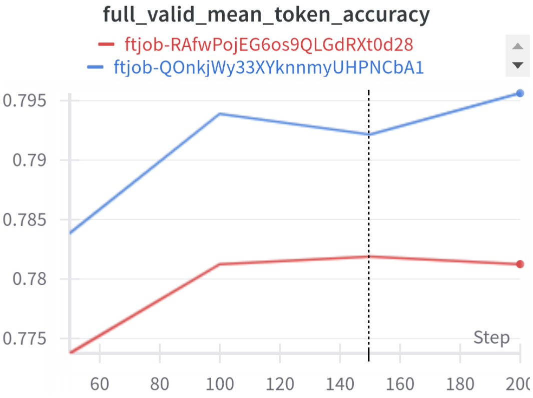


FIGURE A.4: Full Validation Mean Token Accuracy

A.4 Results Summary by PERSONA

TABLE A.2: Results Summary for Persona 1

Variable	Approach	Precision	Recall	F1-Score
emotion	NC	0.495 ± 0.02	0.442 ± 0.02	0.461 ± 0.02
	ZSL	0.496 ± 0.02	0.51 ± 0.03	0.493 ± 0.03
	FSL-5	**0.507 ± 0.03**	0.468 ± 0.02	0.478 ± 0.02
	FSL-10	0.491 ± 0.03	0.504 ± 0.03	0.488 ± 0.02
	FT	0.472 ± 0.02	**0.57 ± 0.03**	**0.507 ± 0.02**
emotion class	NC	0.766 ± 0.0	0.735 ± 0.02	0.75 ± 0.01
	ZSL	0.775 ± 0.02	0.771 ± 0.02	0.765 ± 0.01
	FSL-5	0.748 ± 0.03	0.719 ± 0.03	0.733 ± 0.03
	FSL-10	0.764 ± 0.03	0.753 ± 0.03	0.757 ± 0.03
	FT	**0.814 ± 0.02**	**0.819 ± 0.02**	**0.803 ± 0.01**
trigger level	NC	0.16 ± 0.04	0.239 ± 0.01	0.122 ± 0.0
	ZSL	0.119 ± 0.19	0.219 ± 0.01	0.105 ± 0.01
	FSL-5	0.245 ± 0.1	0.169 ± 0.03	0.137 ± 0.02
	FSL-10	0.279 ± 0.12	0.187 ± 0.05	0.192 ± 0.07
	FT	**0.292 ± 0.05**	**0.275 ± 0.03**	**0.263 ± 0.04**
context (strict)	NC	0.001 ± 0.0	0.086 ± 0.01	0.002 ± 0.0
	ZSL	0.001 ± 0.0	**0.095 ± 0.02**	0.002 ± 0.0
	FSL-5	0.001 ± 0.0	0.09 ± 0.03	0.002 ± 0.0
	FSL-10	0.001 ± 0.0	0.084 ± 0.04	0.002 ± 0.0
	FT	0.001 ± 0.0	0.092 ± 0.04	0.002 ± 0.0
context (proportional)	NC	0.102 ± 0.03	0.276 ± 0.05	0.149 ± 0.03
	ZSL	0.189 ± 0.02	0.348 ± 0.04	0.245 ± 0.03
	FSL-5	0.274 ± 0.01	0.349 ± 0.01	0.307 ± 0.01
	FSL-10	**0.291 ± 0.03**	0.373 ± 0.04	**0.327 ± 0.03**
	FT	0.242 ± 0.03	**0.424 ± 0.06**	0.308 ± 0.04

TABLE A.3: Results Summary for Persona 2

Variable	Approach	Precision	Recall	F1-Score
emotion	NC	0.53±0.01	0.434±0.02	0.466±0.02
	ZSL	0.633±0.01	0.516±0.02	0.523±0.01
	FSL-5	**0.642±0.01**	0.5±0.03	0.54±0.03
	FSL-10	0.578±0.05	0.536±0.03	0.522±0.04
	FT	0.569±0.02	**0.588±0.01**	**0.57±0.0**1
emotion_class	NC	0.682±0.01	0.671±0.03	0.676±0.02
	ZSL	0.817±0.0	0.775±0.02	0.763±0.01
	FSL-5	0.788±0.03	0.699±0.05	0.739±0.04
	FSL-10	0.801±0.01	0.779±0.01	0.784±0.01
	FT	**0.863±0.01**	**0.843**±0.02	**0.849±0.0**
trigger_level	NC	0.273±0.18	0.211±0.03	0.136±0.04
	ZSL	0.263±0.15	0.217±0.03	0.125±0.06
	FSL-5	**0.352±0.03**	0.287±0.03	**0.261±0.04**
	FSL-10	0.176±0.03	0.199±0.03	0.163±0.02
	FT	0.167±0.04	**0.305±0.04**	0.204±0.03
context (strict)	NC	0.001±0.0	0.068±0.02	0.002±0.0
	ZSL	0.001±0.0	0.074±0.03	0.002±0.0
	FSL-5	0.001±0.0	0.092±0.01	0.003±0.0
	FSL-10	0.002±0.0	0.104±0.03	0.003±0.0
	FT	0.002±0.0	**0.151±0.04**	**0.005±0.0**
context (proportional)	NC	0.087±0.01	0.365±0.03	0.141±0.02
	ZSL	0.168±0.02	0.461±0.05	0.246±0.03
	FSL-5	0.202±0.02	0.44±0.03	0.276±0.02
	FSL-10	0.214±0.03	0.455±0.05	0.291±0.04
	FT	**0.268±0.01**	**0.468±0.03**	**0.341±0.01**

Bibliography

Acheampong, Francisca Adoma, Henry Nunoo-Mensah, and Wenyu Chen (2021). "Transformer Models for Text-Based Emotion Detection: A Review of BERT-Based Approaches". In: *Artificial Intelligence Review* 54.8, pp. 5789–5829. DOI: 10.1007/s10462-021-09958-2. URL: https://doi.org/10.1007/s10462-021-09958-2.

Aggarwal, Charu C. (2018). "Opinion Mining and Sentiment Analysis". In: *Machine Learning for Text*. Ed. by Charu C. Aggarwal. Cham: Springer International Publishing, pp. 413–434. DOI: 10.1007/978-3-319-73531-3_13. URL: https://doi.org/10.1007/978-3-319-73531-3_13.

Asghar, Muhammad Zaheer et al. (2019). "Performance Evaluation of Supervised Machine Learning Techniques for Efficient Detection of Emotions from Online Content". In: DOI: 10.20944/preprints201908.0019.v1. URL: https://doi.org/10.20944/preprints201908.0019.v1.

Bakir, Vian and Andrew McStay (2017). "Fake News and The Economy of Emotions: Problems, causes, solutions". In: *Digital Journalism* 6, pp. 1–22. DOI: 10.1080/21670811.2017.1345645. URL: https://doi.org/10.1080/21670811.2017.1345645.

Batbaatar, Enkhbold, Maoguo Li, and Keehoon Ryu (2019). "Semantic-Emotion Neural Network for Emotion Recognition From Text". In: *IEEE Access* 7, pp. 111866–111878. DOI: 10.1109/ACCESS.2019.2934529. URL: https://doi.org/10.1109/ACCESS.2019.2934529.

Becker, Karina, Viviane P. Moreira, and Ana G. L. dos Santos (2017). "Multilingual emotion classification using supervised learning: Comparative experiments". In: *Information Processing & Management* 53, pp. 684–704. DOI: 10.1016/j.ipm.2016.12.008. URL: https://doi.org/10.1016/j.ipm.2016.12.008.

Bhaumik, Ankita et al. (2023). "Adapting Emotion Detection to Analyze Influence Campaigns on Social Media". In: *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*. Ed. by Jeremy Barnes, Orphée De Clercq, and Roman Klinger. Association for Computational Linguistics. Toronto, Canada, pp. 441–451. DOI: 10.18653/v1/2023.wassa-1.38. URL: https://doi.org/10.18653/v1/2023.wassa-1.38.

Chatterjee, Abhijit et al. (2019). "Understanding Emotions in Text Using Deep Learning and Big Data". In: *Computers in Human Behavior* 93, pp. 309–317. DOI: 10.1016/j.chb.2018.12.029. URL: https://doi.org/10.1016/j.chb.2018.12.029.

Da San Martino, Giovanni et al. (2019). "Fine-Grained Analysis of Propaganda in News Article". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 5635–5645. DOI: 10.18653/v1/d19-1565. URL: https://doi.org/10.18653/v1/d19-1565.

Demszky, Dorottya et al. (2020). "GoEmotions: A Dataset of Fine-Grained Emotions". In.

Devika, M. Devi, C. Sunitha, and A. Ganesh (2016). "Sentiment Analysis: A Comparative Study on Different Approaches". In: *Procedia Computer Science* 87. Fourth International Conference on Recent Trends in Computer Science & Engineering

(ICRTCSE 2016), pp. 44–49. DOI: 10.1016/j.procs.2016.05.124. URL: https://doi.org/10.1016/j.procs.2016.05.124.

Eady, Gregory et al. (2019). "How Many People Live in Political Bubbles on Social Media? Evidence From Linked Survey and Twitter Data". In: *Sage Open* 9.1. DOI: 10.1177/2158244019832705. URL: https://doi.org/10.1177/2158244019832705.

(EEAS), European External Action Service (2021). *Report on Stratcom Activities 2021*. URL: https://www.eeas.europa.eu/sites/default/files/documents/Report%20Stratcom%20activities%202021.pdf.

— (2023). *EEAS Data Team Threat Report 2023*. URL: https://www.eeas.europa.eu/sites/default/files/documents/2023/EEAS-DataTeam-ThreatReport-2023.pdf.

— (2024). *2nd Report on FIMI Threats - January 2024*. URL: https://www.eeas.europa.eu/sites/default/files/documents/2024/EEAS-2nd-Report%20on%20FIMI%20Threats-January-2024_0.pdf.

Ekman, Paul (1992). "An argument for basic emotions". In: *Cognition and Emotion* 6.3-4, pp. 169–200. DOI: 10.1080/02699939208411068. URL: https://doi.org/10.1080/02699939208411068.

Ghanem, Bilal, Paolo Rosso, and Francisco Rangel (2020). "An Emotional Analysis of False Information in Social Media and News Articles". In: *ACM Transactions on Internet Technology* 20.2, pp. 1–18. DOI: 10.1145/3381750. URL: https://doi.org/10.1145/3381750.

Ghanem, Bilal et al. (2021). "FakeFlow: Fake News Detection by Modeling the Flow of Affective Information". In.

Giachanou, Anastasia, Paolo Rosso, and Fabio Crestani (2021). "The impact of emotional signals on credibility assessment". In: *Journal of the Association for Information Science and Technology* 72, pp. 1117–1132. DOI: 10.1002/asi.24480. URL: https://doi.org/10.1002/asi.24480.

Hamed, Shadi Kamal, Mohammad Jusoh Ab Aziz, and Mohd Rizal Yaakub (2023). "Fake News Detection Model on Social Media by Leveraging Sentiment Analysis of News Content and Emotion Analysis of Users' Comments". In: *Sensors* 23.4, p. 1748. DOI: 10.3390/s23041748. URL: https://doi.org/10.3390/s23041748.

Hasan, Mahbub, Elke Rundensteiner, and Emmanuel Agu (2019). "Automatic emotion detection in text streams by analyzing Twitter data". In: *International Journal of Data Science and Analytics* 7, pp. 35–51. DOI: 10.1007/s41060-018-0096-z. URL: https://doi.org/10.1007/s41060-018-0096-z.

Hu, Tingsong and Nigel Collier (2024). "Quantifying the Persona Effect in LLM Simulations". In.

Jain, Vishal K., S. Kumar, and S. L. Fernandes (2017). "Extraction of emotions from multilingual text using intelligent text processing and computational linguistics". In: *Journal of Computational Science* 21, pp. 316–326. DOI: 10.1016/j.jocs.2017.01.010. URL: https://doi.org/10.1016/j.jocs.2017.01.010.

Kolev, Vladislav, Gerhard Weiss, and Gerasimos Spanakis (2022). "FOREAL: RoBERTa Model for Fake News Detection Based on Emotions". In: *Proceedings of the 14th International Conference on Agents and Artificial Intelligence*. SCITEPRESS - Science and Technology Publications, pp. 429–440. DOI: 10.5220/0010873900003116. URL: https://doi.org/10.5220/0010873900003116.

Kühne, Rinaldo and Christian Schemer (2015). "The Emotional Effects of News Frames on Information Processing and Opinion Formation". In: *Communication Research* 42.3, pp. 387–407. DOI: 10.1177/0093650213514599. URL: https://doi.org/10.1177/0093650213514599.

Labelbox (2024). https://labelbox.com/. Accessed: 2024-05-05.

Lee, S.Y.M. et al. (2010). “Emotion Cause Events: Corpus Construction and Analysis”. In.

Li, Wei et al. (2023). “ECPEC: Emotion-Cause Pair Extraction in Conversations”. In: *IEEE Transactions on Affective Computing* 14.3, pp. 1754–1765. DOI: 10.1109/TAFFC.2022.3216551.

Li, Zhen et al. (2021). “Word-level emotion distribution with two schemas for short text emotion classification”. In: *Knowledge-Based Systems* 227, p. 107163. DOI: 10.1016/j.knosys.2021.107163. URL: https://doi.org/10.1016/j.knosys.2021.107163.

Liu, Zhiwei et al. (2024). “Emotion Detection for Misinformation: A Review”. In: *Information Fusion* 107, p. 102300. DOI: 10.1016/j.inffus.2024.102300. URL: https://doi.org/10.1016/j.inffus.2024.102300.

Luceri, Luca, Eric Boniardi, and Emilio Ferrara (2023). “Leveraging Large Language Models to Detect Influence Campaigns in Social Media”. In: *arXiv*. URL: https://doi.org/10.48550/ARXIV.2311.07816.

Mohammad, Saif M. and Peter D. Turney (2013). *Crowdsourcing a Word-Emotion Association Lexicon*. URL: https://doi.org/10.48550/arXiv.1308.6297.

Mondal, Avishek and Swapna Gokhale (2020). “Mining Emotions on Plutchik’s Wheel”. In: *2020 IEEE/ACM International Conference on Social Networks Analysis and Mining (SNAMS)*. DOI: 10.1109/SNAMS52053.2020.9336534. URL: https://doi.org/10.1109/SNAMS52053.2020.9336534.

Mostafazadeh, Nasrin et al. (2016). “A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kevin Knight, Ani Nenkova, and Owen Rambow. Association for Computational Linguistics. San Diego, California, pp. 839–849. DOI: 10.18653/v1/N16-1098. URL: https://doi.org/10.18653/v1/N16-1098.

Nae, Adrian (2022). *Russian Propaganda on Telegram: Narratives Targeting the Republic of Moldova*. URL: https://russianstudiesromania.eu/2022/06/01/russian-propaganda-on-telegram-narratives-targeting-the-republic-of-moldova/.

Nandwani, Priya and R. Verma (2021). “A review on sentiment analysis and emotion detection from text”. In: *Social Network Analysis and Mining* 11.81. DOI: 10.1007/s13278-021-00776-6. URL: https://doi.org/10.1007/s13278-021-00776-6.

OHCHR (2024). *2024 elections are testing democracy’s health*. URL: https://www.ohchr.org/en/stories/2024/03/2024-elections-are-testing-democracys-health.

OpenAI (2024). *Fine-tuning Guide*. https://platform.openai.com/docs/guides/fine-tuning. Accessed: 2024-05-04.

Parrott, W. Gerrod (2001). *Emotions in Social Psychology*. Key Readings in Social Psychology. Philadelphia: Psychology Press. ISBN: 978-0863776830.

— ed. (2014). *The Positive Side of Negative Emotions*. Guilford Publications.

Plutchik, Robert (2003). *Emotions and Life: Perspectives from Psychology, Biology, and Evolution*. Washington, DC: American Psychological Association.

Rabeya, Tasnia et al. (2017). “A survey on emotion detection: A lexicon based backtracking approach for detecting emotion from Bengali text”. In: *2017 20th International Conference of Computer and Information Technology (ICCIT)*, pp. 1–7. DOI: 10.1109/ICCITECHN.2017.8281855. URL: https://doi.org/10.1109/ICCITECHN.2017.8281855.

Ruffo, Giancarlo et al. (2023). “Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language”. In: *Computer Science Review* 47, p. 100531. DOI: 10.1016/j.cosrev.2022.100531. URL: https://doi.org/10.1016/j.cosrev.2022.100531.

Salewski, Lasse et al. (2023). “In-Context Impersonation Reveals Large Language Models’ Strengths and Biases”. In: *arXiv*. URL: https://doi.org/10.48550/ARXIV.2305.14930.

Saravia, Elvis et al. (2018). “CARER: Contextualized Affect Representations for Emotion Recognition”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff et al. Association for Computational Linguistics. Brussels, Belgium, pp. 3687–3697. DOI: 10.18653/v1/D18-1404. URL: https://doi.org/10.18653/v1/D18-1404.

Shaver, Phillip et al. (1987). “Emotion knowledge: further exploration of a prototype approach”. In: *Journal of Personality and Social Psychology* 52.6, pp. 1061–1086. DOI: 10.1037/0022-3514.52.6.1061.

SimilarWeb (2024). *Worldwide Messaging Apps Market Research*. URL: https://www.similarweb.com/blog/research/market-research/worldwide-messaging-apps/.

Singh, Manvi, Anil Kumar Jakhar, and Sandeep Pandey (2021). “Sentiment analysis on the impact of coronavirus in social life using the BERT model”. In: *Social Network Analysis and Mining* 11.33. DOI: 10.1007/s13278-021-00737-z. URL: https://doi.org/10.1007/s13278-021-00737-z.

Singh, Smriti, Cornelia Caragea, and Junyi Jessy Li (2023). “Language Models (Mostly) Do Not Consider Emotion Triggers When Predicting Emotion”. In: *arXiv*. URL: http://arxiv.org/abs/2311.09602.

Su, Xinxin et al. (2023). “Recent Trends in Deep Learning Based Textual Emotion Cause Extraction”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* PP, pp. 1–26. DOI: 10.1109/TASLP.2023.3254166. URL: https://doi.org/10.1109/TASLP.2023.3254166.

Tseng, Hsiao-Ting et al. (2024). “Emotional Reactions in Information Dissemination Through the Lens of SOR Theory”. In: *arXiv*. URL: https://arxiv.org/abs/ND.

Tucker, Joshua et al. (2018). “Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature”. In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3144139. URL: https://doi.org/10.2139/ssrn.3144139.

Vosoughi, Soroush, Deb Roy, and Sinan Aral (2018). “The Spread of True and False News Online”. In: *Science* 359.6380, pp. 1146–1151. DOI: 10.1126/science.aap9559. URL: https://doi.org/10.1126/science.aap9559.

Wang, Fanfan et al. (2024). “SemEval-2024 Task 3: Multimodal Emotion Cause Analysis in Conversations”. In: *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Mexico City, Mexico: Association for Computational Linguistics, pp. 2022–2033. URL: https://aclanthology.org/2024.semeval2024-1.273.

WatchDog.md (2024). *WatchDog.md - Think-Tank and Civil Society Community*. URL: https://watchdog.md/en/.

Xia, Rui and Zixuan Ding (2019). “Emotion-Cause Pair Extraction: A New Task to Emotion Analysis in Texts”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Association for Computational Linguistics. Florence, Italy, pp. 1003–1012. DOI: 10.18653/v1/P19-1096. URL: https://doi.org/10.18653/v1/P19-1096.

Zhan, Hongli et al. (2022). “Why Do You Feel This Way? Summarizing Triggers of Emotions in Social Media Posts”. In: *arXiv*. URL: http://arxiv.org/abs/2210.12531.

Zhang, Xueying et al. (2021). “Mining Dual Emotion for Fake News Detection”. In: *Proceedings of the Web Conference 2021*, pp. 3465–3476. DOI: 10.1145/3442381.3450004. URL: https://doi.org/10.1145/3442381.3450004.

Zheng, Mingqian, Jiaxin Pei, and David Jurgens (2023). “Is ‘A Helpful Assistant’ the Best Role for Large Language Models? A Systematic Evaluation of Social Roles in System Prompts”. In: *arXiv*. URL: https://doi.org/10.48550/ARXIV.2311.10054.