Detecting deepfakes and false ads through analysis of text and social engineering techniques

Alicja Martinek
NASK National Research Institute
ul. Kolska 12, 01-045 Warszawa
AGH University of Kraków
al. Mickiewicza 30, 30-059 Kraków
alicja.martinek@nask.pl

Ewelina Bartuzi-Trokielewicz
NASK National Research Institute
ul. Kolska 12, 01-045 Warszawa
Warsaw University of Technology
Plac Politechniki 1, 00-661 Warszawa
ewelina.bartuzi@nask.pl

Abstract

Existing deepfake detection algorithm frequently fail to successfully identify fabricated materials. These algorithms primarily focus on technical analysis of video and audio, often neglecting the meaning of content itself. In this paper, we introduce a novel approach that emphasizes the analysis of text-based transcripts, particularly those from AI-generated deepfake advertisements, placing the text content at the center of attention. Our method combines linguistic features, evaluation of grammatical mistakes, and the identification of social engineering techniques commonly used in fraudulent content. By examining stylistic inconsistencies and manipulative language patterns, we enhance the accuracy of distinguishing between real and deepfake materials. To ensure interpretability, we employed classical machine learning models, allowing us to provide explainable insights into decision-making processes. Additionally, zero-shot evaluations were conducted using three large language model based solutions to assess their performance in detecting deepfake content. The experimental results show that these factors yield a 90% accuracy in distinguishing between deepfake-based fraudulent advertisements and real ones. This demonstrates the effectiveness of incorporating content-based analysis into deepfake detection, offering a complementary layer to existing audio-visual techniques.

1 Introduction

Disinformation is one of the biggest plagues of 21st century. With the rise of deep learning technologies, false narratives can take more sophisticated forms such as deepfakes and can spread faster, causing more damage. According to the report (Digital, 2024), incidents using deepfake technology have seen a dramatic rise, with a threefold increase in video deepfakes and an eight-fold growth in voice deepfakes between 2022 and 2023. This surge resulted in an estimated 500,000 deepfakes shared on social media platforms in 2023 alone.

Deepfake technology is used in a variety of ways, particularly in financial scams. These schemes often target vulnerable groups, including young internet users and seniors, who may be lured by unrealistically attractive offers. Especially for the latter group, the desire for financial independence from relatives can lead to risky decisions, making them prime targets for scams.

The scripts used in deepfake scams tend to follow a specific and well-structured pattern. These scams typically start with convincing deepfake videos or voice recordings of high-profile celebrities promoting financial opportunities, then tempt victims with promises of quick and easy wealth. This well-planned scenario exploits psychological factors such as fear of missing out (FOMO) or the urgent need to act quickly before the opportunity disappears. Alongside financial frauds, fake advertisements of medical treatments and procedures are prevalent in the deepfake landscape. These ads often promote unproven or unsafe treatment, using fake recommendations from alleged medical professionals or celebrities to manipulate vulnerable individuals into trusting the fraudulent claims.

Disinformation is spreading faster than ever, and language analysis plays a key role in detecting manipulation. One way to strengthen the impact of a message is to pair it with imagery. As a consequence, manipulated audiovisual materials are increasingly used to spread disinformation, in which authorities are presented as sources of false content. As Sander van der Linden points out in his book Fake News (van der Linden, 2024), even when false information is repeatedly corrected, it can still influence the beliefs of the recipient. This trend is due to the fact that disinformation is more likely to perpetuate itself when it fits the existing expectations and vision of the world in the minds of the recipient (Biela, 2016). In detecting deepfakes and

Proceedings of the 31st International Conference on Computational Linguistics, pages 8432–8448
January 19–24, 2025.

fake advertisements, language analysis serves not only as a technical tool but also as a psychological one, aiding in understanding and identifying manipulative audiovisual content by focusing on the text and context of the message.

Fake advertisements often exhibit numerous audiovisual errors like mismatched lip movements and spoken content, blurring around the mouth, visible processing gaps, unnatural shadows around the contours of the face, body inconsistencies, language changes, mismatches between the emotional tone of the voice and the content or gestures, audio cuts, abnormal pauses, crackling noises (so-called digital artifacts), voice tones that differ from the original identity, audible and unusual breathing, accent issues, robotic-sounding voices, a speech tempo that feels like reading, and pitch changes. It is important to note that cybercriminals often use various techniques to mask these distortions, such as adding background music, noise, image blurring, or using texture noise in the video. However, in this article, we focus not on the technical aspects of audiovisual analysis, but rather on sentence structure, grammatical and logical errors, as well as social engineering techniques used by fraudsters, which allow us to identify patterns commonly used in these fraudulent schemes.

While technical text analysis has been widely explored, linguistic analysis of deepfake content remains under-researched. To the best of our knowledge, no significant research has been conducted to differentiate fake advertisements from real ones based on linguistic analysis. While most of papers focus on spam/ham classification or general fake news detection, our research explores:

- nuances of stylometric features and their bound to classes of data,
- categorization and examples of social engineering techniques prevalent in deepfake materials,
- the role of linguistic patterns, such as grammar mistakes, sentiment, and part-of-speech tagging, in distinguishing deepfake advertisements from legitimate content.

In addition to these linguistic-focused approaches, we have incorporated explainable classical machine learning models to ensure interpretability and transparency in decision-making. These models allow for deeper insights into the reasoning behind classification results, which is critical in the context of sensitive applications like deepfake detection. Furthermore, to evaluate the potential of advanced techniques, we conducted zero-shot testing using three solutions based on large language models. This dual approach enables us to leverage the strengths of both traditional explainable methods and deep neural architectures.

2 Related work

Deepfakes rely on artificial intelligence and deep learning algorithms to create highly realistic synthetic audiovisual materials, such as images, videos, and voices, depicting events that never happened. In the context of fake advertisements, deepfakes can be used to manipulate the images of public figures or celebrities who unknowingly promote products or services they have no connection with. An example could be fake videos featuring well-known individuals encouraging viewers to invest in fraudulent financial schemes, buy unverified products, or support controversial campaigns. This can mislead audiences and damage trust in the associated brand (Di Domenico and Ding, 2023). Although fake news lack the advanced multimedia technologies used in deepfakes, its influence on consumer trust can be equally damaging (Botha and Pieterse, 2020).

Social engineering techniques are critical to the success of both deepfake and fake advertisements, exploiting human psychology to manipulate individuals into specific actions. Common tactics include urgency (the need to act quickly), scarcity (limited-time offers), and appeals to authority (celebrity endorsements), all of which are amplified by the realism of deepfake content. According to Cialdini's principles of persuasion (Cialdini, 2007), methods such as reciprocity, scarcity, authority, consistency, liking, and consensus play a central role in influencing consumer behavior, with deepfakes adding a layer of realism that enhances the psychological impact.

Research highlights the frequent use of linguistic manipulation in scams, fake news, and fraudulent advertisements, employing exaggerated promises, emotional appeals, and fabricated testimonials to deceive consumers (Biela, 2016; Ferreira et al., 2015). Studies show that fake ads often feature linguistic markers like incorrect grammar, illogical sentence structures, and inconsistent narratives, which can be detected through automated analysis (Prelipceanu, 2013). Scam communications, including deepfake-generated ads, use persuasive language, urgency, and social proof to tempt victims with promises of financial rewards, making

linguistic features key indicators of fraudulent and disinformative content (Anafo and Ngula, 2020; Modzelewski et al., 2024).

To the best of our knowledge, no significant research has been conducted specifically on detecting fake advertisements through text analysis. While a substantial amount of work has been done in the field of fake news detection, existing studies generally focus on the analysis of articles, headlines, or social media posts, rather than commercial or advertising content. Fake news detection methods generally utilize a combination of linguistic, statistical, and semantic features to distinguish between real and fake content.

Linguistic analysis plays a crucial role in detecting fraudulent content. For example part-of-speech tagging (POS) (Hassan et al., 2015), Named Entity Recognition (NER) (Boididou et al., 2018), and stylometric features (Okulska et al., 2023; Wood, 2024) have been widely used to identify key grammatical patterns and sentence schemes that differentiate fake from real text. These features allow system to capture syntactic and stylistic anomalies, which are often presented in AI-generated content.

Sentiment analysis has been another valuable tool in fake news detection. Fake news often contain provocative or emotionally charged language designed to manipulate readers. Studies such as those by Sun (Sun et al., 2013) and Wu (Wu et al., 2015) have used tools like Linguistic Inquiry and Word Count (LIWC) (Kwon et al., 2013; Pérez-Rosas et al., 2018) and sentiment lexicons (Ma et al., 2015) to extract features related to the emotional tone of the text. These approaches detect sentiment-based cues, such as the frequency of negative or extreme adjectives, which are typically used in misleading advertisements.

The use of language models to calculate perplexity has also shown a promise in identifying whether a text follows a predictable or natural linguistic structure. Fake news, particularly those generated using AI, often exhibit higher perplexity due to their artificial construction. These differences in predictability can be exploited to flag potentially fake content (Lee et al., 2020).

Text vectorization techniques, such as Term Frequency-Inverse Document Frequency (TF-IDF), have been widely used in binary classification tasks like spam-ham detection (Shahariar et al., 2019; Gilda, 2017; Ahmed et al., 2017). By quantifying the relevance of worlds in a document compared to the entire corpus, TF-IDF helps identify words or phrases that are common in fake news. Such techniques are effective when combined with machine learning classifiers to automatically differentiate between fake news and real ones.

In the context of binary classification, previous research has shown the effectiveness of applying supervised machine learning models to distinguish between real and fake content. Techniques such as logistic regression (Patel and Meehan, 2021; Lai et al., 2022), support vector machines (SVM) (Hussain et al., 2020; Baarir and Djeffal, 2021), and more recently deep learning models like LSTMs and RNNs, have been applied to classify fake news based on text features (Chen et al., 2018; Rashkin et al., 2017; Wang et al., 2018; Volkova et al., 2017; Popat et al., 2018; Kumari et al., 2021). However, their application domain of false advertisement remains under-explored.

3 Experimental dataset

Dataset for this study comes from two main sources. The first is the UKNF (Polish Financial Supervision Authority), which monitors the Internet for fraudulent financial activity, providing a rich source of data for identifying deepfake scams. The second source involves internal searches for deepfake materials across the Internet.

Internal searches targeted controversial topics such as alternative medicine and government subsidies, as well as content from specific influencers. Legitimate advertisements, primarily from YouTube and centered around the financial sector, were gathered and filtered to include only those with spoken language. Over 70 keywords commonly associated with deceptive ads (e.g. “guaranteed income”, “quick profit”) were used to identify deepfake scams that often make false promises. Additionally, major social media platforms like Facebook, Instagram, X, YouTube, and TikTok were monitored to collect deepfake videos flagged as fraudulent or disinformative.

The collected video data underwent audio processing, with transcripts generated using Whisper by OpenAI (OpenAI, 2022; Radford et al., 2022). Despite transcription errors, especially with cloned voices, the data was manually corrected. In total, the dataset contains 5966 fake and 858 real sentences across 252 and 114 transcripts, respectively. Extensive metadata labeling helped categorize the materials based on topics such as financial instruments, medicine, disinformation, and fraud, provid-

topic	count	fakes count	real count
investing	152	135	17
medicine	92	72	20
other	36	7	29
help	30	15	15
gambling	10	10	0

Table 1: Five most popular topics of ads in the dataset.

ing insights into which areas are most vulnerable to manipulation, as shown in Table **??**. Before annotation, a detailed manual review of all collected data was conducted to ensure that only relevant materials were included in the dataset. This step reflects a data-centric approach, highlighting the importance of thorough analysis and understanding of the dataset. Clear and detailed instructions were provided to annotators, outlining how to address transcription errors, label metadata, and categorize topics. A detailed set of instructions was prepared for annotators, specifying how to handle transcription errors, label metadata, and categorize topics. For instance, topics were assigned based on clear criteria, such as explicit mentions of financial schemes or references to medical treatments.

A subset of the dataset (35%) was annotated independently by multiple annotators to measure inter-annotator agreement. The resulting Cohen's kappa score of 0.88 indicates a high level of consistency in the annotations. After automated transcription, all transcripts were manually reviewed to correct errors and verify alignment with the audio content. Particular attention was paid to cloned voices, where errors were more prevalent, and sentences containing grammatical inconsistencies or logical errors were carefully documented.

To our best knowledge, there are no open datasets containing transcripts of deepfake materials, not to mention deepfake fraudulent advertisements. We believe that our data set, even tough in Polish, is still valuable and unique.

3.1 Most common mistakes

In our study, we identified several common errors in fake ads. Each of these errors reveals shortcomings in the methods used by cybercriminals and can help detect fake content. These errors fall into four main categories: logical inconsistencies, gender misalignment, lack of declension, mishaps with numbers and dates. Examples of these mistakes can be found in Appendix B.

Analysis of these common mistakes allows to draw conclusion that cyberattackers distributing these fake ads are likely operating from multiple countries and often lack the necessary cultural and linguistic background of the regions they are attacking. This is a common problem, with criminals from different parts of the world distributing fake content in multiple countries, often without a deep understanding of the local context. As a result, their content may seem effective on a broad scale, but it often contains errors that reveal their ignorance of the specific cultural and linguistic environment of the target area.

3.2 Manipulation and social engineering

Deepfakes appearing in social media, especially as advertisements, follow all guidelines for successful marketing manipulation. In our study, based on collected dataset, we categorized the manipulation techniques in four general groups: tension-building, language-style layer, motivational and sociological. These include: **time pressure** - such technique requires person to act fast on the advertisement while reducing the time to stop and deeply think about undertaken actions; **evoking certain emotions** - action designed to purposely force people into strong emotions, often extreme like fright, shock or happiness or inducing them to take specific actions; **revealing conspiracy** - exposing fraud, plotting, disclosing secrets, hiding the truth, concealing facts and statistics; **stunts** - emotionally charged messages that are exaggerated, shocking, or dramatic, intended to surprise or confuse the recipient; **repetitions** - multiple mentions of the same statements in order to amplify the message; **superlatives** - presenting something in a good light, praising, inflating its advantages, giving positive statements without a proof; **discrediting others** - negatively presenting the competition, creating an exaggerated false picture; **competitiveness** - referring to the advances in technology, automation and methodology of advertised product, presenting own solution as superior to others; **transparency** - describing self as trustworthy, reliable and giving statements of positive intentions - such behavior is aimed to reduce alertness; **ease of reaching the goal** - in most cases ease of earning money, presenting own product as user friendly and the process of getting desired output as an instant one; **offer's luxury and limitation** - giving the illusion of restricted availability makes people want something more; **risk minimization** - ensuring about safety, success rate and presenting success stories of other people; **giving a promise** - providing assurance

about the effectiveness and validity of the actions taken, which end with certain reward; **call to action** - persuading into taking certain steps, gives very simple instructions which are easy to follow, hence the chance of completing them increases; **altruism** - assurance of selfless and good intentions, early and extensive testing of solutions, combating global social problems; **targeting normal people's needs** - creating the illusion of idyllic life in which everyone can fulfil their dreams without work as well as addressing and presenting solutions to everyday life problems and worries; **direct talk** - straightforward questions, shortening the interpersonal distance, identifying with the recipient, pretending to be equals; **social endorsement** - confirmation by other people about the effectiveness of the methods, reviews and recommendations, often false testimonies and personal stories, reference to other users or famous people; **addressing possible concerns** - challenging stereotypes, anticipating the recipient's possible fears and doubts and addressing them before they become serious. Figure 1 shows the frequency of techniques in the dataset, expressed as the count of a given technique divided by number of sentences in a transcript. This provides average density of social engineering among transcripts for both classes, real and deepfake ads.

4 Methodology

4.1 Explainable approaches

The primary focus of our data processing was on linguistic and stylistic analysis, which required clean and structured text data. After obtaining the initial transcripts, further processing steps were applied to prepare the data for analysis. We employed the SpaCy library, which provides open-source tools for text processing across multiple languages, including Polish. The `pl_core_news_lg` model was used for lemmatization and **Part-Of-Speech** (POS) tagging to standardize word forms and identify syntactic roles within sentences.

Furthermore, the analysis involved measuring the **sentiment** associated with each transcript. This calculation was conducted using the `eevvgg/bert-polish-sentiment-politics` model, available on Hugging Face (Gajewska and Konat, 2023). This model is trained specifically to predict sentiment in Polish, with special attention drawn to political topics. Such model makes a perfect fit for data in given study as the substantial share of deepfakes is related to politics. There were cases where length of transcript exceeded computational possibilities of models. In order to mitigate this issue, transcripts were chunked into bits that included around 300 tokens (with respect to the sentence level, hence it was not always exactly 300). Results describing whole transcript were derived as an average result of all chunks.

Perplexity of text is another metric that can be used to describe characteristics of a transcript. LM-PPL library (Asahi417 GitHub user, 2023) implements such calculations with variety of models. Among others, recurrent language model `facebook/opt-30b` was selected for computations. Perplexity can be seen as the inverse of the geometric mean of the probability of each word $P(W)$.

$$Perplexity(W) = P(W)^{-\frac{1}{N}} \quad (1)$$

Analysis of **style and unique features** associated with it can be a source of meaningful features. Stylometrix library (Okulska et al., 2023) builds text embedding reflecting the author's style in Polish, calculating over 170 features. Principal Component Analysis helps visualize the separability of real and deepfake data based on style.

The next step included calculating **TF-IDF**, a text vectorization method that assigns values to words (terms) in a document based on the formula:

$$TF\text{-}IDF(t, d, D) = TF(t, d) * IDF(t, D), \quad (2)$$

where t denotes term, d a document and D means collection of documents (corpora). TF measures term frequency in a document, while IDF penalizes terms common across all documents.

TF-IDF results, from the lemmatized corpus were used for classification. **Logistic Regression** (LR) was used in order to decide if given document (transcript) represents a real advertisement or a fake one. Classification was run on three different sets of features, each time adding new information. The simplest classifier was built only with average sentence density values for social engineering techniques (SE). More advanced data included TF-IDF vectors describing each transcript. Last set of features - *linguistic* was further extended to also cover perplexity, sentiment and percentages of POS tags.

Keeping in mind narrow applicability and generalization of such analysis, it was decided to run same set of calculations on translated transcripts. Translation was completed with the use of `gpt-4o-mini`. We are aware that translating polish-specific transcripts may lose some nuances and details of unique style. However, we

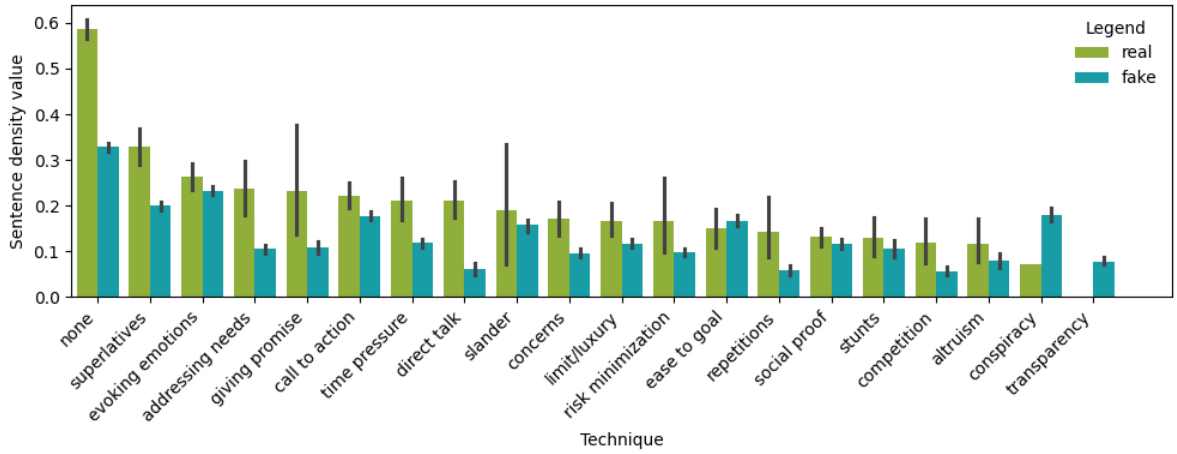

Figure 1: Social engineering techniques - average occurrences per sentence in real and deepfake materials.

made efforts to ensure that the translations preserved the original context and style, allowing for a direct comparison of the results in both Polish and English. Models used for consecutive processing steps were replaced with those dedicated to English. As a result `en_core_news_lg` and `distilbert-base-uncased-finetuned-sst-2-english` were used for POS tagging, lemmatization and for sentiment analysis correspondingly.

Last step of processing pipeline, which was unique only to Polish transcripts, utilized large language model to perform analysis of social engineering techniques. OpenAI's `gpt-4o-mini` was provided with a single transcript at a time and prompt describing the context in details. Further information about the prompting phase can be found in Section A within Supplementary Materials.

4.2 LLMs-based experiments

In this part of the study, we explored two approaches leveraging LLMs to classify advertisements as real or fake based on their content: a zero-shot classification using LLMs and a fine-tuned BERT-based classifier.

Zero-shot classification using LLMs - the first approach used 2 advanced LLMs, OpenAI's `GPT-4o-mini` and Mistral's `Ministral-8B-Instruct`, to classify advertisement transcripts as real or fake without requiring any task-specific training.

The prompt used for interaction with both models was as follows: `You are analyzing transcripts from ads found on the Internet. Answer with one word: 'REAL' if you think that the transcript comes from a real ad or with word 'FAKE' if the transcript comes from a scam.`

Each transcript was processed individually using GPT-4o-mini and Mistral models. Both models returned classifications that were compared with ground truth labels to measure accuracy.

BERT-based feature extraction with Logistic Regression - the second approach combined a pretrained BERT model for feature extraction with a Logistic Regression classifier for final classification. This hybrid method leveraged BERT's ability to capture contextual embeddings, the simplicity and interpretability of logistic regression.

5 Experimental results

The experiments conducted in this research focused on identifying characteristic data features that can differentiate between transcripts of two classes: real advertisements and fake advertisements generated using deepfake techniques.

In the first step of analysis, perplexity scores were calculated for both classes. Real advertisements had a mean perplexity of 57.18 $\pm$ 73.72, while for fake advertisements was around 20.84 $\pm$ 13.98. Perplexity measures text predictablility, with lower scores indicating more structured patterns. Real ads show a wider range of predictability, as evidenced by the broader distribution of perplexity. In contrast, fake transcripts exhibit a narrower distribution, with a clear peak, suggesting that deepfake scripts tend to be more predictable and repetitive. This distinction highlights the repetitive nature of deepfake scripts and suggests that perplexity, combined with other linguistic features, can effectively aid in early detection of deepfake content.

The next set of features refers to the grammatical structure of sentences. Figure 2 shows average dis-

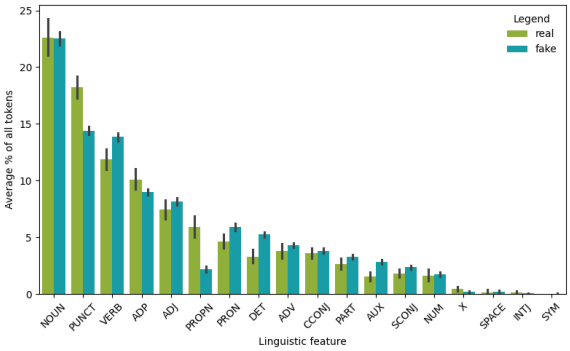

Figure 2: Part-of-Speech tagging in Polish, showing the average share of each token type per advertisement. Data is sorted by values of the *real* data class.

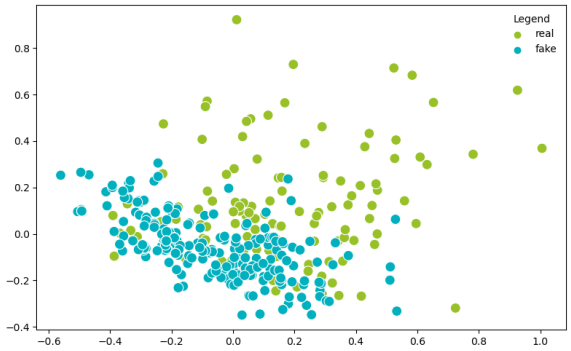

Figure 3: Visualization of Principal Component Analysis based on the StyloMetrix output for Polish corpora.

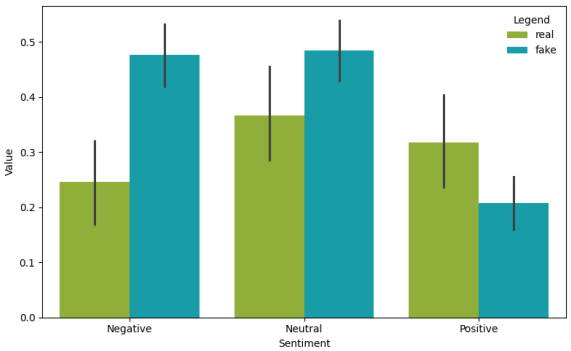

Figure 4: Average sentiment analysis for transcripts in the Polish language.

tributions of POS tags across tokens within a transcript, with regards to the dataset labels. The most noticeable disproportion between classes is observable in case of PROPN tag (*proper noun*), where in real advertisements its share is over 6%, whereas for deepfakes its below 2.5%. This discrepancy likely stems from the fact that real advertisements often reference specific products, brands, or people, necessitating a higher frequency of proper nouns. In contrast, deepfake ads may avoid such specificity to maintain a broader, more generalized appeal.There are also significant difference in the use of PUNCT (*punctuation*) and VERB (*verb*) tags. The higher share of verbs in false ads might be attributed to the frequent use of *call to action* technique, where the audience is encouraged to perform specific actions, such as clicking a link, making a purchase, or signing up for a service. This tactic typically requires a more action-oriented language, which explains the elevated presence of verbs.

A more advanced analysis of the linguistic and stylometric characteristic of the text was performed with use of StyloMetrix library. The resulting feature vectors were subsequently processed through the Principal Component Analysis algorithm. The visualizations of two most significant components is present in Figure 3. The arrangement of the green dots resembling real ads, suggests that these texts are more diverse in nature, whereas blue dots (representing fake ads) are creating a cluster-like structure on the plot. This indicates that fake advertisements share more similar linguistic and stylometric traits, potentially due to their repetitive or templated nature. However, it is important to note that, in this setting, the two classes are not linearly separable. This suggests that while there are observable differences in the stylometric properties of real and fake ads, these distinctions may require more sophisticated or nonlinear methods to achieve a clear separation between the two classes.

The sentiment of transcriptions was another key measure under evaluation in this research. The sentiment analysis model built for the Polish language categorized into three groups: negative, neutral and positive. Figure 4 presents some intuitive observations - false advertisements have much more negative tone due to their nature. Fake ads often employ slander and divisive language to polarize the audience. On the other hand, legitimate advertisements exhibit a higher positive sentiment, as they typically avoid negative marketing strategies and focus on promoting products or services in a more constructive manner.

During the exploration stage of with LLM, ChatGPT was tasked with generating two specific statistics at the transcript level. This decision stemmed from key observations made during in-depth review of deepfake transcripts. It was noticed that, along with numerous grammatical errors, there were frequent references to nationality, specifically related to Poland in our case, throughout the texts. "Poland," "Poles," and "citizens of Poland" were frequently used throughout the fake advertisements. This pattern suggests an intentional targeting of the national audience, leveraging national identity in an effort to manipulate and appeal to Polish viewers.

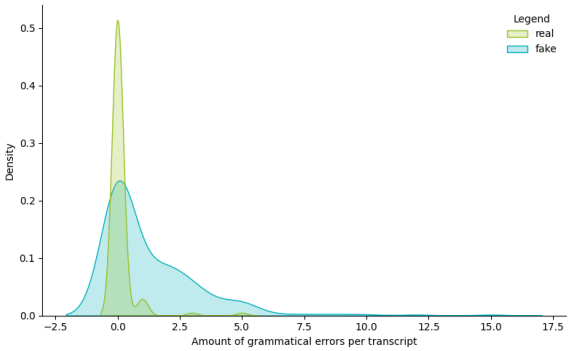

Figure 5: Distribution of grammatical errors in Polish transcripts, including issues with verb conjugations, declensions, and inflections.

Grammatical errors often involved incorrect verb conjugations, declensions, and misused inflections, which are particularly challenging in Polish due to its complex grammar. Results of this countings are depicted in Figure 5.

Having extracted various features and embeddings that describe the dataset, a simple classifier was built to explore their potential in distinguishing between transcripts of false and real advertisements. Logistic Regression was chosen for this task due to its simplicity, interpretability, and effectiveness as a baseline model in binary classification problems.

Since the problem at hand is binary (false vs. real), Logistic Regression was applied with three different sets of features: **SE embeddings**, **SE embeddings combined with TF-IDF features**, and **SE embeddings combined with both TF-IDF and linguistic features** including POS tag counts, average perplexity and sentiment. These sets were selected to explore whether traditional linguistic indicators or more advanced embeddings perform better in identifying deepfake content. Table 2 summarizes achieved performance and displays confusion matrices for each classifier.

Table 2 presents the accuracy and standard deviation results for various methods tested on both Polish and English datasets, comparing classical feature-based classifiers with zero-shot LLM-based approaches in detecting deepfake advertisements.

The classical methods utilized Logistic Regression as the final classifier, employing three different sets of features: SE embeddings, SE + TF-IDF embeddings, and SE + TF-IDF + linguistic features.

Social Engineering (SE) features alone served as a baseline, achieving 80.91% accuracy for Polish and 79.36% for English. While this approach demonstrated solid performance, its limited scope highlighted the need for additional feature sets

	Polish		English	
Model	Acc	Std	Acc	Std
SE	80.91	5.54	79.36	2.76
SE + TF-IDF	89.91	3.21	89.36	2.76
SE + TF-IDF + Linguistic	**90.55**	2.88	**90.82**	1.93
GPT-4o-mini	**94.86**	6.89	**95.40**	5.68
Ministral-8B-Instruct	80.54	9.57	84.59	19.78
BERT + Logistic Regression	80.11	7.50	87.08	4.80

Table 2: Accuracy and standard deviations for LLM-based and feature-based methods on Polish and English datasets, averaged over 10-fold cross-validation.

to improve classification accuracy. Adding TF-IDF vectors to SE embeddings significantly enhanced performance. The accuracy increased to 89.91% for Polish and 89.36% for English, indicating that combining SE features with TF-IDF provides deeper insight into the structure and content of the transcripts. Further enrichment with linguistic features, yielded the best results among feature-based approaches. The accuracy reached 90.55% for Polish and 90.82% for English, with reduced variability (standard deviations of 2.88 and 1.93, respectively). These results underscore the value of combining multiple feature sets to capture the nuanced patterns of deepfake content while maintaining interpretability.

GPT-4o-mini, a zero-shot LLM, achieved the highest accuracy on both Polish (94.40%) and English (94.67%) transcripts. Its low standard deviation values (4.00 for Polish and 4.59 for English) indicate consistent and robust performance across different data folds. These results demonstrate the model's strong generalization capability, even without task-specific fine-tuning, making it highly effective at detecting nuanced manipulations and stylistic inconsistencies in the advertisements. However, despite its performance, the model operates as a "black box," offering no explainability for its decision-making process. This lack of interpretability poses challenges for understanding why certain advertisements are classified as fake or real, which could be problematic for high-stakes applications or when transparency is required.

Ministral-8B-Instruct, another LLM tested in a zero-shot setting, showed moderate accuracy, with 80.54% for Polish and 84.59% for English. However, its relatively high standard deviations (9.57 and 19.78, respectively) suggest variability in its ability to generalize across different data folds.

BERT with Logistic Regression demonstrated strong performance, particularly on the English

dataset, achieving an accuracy of 87.08%. On the Polish dataset, the model achieved 80.11% accuracy. Unlike LLMs, this approach offers greater interpretability by leveraging task-specific features. This transparency allows for better understanding and debugging of the model's classifications, making it a preferred option in contexts requiring explainable decisions.

5.1 Contrasting results after translation

To further evaluate the robustness of the model and the features used, an additional experiment was conducted by translating the original Polish transcripts into English while maintaining the context and style of the statements. The goal was to determine whether the results obtained from the Polish dataset would hold after translation, given that certain linguistic nuances and cultural references might not translate seamlessly between languages.

The results showed minimal variation in accuracy, with the classifier performing similarly across all feature sets in both languages. The SE + TF-IDF feature set, accuracy for the Polish data was 89.91%, while for the English-translated dataset it reached 89.36%. This indicates that the distinguishing characteristics of false and real advertisements are language independent to a large extent. The similarities between the Polish and English results suggest that the models and features developed in this study can be applied across different languages, enhancing their generalizability and potential for broader use in detecting fake ads globally.

6 Conclusions

This study presents a novel approach to detecting deepfake advertisements by combining linguistic, social engineering, and stylometric features. The results demonstrate the effectiveness of these features in distinguishing real from false ads in both in the original Polish and in English-translated dataset. Social engineering features alone provide a strong baseline, with 80% accuracy. Addition of TF-IDF vectors significantly improves the model's performance, raising the accuracy to 89.91%. Further enrichment with linguistic features, including POS tags, perplexity, and sentiment analysis, boosts accuracy to 90.55%, underscoring the importance of combining multiple feature sets to capture the complexity of deepfake content and improve detection.

It is important to highlight that a comprehensive approach to deepfake detection is essential. While technical analysis of audiovisual materials is critical for spotting manipulations in deepfake videos and audio, textual analysis of transcripts offers a crucial additional layer of verification. By focusing on language, sentence structure, and textual inconsistencies, this study demonstrates how text-based methods can enhance the overall deepfake detection process, uncovering common mistakes, linguistic anomalies, and manipulative tactics that might not be immediately apparent through technical analysis alone.

The results highlight the potential of LLMs for detecting deepfake advertisements, with models like GPT-4o-mini achieving the highest accuracy across both Polish and English datasets, exceeding 94% in a zero-shot setting. However, their lack of explainability limits their practical application in scenarios demanding transparency. In contrast, feature-based approaches like the combination of SE features, TF-IDF vectors, and linguistic attributes, provide a more interpretable solution. Achieves high accuracy (over 90.5%), this method offers valuable insights into decision-making processes, making it a robust alternative for use cases where explainability and are crucial.

Importantly, this study highlights the cross-linguistic effectiveness of the proposed approaches. The results from English-translated transcripts showed minimal variation in performance compared to the original Polish dataset, indicating that many of the characteristics that distinguish fake from real advertisements are language-independent. This suggests that the proposed methodology developed can be successfully extended to other languages, broadening its applicability and making it a valuable tool in global efforts to detect and mitigate the spread of deepfake ads. Common mistakes in deepfake ads, such as grammatical inconsistencies, gender mismatches, and logical errors, further demonstrates that cybercriminals often operate without a thorough understanding of the linguistic and cultural nuances of their target regions. These mistakes can serve as critical indicators in the detection of fake content, providing another layer of defense against the spread of disinformation.

In conclusion, the results of this study emphasize the importance of a multifaceted approach to deepfake detection. Beyond technical analysis of the manipulated audiovisual content, the investigation of textual features such as grammar, style, and linguistic manipulations plays a crucial role in enhancing the content verification process.

7 Limitations

The biggest limitation of this study is the language barrier, as the dataset is primarily in Polish. However, translating the sentences into English does not alter the core results or findings of the analysis, as the manipulation techniques and structures remain consistent across languages.

Another limitation is the dataset size. While we have collected a significant amount of both fake and real sentences, the dataset is still limited in comparison to larger, global studies. This constraint may affect the generalizability of the findings, especially across different cultural contexts or languages beyond Polish. Further research with larger and more diverse datasets would be necessary to validate the conclusions on a broader scale.

Acknowledgments

This work was partially supported by the program "Excellence initiative - research university" for the AGH University of Kraków and by Grant for Statutory Activity from the Faculty of Physics and Applied Computer Science of the AGH University of Kraków.

This work was also partially supported by the program "Science4Business – Nauka dla Biznesu" under the task no. 1, "Inkubator Rozwoju," as part of the initiative to foster collaboration between science and industry. Additional support was provided through the "Excellence Initiative – Research University" program.

References

Hany Ahmed, Issa Traore, and Sherif Saad. 2017. Detection of online fake news using n-gram analysis and machine learning techniques. In Issa Traore, Isaac Woungang, and Ahmed Awad, editors, *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, volume 10618 of *Lecture Notes in Computer Science*, pages 127–138. Springer, Cham.

Comfort Anafo and Richmond S Ngula. 2020. On the grammar of scam: transitivity, manipulation and deception in scam emails. *Word*, 66(1):16–39.

Asahi Asahi417 GitHub user. 2023. Language Model Perplexity (LM-PPL). https://github.com/asahi417/lmppl.

Nihel Fatima Baarir and Abdelhamid Djeffal. 2021. Fake news detection using machine learning. In *2020 2nd International workshop on human-centric smart environments for health and well-being (IHSH)*, pages 125–130. IEEE.

Bogdan Biela. 2016. Typy i mechanizmy manipulacji w mediach. *Studia Pastoralne*, 12:286–311.

Chrysoula Boididou, Stuart E. Middleton, Zhiwei Jin, et al. 2018. Verifying information with multimedia content on twitter. *Multimedia Tools and Applications*, 77(12):15545–15571.

Johnny Botha and Heloise Pieterse. 2020. Fake news and deepfakes: A dangerous threat for 21st century information security.

Tianyu Chen, Xue Li, Hongzhi Yin, et al. 2018. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 40–52. Springer.

Robert B Cialdini. 2007. *Influence: The psychology of persuasion*, volume 55. Collins New York.

Giandomenico Di Domenico and Yu Ding. 2023. Between brand attacks and broader narratives: How direct and indirect misinformation erode consumer trust. *Current Opinion in Psychology*, 54:101716.

Redline Digital. 2024. Fake news statistics facts. https://redline.digital/fake-news-statistics/.

Ana Ferreira, Lynne Coventry, and Gabriele Lenzini. 2015. Principles of persuasion in social engineering and their use in phishing. In *Human Aspects of Information Security, Privacy, and Trust: Third International Conference, HAS 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015. Proceedings 3*, pages 36–47. Springer.

Ewelina Gajewska and Barbara Konat. 2023. Parestw: Bert for sentiment detection in polish language. https://huggingface.co/eevvgg/PaReS-sentimenTw-political-PL.

S. Gilda. 2017. Notice of violation of ieee publication principles: Evaluating machine learning algorithms for fake news detection. In *2017 IEEE 15th Student Conference on Research and Development (SCOReD)*, pages 110–115. IEEE.

Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, pages 1835–1838. ACM.

Md Gulzar Hussain, Md Rashidul Hasan, Mahmuda Rahman, Joy Protim, and Sakib Al Hasan. 2020. Detection of bangla fake news using mnb and svm classifier. In *2020 International Conference on Computing, Electronics & Communications Engineering (iCCECE)*, pages 81–85. IEEE.

S. Kumari, H. K. Reddy, C. S. Kulkarni, et al. 2021. Debunking health fake news with domain specific pre-trained model. *Global Trans. Proc.*, 2(2):267–272.

Sejeong Kwon, Meeyoung Cha, Kyomin Jung, et al. 2013. Prominent features of rumor propagation in online social media. In *2013 IEEE 13th International Conference on Data Mining (ICDM)*, pages 1103–1108. IEEE.

Chun-Ming Lai, Mei-Hua Chen, Endah Kristiani, Vinod Kumar Verma, and Chao-Tung Yang. 2022. Fake news classification based on content level features. *Applied Sciences*, 12(3):1116.

Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 2020. Misinformation has high perplexity. *Preprint*, arXiv:2006.04666.

Jun Ma, Wei Gao, Zhenzhen Wei, et al. 2015. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1751–1754. ACM.

Arkadiusz Modzelewski, Giovanni Da San Martino, Pavel Savov, Magdalena Wilczyńska, and Adam Wierzbicki. 2024. Mipd: Exploring manipulation and intention in a novel corpus of polish disinformation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19769–19785.

Inez Okulska, Daria Stetsenko, Anna Kołos, Agnieszka Karlińska, Kinga Głąbińska, and Adam Nowakowski. 2023. Stylometrix: An open-source multilingual tool for representing stylometric vectors. *Preprint*, arXiv:2309.12810.

OpenAI. 2022. Whisper: Automatic speech recognition. `https://github.com/openai/whisper`.

Ankitkumar Patel and Kevin Meehan. 2021. Fake news detection on reddit utilising countvectorizer and term frequency-inverse document frequency with logistic regression, multinominalnb and support vector machine. In *2021 32nd Irish signals and systems conference (ISSC)*, pages 1–6. IEEE.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, et al. 2018. Declare: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 22–32. Association for Computational Linguistics.

Cristina-Maria Prelipceanu. 2013. Advertising and language manipulation. *Diversité et Identité Culturelle en Europe, X (2)*, pages 247–254.

Veronica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, et al. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 3391–3401. ACL.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, et al. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2931–2937. Association for Computational Linguistics.

G. M. Shahariar, Swapnil Biswas, Faiza Omar, Faisal Muhammad Shah, and Samiha Binte Hassan. 2019. Spam review detection using deep learning. In *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE.

Siqi Sun, Hongchao Liu, Jing He, et al. 2013. Detecting event rumors on sina weibo automatically. In *Proceedings of the Asia-Pacific Web Conference*, pages 120–131. Springer.

Sander van der Linden. 2024. *Fake News: Understanding its Impact and How to Combat it*. Dom Wydawniczy Rebis.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, et al. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 647–653. Association for Computational Linguistics.

Yaqing Wang, Fanhai Ma, Zhiwei Jin, et al. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 849–857. ACM.

T.A. Wood. 2024. Fast stylometry (computer software), version 1.0.4.

Kai Wu, Song Yang, and K. Q. Zhu. 2015. False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st International Conference on Data Engineering (ICDE)*, pages 651–662. IEEE.

A Prompt

The prompt requested the model to display results in a table format, where columns represented the filename, sentence, identified technique, and a justification for why that technique was recognized. In addition, the prompt instructed the model to provide a second table summarizing the overall transcript, including the topic of the advertisement, total grammar mistakes, and any mentions of Polish nationality or culture. This textual cue was adjusted accordingly to the class of analyzed transcript. `(...)` contained the detailed listing of social engineering techniques along with examples

and definitions. The full prompt is structured as follows:

```
You are analyzing the transcript of real/fake advertisement found on the Internet. Analyze it sentence by sentence and display results of your analysis in a table. It should have following columns: filename, sentence, technique, justification. The possible manipulation techniques are listed below. In brackets you will see examples and definitions of these techniques. Do not come up with new techniques. In case when no technique was used write 'none'. (...) Now, analyze the whole transcript as one entity. In second table that has columns: filename, topic, grammar mistakes, polish - present aggregated data for the whole transcript. Possible values for column topic are: gambling, cryptocurrencies, currency exchange, investing, medicine, help, institutions, subsidies. In case when no category matches the topic, assign a new label. In column 'grammar mistakes' count each occurrence of an error. Column 'polish' should include number of words that stem from word Poland/Polish that were apparent in the text.
```

B Examples of common mistakes

In Appendix B, we present examples of common mistakes found in deepfake-generated content. The table below provides specific examples of such errors, grouped by type, with accompanying explanations. The errors include logical inconsistencies, gender misalignment, grammatical issues like lack of declension, and mishaps involving numbers and dates.

C Examples of techniques

Appendix C presents examples of sentences that illustrate various manipulation techniques found in deepfake content. These examples are categorized into four distinct groups based on the nature of the manipulation. The sentences serve as clear examples of how different linguistic and psychological strategies are employed to influence or deceive the audience. The tables below provide representative sentences for each group of manipulation techniques.

type of error	examples	explanation
logical inconsistencies	I couldn't walk with my grandchildren*, go shopping, or climb stairs properly.	* It is spoken by a well-known Polish clergyman, who according to the law, cannot have children, let alone grandchildren.
gender misalignment	My name is Kazimierz Nycz* and I have had knee problems for twenty years. I have seen** many massage therapists and chiropractors, but none of them were able to cure my aching joints.	* Deepfake features celebrity (woman, around 40 years old) but the text refers to the identity of a well-known Polish priest. ** Usage of the masculine form of the verb, while the voice belongs to a female.
lack of declension	The project is intended for only one first* hundred thousand participants.	* The numeral one hundred should be changed in order agree with the rest of the sentence, which is a common rule in Polish grammar.
mishaps with numbers, dates	Ninety-nineties* percent of people lose money during a crisis.	* Phrase contains a pronunciation error, which cannot be transcribed correctly.

Table B.1: Examples of common mistakes in deepfakes.

technique	examples
time pressure	If you are watching this video now, we still have a few spots available.
	Hurry up with submitting your application online. The number of available spots is limited. Change your life today.
	But there's a catch. The project is designed for only one hundred thousand participants.
	If you hesitate, you'll likely end up like ninety-nine percent of other people.
	Read it before it gets removed.
evoking certain emotions	I didn't know what home was. I lived in an orphanage until I was sixteen, then in barracks or with friends, and now I finally bought my own apartment. It was my dream.
	Save yourself and your loved ones before it's too late.
	Doctors kill thousands of men every day.
	My life changed radically after getting to know the program from Whatsapp.
	The most famous Polish goalkeeper has serious health problems, and continuing his football career threatens his life.
revealing conspiracy	The government prematurely terminated the contract for Russian gas supplies from Russia.
	I'm not sure how long this film will be available, as pressure from the pharmaceutical mafia is increasing and it will likely be removed soon.
	A drug that destroys diabetes at the cellular level within a few weeks has been concealed by deceitful pharmaceutical companies.
	I hope this film will open everyone's eyes to the truth and show that we've all been deceived for years.
	Don't let the government rob you and your children.
sensations	Today, before your eyes, I'm starting a financial revolution.
	You'll be shocked; it's possible now.
	But be careful, many people start crying tears of joy when they hear this.
	The absurdity. The absurdity of this situation.
	Scandal of the month. What's happening in our country?

Table C.1: Example sentences for manipulation techniques from tension-building group.

technique	examples
repetitions	Just think. Thousands today, a thousand tomorrow, and a thousand every day.
	It can help people. Actually help people.
	Do we need it? Yes. Is it a worthwhile investment? Yes. Is it worth it? Yes. Is it worth spending a lot of money on it? Yes.
	Stop suffering. Stop suffering and waiting for complications.
	Again, I win money every day.
superlatives	A brilliant app on my phone automatically resells stocks on exchanges, and I receive all the profits.
	The Tesla X is now available in Poland and can provide every Polish citizen with a passive income of up to three hundred euros per day.
	This software trades stocks and does it so well that it wins trades with a ninety-eight percent success rate.
	Just one cup a day is enough to boost your metabolism, lose weight five times faster, and visibly transform yourself.
discrediting others	It has nothing to do with the other nonsense you see everywhere.
	Most doctors don't care about it.
	To them, you're just a person who brings them money.
	It would be a lie to promise millions without any work, as others do.
	I'm utterly exhausted from watching ordinary people struggle, save, and restrict themselves while officials stuff their pockets with our money, indulge in expensive treats at our expense, and take their fat behinds to the best resorts, while we hunch over under the weight of taxes.
competitiveness	It's something completely different from Forex.
	What is the difference between these scammers and me?
	It's different from all other games, it won't force you to watch an ad.
	We have developed a powerful software. This is the world's first software.
	This powerful computing device is nothing like a regular home computer.
transparency	I respect you, my reputation is impeccable, and I want to earn your trust.
	This isn't another video where someone tries to scam you out of money, because I respect you and want to earn your trust.
	I'm not promising hundreds of millions of dollars; that's unrealistic.
	However, I assure you that this has been the plan from the very beginning.
	Our remedy has successfully passed all clinical trials and has been recognized as the most effective treatment for joints.

Table C.2: Example sentences for manipulation techniques from language style related group.

technique	examples
ease of reaching the goal	Click Start and earn 500 euros every day.
	Anyone with a smartphone can earn money passively.
	In reality, my team will do everything instead of you.
	It really is that simple.
	You don't need any computer or programming skills.
offer's luxury and limitation	Spaces are limited. Only one hundred and fifty lucky individuals.
	Out of 100 invited people, only the 50 fastest and most ambitious will make it through the competition.
	You're one of the few who will get a chance to change your life.
	The number of spots on the platform is limited, so act quickly.
	I've sent this invitation to exactly 100 people.
risk minimization	Many people have already tried this method, and it is 100% safe.
	The method is completely legal and ethical.
	There is a possibility of recovering lost funds.
	It's the safest way to invest in Poland.
	It's a simple game with a high probability of winning.
giving a promise	I guarantee that if you join me today, you'll earn at least 4000 zlotys within the next 24 hours.
	Now every Pole will be able not only to supply their home with gas but also to earn from four hundred euros a day through the company's shares.
	By the end of the week, you'll have over 3000 euros in your account.
	You won't regret coming across this video today.
	From this moment on, your new, happy life will become a reality.
call to action	Submit an official application for financial independence and start earning today.
	You urgently need to click on the link below to get the real cure.
	Press the "Read More" button.
	To do this right now, click the "Learn More" button under the video, read the article about this product, and get the remedy that will truly heal all your aching joints, get rid of back and knee pain, and give you a second youth.
	To take advantage of it, simply register on the project's website by providing your details.

Table C.3: Example sentences for manipulation techniques from motivational group.

technique	examples
altruism	It can help people. Really help people.
	Because the project is of a social nature and aims to help the residents of Poland improve their well-being during times of rising global inflation.
	Buddha is a well-known philanthropist. His goal is to give Poles a chance to win ten million zlotys in his online casino.
	In this way, Elon plans to start a global fight against poverty...
	I don't want anyone to experience knee or back pain like I do. I want everyone to be healthy and live a happy life.
targeting normal people's needs	Think about your new home, a new car, your dream trip, the thousands of opportunities that these money could open up for you.
	If, like me, you're a mom and you'd like to add to the household budget while staying at home, I invite you to collaborate.
	This isn't just an investment; it's a step towards financial independence.
	If you're unemployed and at a last-chance crossroads, this offer is for you.
	Regardless of your personal situation, if you want to earn a million zlotys and leave the rat race behind, this offer is for you.
direct talk	I'm just an ordinary guy, like you.
	You, just like me, can get a real medication.
	If you're watching this video, it means you've noticed it yourself.
	I openly admit it, and that's why I'm challenging you.
	She's just like you, so if it worked for her, it will work for you too.
social endorsement	My mom is living proof of its effectiveness.
	Join those who have already experienced its power and forget about blood pressure problems forever.
	Each of them earned over 16,000 zlotys.
	So far, 85,000 Poles have received payments from our program.
	Yes, indeed. For several months now, my acquaintances and I have been earning directly on our mobile phones thanks to this app.
addressing possible concerns	It's best to confirm it with numbers.
	Before you leave thinking I'm talking nonsense, take a moment and listen to me.
	This isn't another promise of easy money.
	I want to assure you that we've just decided to test the app ourselves to avoid risking ordinary citizens' money.
	We had to test everything in detail because we didn't have the moral right to test it on our own citizens.

Table C.4: Example sentences for manipulation techniques from sociological group.