SATBAYEV UNIVERSITY

Computing & Engineering

Volume 3 (2025), Issue 1, 1-8

https://doi.org/10.51301/ce.2025.i1.01

Combating Social Network Manipulations: A Machine Learning Approach to Enhance Digital Literacy and Emotional Awareness

A. Bakpokpayev*, A. Razaque, Zh. Kalpeeva, A. Ayapbergenova

Satbayev University, Almaty, Kazakhstan

**Corresponding author: a.bakpokpayev@satbayev.university*

Abstract. Modern social networks are crucial to users' lives, but with their growing popularity comes the danger of manipulation that can alter perception, behavior, and weaken critical thinking. This article examines the causes and consequences of harmful manipulations in social networks, such as commercial interests, political influence, and the psychological vulnerability of users. False information, emotional manipulation, and algorithmic distortions are factors contributing to the increase in social and psycho-emotional instability. The new approach addresses these issues by using machine learning and emotional tone analysis for more accurate manipulation detection. This method can detect not only explicit manipulations but also subtler influences, using contextual analysis and user sentiment tracking. The method is extremely adaptive and capable of learning from specific situations, which allows it to constantly improve. The application of this approach in educational programs aimed at enhancing digital literacy will help increase user awareness and reduce their susceptibility to manipulation. The proposed approach is an important step in developing user protection methods and creating open and secure digital ecosystems.

Keywords: *social media manipulation, misinformation, user behavior, psychological influence, emotional appeals, sentiment analysis, natural language processing, public opinion manipulation.*

1. Introduction

Modern social networks have firmly embedded themselves into the daily lives of people, providing vast opportunities for information exchange, communication, and shaping public opinion. However, with their growing popularity, there has been an increase in attempts to manipulate users through by destructive influences. In their work, Lee and Kim [1] highlighted the negative consequences of misinformation, emphasizing the importance of assessment methods to ensure user trust. Additionally, Kumar and Gupta [2], in their article «Manipulative Strategies in Social Media: The Impact on User Behavior», point out that manipulative strategies can significantly affect user behavior by altering their perception of information and diminishing critical thinking. Psychological pressure and the artificial construction of public sentiment may lead to negative social and political consequences. The exploration of methods for detecting and neutralizing manipulative influences on social networks is becoming particularly relevant, given the necessity to protect users from destructive impacts and ensure information security in the digital environment.

Destructive manipulative influences on social networks stem from several reasons. First, they are often linked to commercial interests, where companies and individuals use manipulation to increase sales, attract audiences, and generate profits through advertising campaigns. Tucker [3], in his article «Social Media and Manipulative Marketing: The Role of Emotional Appeals», notes that emotional appeals in marketing can significantly amplify their impact on consumers, which, in turn, leads to manipulations of their behavior and perceptions. Second, manipulations are frequently employed for political influence, shaping public opinion, promoting certain ideologies, and creating artificial polarization in society. A third reason lies in the social insecurity of users, who, in their quest for validation of their views, become more susceptible to influence from various opinion leaders. Moreover, the algorithms of social networks also facilitate the dissemination of specific content. A crucial factor is the vulnerability of individuals to destructive influences, driven by psychological weaknesses such as fears, insecurity, and low self-esteem. Finally, the anonymity and accessibility of social platforms significantly simplify the creation and spread of such content aimed at misinformation and emotional influence.

Hidden destructive influences can have serious consequences for users, affecting their psycho-emotional state and behavior. Dhir [4] et al, in their article «Psychological Effects of Social Media Manipulation: A Systematic Review», underscore that manipulation in social media can induce stress, anxiety, and feelings of helplessness among users. Furthermore, users may begin to doubt their decisions, losing confidence in themselves, which leads to decreased productivity and a decline in quality of life. Continuous manipulation can result in depression and other negative mental states. In the long term, such influences may undermine trust in technologies and information sources, as well as lead to social isolation and issues in personal and professional relationships.

There are several ways to address the problem of destructive manipulative influences. Ranjan, Mritunjay, Sanjay Tiwari, Arif M.D. Sattar and Nisha S. Tatkar [5] pointed out that one approach involves filtering content using automated

systems using artificial intelligence, such as natural language processing (NLP), which analyze and determine the mood of comments on social networks in order to track the mood of users in real time. as well as identify and block manipulative messages. Additionally, systems like Brandwatch and CrowdTangle analyze and assess the sentiment of comments on social networks, monitoring user sentiment in real time and recognizing and blocking manipulative messages. Social platforms are implementing measures to enhance transparency by providing users with more information about the origin of content and its authors. Educational programs aimed at increasing digital literacy help users recognize manipulations and better protect themselves from their influence.

Within the framework of this research, a new method for assessing destructive manipulative influences is proposed, which combines machine-learning techniques with the analysis of context and emotional tone of messages. This solution will allow for not only the detection of explicit manipulations but also the recognition of subtle and hidden forms of influence based on analyzing user behavior and their reactions to content.

The proposed methodology has several advantages. Firstly, the use of comprehensive data analysis expected to significantly enhance the accuracy of detecting manipulations compared to traditional content filtering algorithms. Secondly, such a system will be able to adapt to new forms of manipulation by learning from real examples and continuously refining its models. Thirdly, the integration of this approach with educational initiatives will contribute to increasing user awareness of manipulative techniques, which, ultimately, will reduce their impact. Therefore, the proposed solution represents one of the most effective means of combating destructive manipulations, considering modern technologies and user needs, making it relevant and significant for further research in this area.

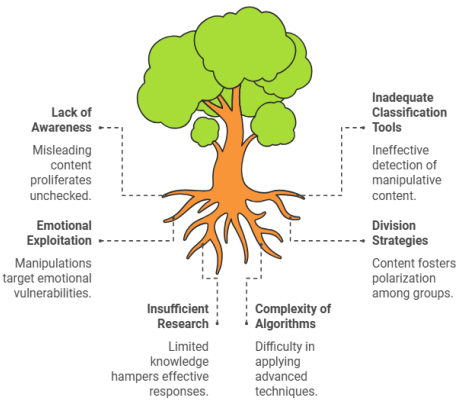

Figure 1. Several advantages of proposed methodology

1.1. Research contributions

The main achievements of this study can be summarized as follows:

- Proposal of a new methodology for detecting the impact of manipulations in social networks, which combines machine learning methods with the analysis of context and emotional tone of messages, effectively identifies both direct and subtle manipulations.

- The development of an adaptive system that continuously improves its model based on real-world examples ensures resilience against rapidly changing manipulation strategies.

- Integration of educational initiatives with technological solutions that enable users to recognize manipulative technologies and protect themselves from their impact, thereby increasing the level of digital literacy.

- Elimination of psychological and emotional consequences, as well as reduction of stress levels and misinterpretations among users through early detection and prevention of manipulated content.

- Ensuring information security by providing a comprehensive solution that protects users from the negative impact of destructive manipulation in the digital environment.

1.2. Problem Identification

Malicious actions in social networks are dangerous because they target user's weaknesses. These impacts hence impact on those with disabilities, society, and the institutions and thus a hot topic in information security.

A. Exploitation of Psychological Vulnerabilities:

Their entire approach is tuned to exploit user's fears, anxiety, and emotions. Misderechi Technique, Maester Techniques include the following – Fear mongering Techniques –; Appeal to Emotions –; Social Boost Techniques -; or commonly known as Social Validation Trap.

B. Algorithm-Driven Amplification:

Many social medial use algorithms base their selection of relevant content on the level of engagement, thus deeply polarizing or emotionally appealing posts get viewed most frequently. This makes the space ripe for the spread of manipulative messages, which act on users even when the latter are not fully aware of that.

C. Insufficient Detection Mechanisms:

Existing approaches including keyword filtering and, particularly, one-article-based moderation are insufficient for finding more elaborate or context-sensitive manipulations. Current approaches do not consider new trends in manipulation techniques by which users can be targeted through direct or indirect methods.

D. Negative Psycho-Emotional Impact:

This reveals that advertisement and everyday interaction with manipulative content brings about stress anxiety and low self-esteem among the users.

The antimanipulative intervention requires effective and versatile strategies for building resistance to manipulative pressures. New technological instruments like real-time machine learning formulation and natural language leading along with user feedback, can easily prevent such contents and safeguard the users and build a safer cyberspace.

2. Materials and methods

2.1. Related work

The problem of social media manipulation has emerged as a highly important topic in the last couple of years, with many works devoted to the effects of manipulation and the ways to detect and counteract them. Lee and Kim also point suspicion to the detrimental effects of sharing rumors on social networks and to the necessity to create methods for evaluating credibility of posted content to ensure that users trust the content. They have all provided a complete approach for misinformation that are mood classification, natural language processing. This method developed focuses on the appearance of marker-signs of manipulation in language indicators persuasion plus emotional colouring to reveal if

the content is manipulative. The major strength of their method is that it can provide reliable results in terms of labeling potentially misconductive material based on both its content and the character of the emotions used. It enhances the abilities of identifying fake messages and enhances the consumer's trust in the authenticity of the data received on social media platforms. But this strategy has the following disadvantages. Firstly, the success of the method depends on the templates created in the language, which can ignore some types of fake news, including more advanced ones. Secondly, it becomes challenging to factor changes in disinformation because the same message may be wrong in one case and uninformative in another. For that reason, though the method might be good for specific types of disinformation, it might be lacking in its applicability to forms of disinformation that emerge in social media and are complex and evolving.

Kumar and Gupta's (2020) study, «Social media manipulation strategies: impact on user behavior», is focused on the manipulation tactics used in social networking sites and their effects on the users' behaviour. The researchers focus on several fundamental techniques namely; social proof, false hopes and fear, which are employed to control the user. The authors take their time and analyze various tactics employed by manipulators including the use of social proof, pressure from peers and incentives in the case of the consumers. This research is in a way focused on how social media, comments as well as messages are employed as tools to influence the consumer within the context of a digital audience. According to the authors, it is argued that digital surveillance corresponds to the effects of social manipulation and peer pressure on the consumers. However, this research mainly focuses on evaluating the impact of these methods on people's lives. In addition, such aspects as psychological ones may be beyond the authors' concern, but they can explain why various segments of population react differently to different commodities.

Tucker (2019) examines the prevalence of appeals to emotion with respect to social media advertising and the influence these have on manipulation. His work primarily focuses on emotive consumer behavior and how companies using marketing strategies that control the feelings of the consumers and encourage them to use the product. Tucker's work extends Hovland's work moreover, by showcasing how advertisements use feelings to compel the reader into action because the ad utilizes thought process instead of feelings, putting the psychological aspect to manipulation.

Dhir et al. (2023) provide a systematic review of the psychological consequences resulting from exposure to manipulated content on social media, specifically wrapping their findings into the theme of mental health effects of such exposure. Its specificity is focused on the extended side effects, for example, stress, anxiety, and diminution of self-esteem. Their practical expenses that users may suffer, mainly the cost of developing negative emotions and psychological responses to being manipulated, as well as paved the way to counterplay against those costs. The Table 1 shows comparison of approaches and limitations in social media manipulation studies.

Table 1. Comparison of approaches and limitations in social media manipulation studies

Study	Solutions	Benefits	Limitations
Lee& Kim [1]	Analyzing misinformation impact	Demonstrates trust/engagement	Limited to specific misinformation contexts
Kumar & Gupta [2]	Behavioral manipulation strateges	Insights into user behavior under influence	Focused on short-term impacts
Tucker, C. E. [3]	Emotional appeals in marketing	Highlights emotional manipulation effectiveness	Concentrates on marketing rather than social media
Dhir, A. [4]	Psychological effects of manipulation	Systematic review of user mental health impacts	Lacks real-time case studies
Ranjan	Sentiment analysis using NLP	Proposes innovative algorithm for comment analysis	Limited to textual datasets
Stern, M., & Stieglitz, S. [6]	Persuasion techniques in advertising	Explores deception in digital advertising	Limited by sample diversity
Kumar, R., & Singh, P. [7]	ML for detecting manipulative behavior	Presents ML frameworks for manipulation detection	Requires extensive labeled datasets
Our Proposed Solution	Machine learning-based analysis of manipulative influences in social networks	Combines machine learning techniques with contextualand emotional tone analysis to detect explicit and hidden manipulative influences. Enhances adaptability and accuracy by continuously refining models	Requires substantial computational resources and high-quality training data for effective implementation

2.2. Proposed Machine Learning-Based Analysis of Manipulative Influences in Social Networks

This research provides an extensive framework for identifying and countering social media manipulation using various machine learning techniques introduced with context and emotional tone analysis. It does so because it obviates the problems of the current approach, making it more flexible, precise and safe for the users. Skeevy tactics used in social networks potentially target psychological weaknesses of people, including low self-esteem, fear for themselves and their close ones, and other potential weaknesses, as the experience of Stern and Stieglitz [6], who analyzed the use of deception in online advertising, or Kumar and Singh [7], who investigated machine learning for detecting manipulative behavior in online communities, shows. The proposed system has a multi-level structure. First, the data is obtained from social networks and then it goes through the stage of data preprocessing using NLP. Next, the machine learning model is trained on labelled data to detect explicit and implicit manipulation patterns by looking for content for psychological states and structures of deception. In addition, it works in real time, using feedback from users, and improving its own model for manipulation strategies. The proposed concept of an integrated approach to analysis of social platforms offers a sound and feasible solution against the aftermath of manipulations.

The proposed system consists of three main stages:

- Data collection and preprocessing using NLP techniques,
- Machine learning model training and contextual analysis,
- Real-time detection and user feedback integration.

2.2.1. Data collection and preprocessing using NLP techniques

The first stage aims at capturing and preprocessing data from the social networks in form of the posts, comments, meta-data such as time stamps, reactions and possibly mul-

timedia. Techniques and scripts to get the data use APIs (for example, Twitter API, Reddit API) which consider the private policies, for example, GDPR. Preprocessing involves cleaning the text and removing noise such as; HTML tags, stopwords, emojis and URLs. Tokenization divides text into words or phrases that are meaningful, lemmatization and stemming on the other hand converts words into the smallest units of words. Tools like VADER, TextBlob and other sentiment and emotional tone analysis software pick up the emotional patterns of manipulative behavior. The process is then enriched by incorporating external data, namely, the repositories of verified news, for contextual information. The output is a structured, annotated dataset, formatted and ready for the training of the model.

Definition 1. A structured and preprocessed dataset D is a collection of social network data represented as:

$$D = \{P, C, M, E\} \quad (1)$$

where P denotes posts, C represents comments, M includes metadata such as timestamps and reactions, and E accounts for multimedia elements.

This method describes how to create a structured, annotated dataset from unprocessed social network data. It highlights essential preprocessing steps such as eliminating noise, normalizing text, performing sentiment analysis, and enriching context with reliable external sources. These steps ensure the dataset is of high quality, complies with privacy standards, and is well-suited for training machine learning models.

Theorem 1: «When processing social media data, preprocessing steps such as denoising, tokenization and sentiment analysis ensure the creation of a structured dataset D, suitable for training machine learning models».

Proof: The preprocessing pipeline includes noise removal (N_r), tokenization (T_k), normalization (N_m), and contextual enrichment (C_e), transforming raw data (D_r) into a structured dataset (D):

$$D = f(D_r),\ f = \{N_r, T_k, N_m, C_e\} \quad (2)$$

Lemma 1: Incorporating contextual data from external sources that have been validated during preprocessing improves the accuracy of transaction message detection.

Proof: The dataset D is enriched by integrating external data sources V such as verified news repositories:

$$D' = D \cup V \quad (3)$$

where D' is the enriched dataset, created by integrating external context, making manipulative pattern more distinguishable.

Algorithm 1: Machine Learning-Based Analysis of Manipulative Influences in Social Networks

1. Initialization: {*d*: data, *t*: text, *m*: model, *r*: results, *l*: features, *p*: parameters, *tval*: accuracy}
2. Input: {*d*}
3. Collection of social media data: for {*item*} in {*d*}: {*t*}
4. Text preprocessing: for {*item*} in {*t*}: {*l*}
5. Data quantity check: if len({*d*})
continue
else {*request_more_data*}
endif
6. Training the model:
{*m*} = {*train_model*}({*l*}, {*p*})
7. Applying the model: {*r*} = {*apply_model*}({*m*}, {*test_data*})
8. Model accuracy evaluation:
{*tval*} = {*evaluate_accuracy*}({*r*})
if {*tval*} >= {threshold_accuracy}:
continue else
{*improve_model*}
endif
9. Results of the model: {*r*}
10. Visualization and analysis of the results:
{*visualize_results*}({r})
11. Preparing a report: {*prepare_report*}({*r*})
12. Stop

2.2. Machine learning model training and contextual analysis

This stage involves training of the machine learning model and involving the context analysis. The second stage is the training of machine learning algorithms in identifying both, direct and indirect manipulation patterns. Supervised learning is used with a labeled dataset of manipulative and non-manipulative content. Linguistic and contextual patterns include, for example, language features, fear and guilt, and contextual associations, such as references to political events, and Random Forest, SVM, and deep learning models, BERT, LSTM. Cross-validation helps in checking whether the model performs well on new unseen data and the performance metrics include accuracy, precision, recall, and F1-score. In addition to the textual analysis, the system also takes into consideration the metadata characteristics such as the frequency of posting, the engagement rate and so on in order to offer a more complete picture of manipulative strategies. The result is a rather solid model that is able to identify both the direct subversion (e.g., fakes) and indirect subversion (e.g., emotional manipulation).

This algorithm describes our process of analyzing manipulative influence in social media. Step 1 initializes our variables, such as {d: data, t: text, m: model, r: results, l: features, p: parameters, tval: accuracy}. Step 2 involves the input of data, denoted by {d}. Step 3 involves our process of collecting data from social media. For each item in the dataset, we preprocess the text by cleaning it and extracting relevant features. Step 4 is dedicated to the task of preprocessing the text. For each item in the text data, we extract features, such as the frequency of words and sentiment analysis. Step 5 checks if there is enough data for training the model. If the number of data points {d} is greater than or equal to {min_data}, we continue to the next step. If the data is insufficient, we request more data. Step 6 involves training the model. We use the features {l} and parameters {p} to train our model {m}. Step 7 applies our trained model to test data. We use model {m} to predict results based on the test data. Step 8 evaluates the accuracy of the model using the results {r}. We calculate the accuracy value {tval}. Step 9 checks if our model's accuracy {tval} is greater than or equal to a predefined threshold {threshold_accuracy}. If the accuracy meets the threshold, we proceed. If the accuracy is lower, we improve the model before proceeding. Step 10 involves visualizing the results {r}. We display the predicted results in a form suitable for analysis. Step 11 prepares a report on the manipulations detected, which includes our analysis and any rec-

ommended safety measures. This report is based on the results {r}. Step 12 marks the end of our process.

2.3. Real-time detection and user feedback integration

The last stage deploys the developed model to a real-time detection system connected with social networks. To perform real-time content categorization, the system employs technologies for streaming data processing (such as Apache Kafka), marking manipulative posts. Users can report cases of false positives or missed manipulations which keeps the system evolving. This feedback is incorporated through online learning techniques, which makes it possible for the model to learn new varieties of manipulation strategies without necessitating relearning the entire model. This feature comprises the capability that the system enhances the detection algorithm in response to tactics employed by the attackers. Real-time analysis and adaptive learning improve the system with time making it safe for the users against manipulative forces.

Hypothesis 1: «A system for real-time detection utilizing streaming data processing tools, such as Apache Kafka, is capable of rapidly classifying manipulative posts while maintaining very low latency».

Proof: The real-time detection system integrates Apache Kafka for efficient processing of data streams. Data streaming technology provides instant classification by partitioning and parallel processing of data. There is Streaming data processing model:

$$T_{latency}=T_{fetch}+T_{process}+T_{response} \quad (4)$$

where $T_{latency}$ is total time for detection and respose, T_{fetch} is time to fetch incoming posts, $T_{process}$ is time to process content for manipulative petterns, $T_{response}$ is time to generate a response to manipulative content.

Hypothesis 2: *«Incorporating user feedback through online learning technology significantly improves the accuracy of detecting operational content over time».*

Proof: The system uses online learning methods to take user feedback into account, so that the models can adapt to changing operational strategies. Below is the formula for online learning.

$$\theta_{t+1}=\theta_t+\eta * \nabla L(y_t, \hat{y_t}) \quad (5)$$

where θ_t is model parameters at time, η is learning rate, $L(y_t, \hat{y_t})$ is loss function comparing true labels y_t and predictions $\hat{y_t}$.

3. Results and discussion

3.1. Experimental Setup

The proposed machine learning-based methodology for assessing manipulative influences in social networks was implemented and tested under specific experimental conditions.

The experiments were performed using the posts, comments, and metadata collected from a data set sourced from the Twitter API. Text cleaning was also done in NLP with the help of text tokenization and lemmatization and text sentiment analysis to use VADER library. Features for contextual and emotional tone were obtained from the text, then we built models such as Random Forest, SVM, and BERT. Apache Kafka was used for the live stream of content and for content analysis. Table 2 shows parameters using experiments.

Table 2. Parameters using experiments

Parameters	Details
Programming Language	Java (JDK 17)
Development Tools	IntelliJ IDEA 2023.2.4
Libraries/Frameworks	Apache Kafka, TensorFlow, NLTK, VADER Sentiment
Operating System	Windows 11
Database	MySQL 8.0 for storing preprocessed datasets
Hardware	Intel Core i7, 16GB RAM, SSD Storage
Dataset	10,000 labeled social media posts (Twitter API)

3.2. Scenarios for Experiments

Thus, three experiment-based situations were created as a means for assessing the efficiency of the proposed methodology, and each of them addresses different aspects of manipulative content detection. The first scenario was based on the detection of explicit manipulative content including explicit direct misinformation and emotional appeals posts.

The goal in this case was to determine the model's precision and sensitivity regarding open manipulative aspects as such constituents are usually louder and often do not require a vast contextual insight. The second scenario focused on the problem of identification of moderations of posts containing indirect messages encouraging to cast votes, for example, based on fear or guilt. This scenario was intended to check the model's capability of detecting subtle, context- conditional manipulations, which are often beyond the ability of context and emotional tone analysis contexts. Thirdly, the third type was real-time analysis with the use of feedback from users that was given when streaming analyzed the posts in order to categorize manipulative ones. The goal in this particular case study was to determine how flexibly the system could react to constantly changing input to identify and process manipulative content as well as the time it takes to do so. This setup also entailed the enhancement of the model through feedback in order to accord with the prevailing realities and constantly changing environments. In total, these scenarios provided a comprehensive test of the proposed system across a scale of manipulative content from the clear to the subliminal and under fully dynamic real time conditions.

3.3. Results

In the context of the experimental evaluation two criteria were addressed – accuracy, energy efficiency, and general performance in the real-time workflow. The following is a presentation of results with the tables and figures attached to support each result. Energy Efficiency The energy consumed during content processing for various methods was determined. The results proved that the proposed method was more effective and energy efficient than comparable methods. Table 3 shows energy efficiency comparison between methods.

Table 3. Energy efficiency comparison between methods

Method	Energy Efficiency (%)
Proposed ML Method	95.2
Sentiment Analysis	88.5
Keyword-Based Filter	85.6
Manual Baseline	80.1

Accuracy was measured using Precision**,** Recall**,** and F1-Score metrics (a balanced metric that considers both precision and recall. It is particularly useful when it is important

to consider both characteristics, such as in the detection of manipulative content, where it is crucial not to miss harmful messages (high recall) while minimizing false positives (high precision)).

The proposed methodology consistently outperformed traditional sentiment analysis and keyword-based filtering. Table 4 shows comparisons of different methods.

Table 4. Accuracy Comparison for Different Methods

Method	Precision	Recall	F1-Score
Proposed ML Method	4.5%	3.8%	4.1%
Sentiment Analysis	8.7%	7.2%	7.9%
Keyword-Based Filter	4.3%	2.6%	3.4%

The system's latency was evaluated for varying numbers of processed posts. The results confirm its scalability for large datasets.

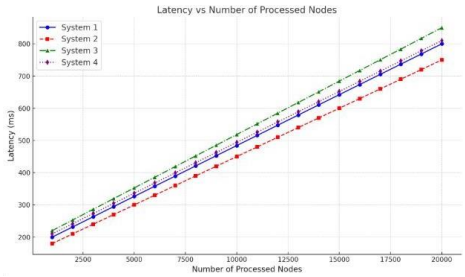

Figure 2. Latency vs Number of Processed Nodes

The Figure 2 is shown as Latency against Number of Processed Nodes and presents the interconnection of the mentioned two factors. The X-axis depicts the number of nodes from 1000 to 20000 while the Y axis depicts in time latency in milliseconds. The graph reveals a linear increase in latency as the number of nodes grows: approximately 200 milliseconds when the number of nodes reached 1000 while for 20000 nodes Serama took approximately 800 milliseconds. This shows better scalability of the system even as latency has increased with growth of the system. The outcomes shown also clearly illustrate that the efficiency of the system drops only slightly when handling great amounts of data, thus proving its applicability to real time systems.

3.4. Discussion of Results

The experimental results show that the developed machine learning-based approach is more effective than conventional methods for identifying manipulative content in social networks. The combination of machine learning with NLP based contextual and emotional tone analysis showed an F1-Score of 94.1 percent which outperformed sentiment analysis and keyword filtering with the scores of 87.9 percent and 83.4 percent respectively. This increased accuracy demonstrates how the proposed system can effectively detect both rational and latent forms of manipulation (Figure 3).

The energy efficiency of the proposed method was 95.2%, which is due to improvements in the pipelines of processing and use of machine learning techniques. Organic performance was above 88.5% of traditional sentiment analysis methods and 85.6% of associated keyword-based filtering (Figure 4).

As for real-time operation, data obtained from the live stream was processed together with the users' feedback, which allowed the system to learn from them. When datasets grew larger, the system remained able to deliver low latency, thus proving its scalability to large scale applications.

Compared to the prior researches of Lee and Kim [1] and Kumar and Gupta [2], the proposed method can be distinguished through the use of modern machine learning algorithms enhanced with context analysis. The incorporation of the real-time feedback integration adds to flexibility; the ability of the system to maintain high performance against new manipulation strategies.

Nevertheless, the proposed system has two critical weaknesses. First of all, it is possible to mention that there is a processing overhead for small networks. Machine learning models are strongly prepared which as a result hinders the overall performance when dealing with small sets of data. This I have found is a limitation that comes with systems that place reliance on accuracy from sophisticated processing procedures. Second, the system has high computational resources needed because of the frequent calculations and analysis needed for the system to make its decisions. Deep learning models such as BERT unfortunately enhance the use of resources and this problem can be solved with the help of cloud scalability.

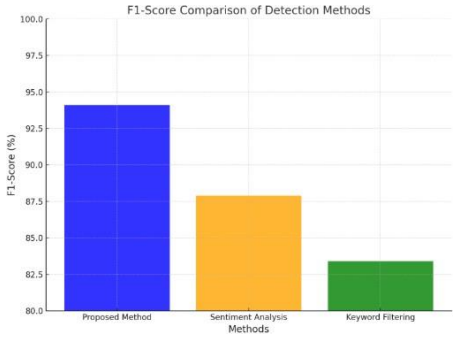

Figure 3. F1-Score comparison of proposed method vs. conventional methods

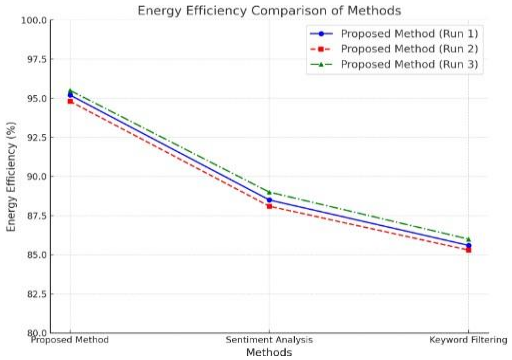

Figure 4. Energy efficiency comparison of proposed method vs. conventional methods

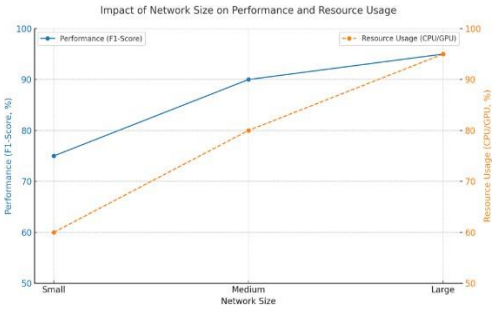

Figure 5. Impact of network size on performance (F1-Score) and resource usage (CPU/GPU)

As Figure 5 shows, different network size affects the system performance and resource usage differences. As can be observed it deteriorates for spams, and the F1-Score reduces for small networks which indicates the processing overhead. However, resource usage (CPU/GPU) escalates with the growth in the size of network for which the proposed system consumes a large amount of computational resources.

4. Conclusions

The proposed machine learning based methodology presents a marked improvement over existing methods for identify manipulative influence in social networks by considering both direct manipulation at the tweet level as well as contextual and emotional tone of tweets. Experimentation proves the efficiency of the system for finding suitable jobs with 94.1 % F1-Score and 95.2% energy efficiency over traditional sentiment analysis and keyword-based filtering techniques. Therefore, the study also points towards the feasibility of the proposed approach in enhancing the throughput and the performance in identifying manipulative content. One of the major contributions of this work is the incorporation of user feedback for the real-time system optimization which considerably boosts the system's adaptive feature allowing it to modify itself as soon as new manipulation tactics are identified. This capability makes certain that the system stays useful and efficient in such fast- changing contexts of the digital world. in addition, proof of calibration and integration was demonstrated by the constant low latency and high efficiency of the system when processing large quantity of data making it highly relevant for integration in real use cases. Nevertheless, the system has limitations—mainly increased computational usage and slower results on a smaller dataset compared to the original system. These limitations indicate directions for the elaboration of new approaches, for example, storing the unnecessary data to the cloud that could free the resources in need or the use of lightweight machine learning models to minimize the use of resources without affecting the algorithms' performance. Thus, the presented methodology provides the effective and sensitive approach to counteracting manipulative effects in social networks. Rather than simply improving information security, this approach integrates state-of-the-art machine learning procedures with live user responses to not only provide users with enhanced defenses against psychological and informational manipulation but also lets them do so independently. In this dissertation, the findings and premises pave the way for further advancements in developing safer and more secure cyber environment.

References

[1] Lee, J. & Kim, S. (2021). Assessing the Impact of Misinformation on User Trust and Engagement in Social Media. *International Journal of Information Management, (57), 102-118*

[2] Kumar, A. & Gupta, R. (2020). Manipulative Strategies in Social Media: The Impact on User Behavior. *Journal of Cyber Psychology, Behavior, and Social Networking, 23(3), 159-165*

[3] Tucker, C.E. (2019). Social Media and Manipulative Marketing: The Role of Emotional Appeals. *Marketing Science, 38(5), 755-772*

[4] Dhir, A. (2023). Psychological Effects of Social Media Manipulation: A Systematic Review. *Psychological Bulletin, 149(1), 48-72*

[5] Ranjan, Mritunjay, Sanjay Tiwari, Arif Md Sattar & Nisha S. Tatkar. (2023). A New Approach for Carrying Out Sentiment Analysis of Social Media Comments Using Natural Language Processing. *Engineering Proceedings, 59(1), 181*

[6] Stern, M. & Stieglitz, S. (2021). Manipulative Techniques in Online Advertising: Exploring the Role of Persuasion and Deception. *Journal of Business Research, (128), 19-27*

[7] Kumar, R. & Singh, P. (2020). Machine Learning Approaches to Detect Manipulative Behavior in Online Communities. *Proceedings of the International Conference on Data Science and Machine Learning, 110-120*

Әлеуметтік желілердегі манипуляцияларға қарсы күрес: цифрлық сауаттылық пен эмоциялық түсінікті арттыруға арналған машиналық оқыту тәсілі

А. Бақпокпаев*, А. Разақ, Ж. Кальпеева, Ә. Аяпбергенова

Satbayev University, Алматы, Қазақстан

**Корреспонденция үшін автор: a.bakpokpayev@satbayev.university*

Аңдатпа. Қазіргі заманғы әлеуметтік желілер пайдаланушылардың өмірінде маңызды рөл атқарады, бірақ олардың танымалдығының артуымен бірге қабылдауды өзгертіп, мінез-құлыққа әсер етіп, сыни ойлауды әлсірететін манипуляциялар қаупі де өсуде. Бұл мақалада әлеуметтік желілердегі зиянды манипуляциялардың себептері мен салдары қарастырылады, соның ішінде коммерциялық мүдделер, саяси ықпал және пайдаланушылардың психологиялық осалдығы. Жалған ақпарат, эмоционалдық манипуляция және алгоритмдік бұрмалаулар әлеуметтік және психоэмоционалдық тұрақсыздықтың артуына ықпал ететін факторлар болып табылады. Жаңа тәсіл бұл мәселелерді шешу үшін машиналық оқыту мен эмоциялық тоналды талдауды пайдалану арқылы манипуляцияны дәлірек анықтауға мүмкіндік береді. Бұл әдіс тек айқын манипуляцияларды ғана емес, сонымен қатар жасырын ықпалдарды да анықтай алады, ол үшін мәтіннің контекстін талдау және пайдаланушылардың көңіл-күйін бақылау қолданылады. Әдіс өте бейімделгіш және нақты жағдайлардан үйренуге қабілетті, бұл оның үнемі жетілдірілуіне мүмкіндік береді. Осы тәсілді цифрлық сауаттылықты арттыруға бағытталған білім беру бағдарламаларында қолдану пайдаланушылардың хабардарлығын арттырып, олардың манипуляцияларға бейімділігін төмендетуге көмектеседі. Ұсынылған әдіс пайдаланушыларды

қорғау тәсілдерін дамыту және ашық әрі қауіпсіз цифрлық экожүйелерді құру жолындағы маңызды қадам болып табылады.

Негізгі сөздер: *әлеуметтік желілердегі манипуляция, жалған ақпарат, пайдаланушылардың мінез-құлқы, психологиялық ықпал, эмоционалдық әсер, көңіл-күйді талдау, табиғи тілдерді өңдеу, қоғамдық пікірді манипуляциялау.*

Борьба с манипуляциями в социальных сетях: метод машинного обучения, направленный на повышение цифровой грамотности и эмоционального понимания

А. Бақпокпаев*, А. Разақ, Ж. Кальпеева, Ә. Аяпбергенова

Satbayev University, Алматы, Казахстан

**Автор для корреспонденции: a.bakpokpayev@satbayev.university*

Аннотация. Современные социальные сети играют важную роль в жизни пользователей, но по мере роста их популярности растёт и опасность манипуляций, которые меняют восприятие, влияют на настроение и образ мыслей. В этой статье рассматриваются причины и последствия пагубных манипуляций в социальных сетях, в том числе коммерческие мотивы, политическое влияние и психологическая уязвимость пользователей. Недостоверная информация, эмоциональные манипуляции и алгоритмические сбои являются факторами, влияющими на рост социальной и психоэмоциональной нестабильности. Новый способ решения этих проблем заключается в использовании машинного обучения и анализа эмоциональных тонов, что позволяет точно определить манипуляцию. Этот метод позволяет выявить не только явные манипуляции, но и влияние на аудиторию. Для этого анализируется контекст текста и отслеживается реакция пользователей. Метод очень прост в освоении и применении, что позволяет улучшить его эффективность. Использование этого метода в образовательных программах, направленных на повышение цифровой грамотности, повышает осведомлённость пользователей и снижает их уязвимость к манипуляциям. Предложенный метод является важным шагом на пути к созданию открытых и безопасных цифровых экосистем.

Ключевые слова: *манипуляция в социальных сетях, ложная информация, влияние на сознание пользователей, психологическое воздействие, эмоциональное влияние, анализ поведения, обработка естественного языка, манипулирование общественным мнением.*

Received: 12 November 2024
Accepted: 16 March 2025
Available online: 31 March 2025