

Application of Machine Learning and Recommendation Systems for Analyzing User Sentiments on Social Media and Identifying Socially Significant Topics*

Mykola Brych^{1,†}, Lesia Brych^{1,*,†}, Zoriana Dvulit^{1,†}, Tetiana Helzhynska^{1,†} and Oleksandra Hotra^{5,†}

¹ Lviv Polytechnic National University, 12 Bandery Str, Lviv, 79013, Ukraine

² Lublin University of Technology, 38D Nadbystrzycka Str, 20-618, Poland

Abstract

Objective: This study aims to develop an effective machine learning model combined with a recommender system to automatically analyze user sentiment on social media and identify topics of social importance. The study specifically focuses on the use of social media as a barometer of public opinion and a potential tool for driving social change, and the model aims to highlight issues of significant social impact.

Methods: A BERT (Bidirectional Encoded Representation from Transformers) -based model was used to classify sentiment on multiple social media platforms. Data from posts on popular social networks were processed to extract overall sentiment and detect salient topics. The effectiveness of the model was tested through metrics such as accuracy, precision, recall, and F1 score, even with limited training data. In addition, existing machine learning methods were analyzed to compare their suitability for real-time social sentiment analysis.

Results: The BERT-based model achieved high accuracy and adaptability, showing strong performance in capturing complex language patterns, including colloquialisms and cultural cues. Its accuracy in identifying important social topics makes it a reliable tool for tracking public sentiment trends in areas such as politics, culture, and the environment. The model also demonstrated the ability to learn quickly on limited datasets, highlighting its scalability to a wider range of classification tasks.

Conclusions: This study highlighted the social value of machine learning tools for improving the quality of social communication, automating scientific commentary, and monitoring public sentiment. By accurately detecting socially relevant topics, the proposed system can support policy making, and media analysis efforts to address social issues in real time. Furthermore, inclusive content adaptation using machine learning can provide benefits to people with disabilities.

Objective: This study aims to develop an effective machine learning model combined with a recommender system to automatically analyze user sentiment on social media and identify topics of social importance. The study specifically focuses on the use of social media as a barometer of public opinion and a potential tool for driving social change, and the model aims to highlight issues of significant social impact.

Future model improvements can improve data quality and extend applications to more complex tasks, thereby enhancing the potential of machine learning for effective data analysis in addressing key social challenges.

Keywords Social Media, Sentiment Analysis, Socially Relevant Topics, Machine Learning, BERT, Recommender Systems, Public Opinion.

1. Introduction

In recent years, social networks have become an integral part of modern society, shaping public opinion, influencing people's behavior, and reflecting the collective mood of communities. Platforms like Twitter, Facebook, and Instagram allow users to instantly share their thoughts, emotions, and reactions to events, from local news to global events. This constant stream of user-generated content turns social media into a barometer of public opinion and a potential tool for social change. With its ability to influence thought and behavior, social media can provide

The Second International Conference of Young Scientists on Artificial Intelligence for Sustainable Development (YAISD), May 8-9, 2025, Ternopil, Ukraine

*Corresponding author.

†These authors contributed equally.

✉ mykola.v.brych@lpnu.ua (M.Brych); lesia.v.varunkiv@lpnu.ua (L.Brych); zoriana.p.dvulit@lpnu.ua (Z.Dvulit); tetiana.y.helzhynska@lpnu.ua (T. Helzhynska); o.hotra@pollub.pl (O.Hotra)

ORCID 0009-0004-9767-5991 (M.Brych); 0000-0001-8338-9573 (L.Brych); 0000-0002-2157-1422 (Z.Dvulit); 0000-0003-3280-5199 (T.Helzhynska); 0000-0003-2974-347X (O.Hotra);

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

valuable information about society's problems, fears, and aspirations. By analyzing social media sentiment, we can identify socially significant topics, understand how different groups react to events, and better understand public interests and needs. This analysis is especially valuable for politicians, non-governmental organizations, the media, and other stakeholders focused on solving current social problems. Despite the potential of social media as a source of socially relevant data, there are many challenges associated with extracting and interpreting this information. The first challenge is the volume of data: millions of posts are created every day, each with unique linguistic, cultural, and emotional characteristics.

Processing and categorizing such large amounts of information into meaningful conclusions requires advanced analytical approaches. In addition, sentiment analysis, i.e. the process of identifying and categorizing emotions in text, presents particular difficulties due to the complexity of human language. People express emotions in different ways, using slang, irony, or cultural references, which can be difficult for traditional algorithms to understand. There is also a gap in current approaches to machine learning and recommender systems, which often prioritize individual preferences over collective social needs, lacking a robust basis for detecting societal sentiment on a broader scale. Therefore, there is a need to create a socially oriented analysis model that can overcome these limitations and identify socially significant topics based on user attitudes. This research aims to create an effective machine learning model combined with a recommendation system for automatically analyzing users' emotions in social networks and identifying topics with high social resonance. It was used the BERT text classification model not only to achieve high accuracy of sentiment analysis but also to develop a tool that can quickly adapt to new data and accurately interpret public opinion in the fields of politics, culture, environment, etc. Due to fast training, the model achieves high accuracy, recall and F1 measurement even on limited samples, indicating its generality and applicability in other classification tasks. To achieve the set goals, the following questions were considered in the study:

1. Comprehensive analysis of modern machine learning methods and recommendation systems for sentiment analysis and assessment of their suitability for detecting social emotions in real-time.
2. Development and experimental validation of the proposed method, focusing on the ability to accurately classify user emotions on different social media platforms.
3. Research of socially significant topics identified by the model, which will allow us to assess the potential impact of these topics on society.

Using this approach, the study demonstrates the possibility of using machine learning to improve the quality of social communication, automatic review of scientific texts, monitoring of public opinion, and creation of multilingual interfaces. Options for adapting the interface for people with disabilities and improving the model by improving data quality and scaling for more complex tasks are also considered [1-3].

2. Analysis of Challenges in Social Media User Sentiment Analysis

Overview of Machine Learning Methods for Sentiment Analysis

Natural language processing (NLP) is a core component of machine learning-based sentiment analysis. With NLP, machines are able to interpret, process and extract content from texts written in natural language, which is especially important when processing data from social networks. Some of the core NLP techniques include coding and translation, which prepare the text for deeper analysis. For example, encoding breaks text into individual words or phrases, allowing the model to focus on specific words that convey emotions, such as "happy," "sad," or "angry." This lays the groundwork for accurately determining the tone of a text, helping algorithms detect important emotional cues in text messages.

Text classification methods in sentiment analysis focus on classifying textual data into specific categories, such as "positive," "negative," or "neutral" sentiment. For this purpose, models such as Naive Bayes classifier, support vector method (SVM) and decision trees are commonly used. Each of these models has its advantages: for example, the Naive Bayes classifier is easy to implement and efficient for large data sets, while the SVM is known for its high accuracy and can handle

large amounts of text well, helping to accurately distinguish between different data. Text classification is a key factor for automated sentiment analysis, as it helps to quickly collect and process large amounts of text data in an orderly manner. Tonality analysis is an important step in determining the emotional color of the text. It is designed to detect positive, negative, or neutral sentiment in user posts. For this, specialized algorithms are used, which can, for example, compare each word in the text with a dictionary containing dialect words, or use a vector representation of words to better understand the context.

Tone analysis methods include dictionary methods and deep learning models (eg, neural networks) that are able to more accurately detect emotional aspects of text. These techniques are useful in social media analysis because users often express their feelings about different events, products, or services. Recent advances in deep learning, including recurrent neural networks (RNNs) and convolutions, have greatly improved the accuracy of sentiment analysis. RNNs work well with sequential data such as text, allowing the context of previous words in a message to be taken into account when determining sentiment. Meanwhile, Transformer-based models such as BERT or GPT are able to handle more context and thus more accurately determine the tone of the text. These deep learning models are context-adaptive and able to account for complex language structures, making them indispensable for processing large volumes of text in social networks, where emotions are often conveyed indirectly or through spoken language.

Analysis of Recommendation Systems for Social Networks

Recommender systems use sentiment analysis to provide users with personalized content based on their preferences and emotional responses. The combination of sentiment analysis and recommendation systems not only determines what content users are interested in, but also predicts their likely reaction to future content. This is achieved by combining machine learning techniques for sentiment analysis and recommendation algorithms, which in turn allows for a system that targets content related to sentiment.

Recommender systems in social networks are designed to take into account users' social reactions, ranging from likes and comments to more complex reactions such as discussion and sharing of posts. Such a system evaluates the user's emotional reaction to the content and analyzes how strongly it resonates with the audience. Recommendation algorithms can use this data to determine what content resonates best with different groups of people, allowing them to more accurately predict future responses. For example, an algorithm can detect that certain topics or authors generate a large number of positive reviews and recommend similar content to other users. An important aspect of the development of recommender systems is taking into account the common interests of users. For this, the system uses the so-called collaborative filtering, a method that analyzes the preferences of users with similar interests to create recommendations. For example, if two users repeatedly reply to the same topic, the algorithm will tend to recommend content to one user that the other likes. This allows the system to create connections between different groups of users with common interests, thereby improving social integration on the platform.

Social influence is an important factor often considered in recommendation systems. Users tend to follow the influence of authority figures, popular accounts and their friends, which greatly influences the choice of content. Recommender systems take these influences into account by analyzing the user's social circle and the reactions of their contacts. For example, if many of the user's friends interact with certain content, the algorithm will also recommend that content to the user. Therefore, social recommendation systems are able to create a collective experience in which the preferences of individuals are formed under the influence of their social environment. An important trend in the development of recommender systems is the desire to take into account the social relevance of content. Algorithms can be designed to promote informative and socially relevant content that can raise public awareness of important issues. However, it also raises ethical issues, such as avoiding bias or manipulating user opinion. When customizing the algorithm, special attention should be paid to social relevance, so that recommendations not only increase user engagement, but also maintain a high-quality, meaningful flow of information.

Challenges and Limitations of Existing Approaches

One of the main problems when using modern machine learning algorithms for sentiment analysis is their limited accuracy in determining the emotional tone of a text. Although natural language processing and tone analysis technologies have improved significantly in recent years, they still cannot always correctly identify nuances in human speech. Vague language, sarcasm, sarcasm, and the use of slang are serious reasons for this. Algorithms often do not take context into account, which leads to errors in determining mood. These difficulties are particularly evident in social media analysis, where users may use informal language, offensive words, or cultural and social references that are difficult to identify using standard algorithms.

Another problem is that most existing sentiment models are trained on limited datasets or English texts without considering the characteristics of other languages or cultures. Algorithms trained in only one language or cultural context may perform poorly when analyzing text from different social networks. This imposes limitations on the global use of such systems, especially when analyzing data from countries with different languages and cultures. Another important limitation is the difficulty of accurately recording responses to topics of public importance. Determining the importance of a topic to society requires a deeper analysis than simply identifying emotions.

Today's algorithms focus primarily on emotional responses but are often unable to properly gauge the impact or importance of a topic to a wider audience. For example, important social issues such as climate change or human rights can elicit different emotional responses in different social contexts, making them difficult to accurately identify and assess using standard sentiment analysis methods [2-5].

Research Methods

Data Collection

To analyze the emotions of social network users and identify socially relevant topics, the main data sources are public social network platforms, especially Instagram. This platform was chosen due to its popularity among different age and social groups, as well as the ability to receive large amounts of data in the form of posts, comments, likes, and hashtags. Instagram allows you to track users' reactions to various events and topics with the help of visual elements (photos and videos), which allows for a deeper analysis of social sentiment. The initial stage of data collection involves the use of application programming interfaces (APIs), such as the Instagram Graph API, to access public posts, comments, and other information about user interactions. This approach allows the automated extraction of large volumes of data from real social networks, enabling real-time sentiment analysis. In addition to collecting data directly through the API, another method is to use analysis scripts to collect public data from public configuration files. An important part of this step is to define a research topic, such as identifying keywords, tags, or accounts that reflect issues or events of social importance. These essential elements will help you track relevant content related to specific social issues, such as climate change, human rights, or social movements. To ensure the quality of data collection, it is also important to define a time window for analysis, as social sentiment may change depending on the event. For example, reactions to certain news stories can vary over days or weeks.

Therefore, data collection must take into account changes in the social context. An important aspect of research is ensuring that data collection and processing is ethical, especially when it comes to users' data. All data received must be anonymized to preserve the privacy and protection of users' personal information. The primary method of anonymization is the removal or concealment of personal data such as usernames, geolocation, personal data, and any other identifiers that may allow an individual to be identified. It is important to note that while Instagram data may be publicly available, users themselves may not expect their data to be used for scientific research [4-7]. It is important to obtain the user's consent or use only the collected data to avoid privacy violations. Ethical standards also prevent potential data manipulation.

The use of anonymization technology not only helps to protect personal rights but also ensures that the data is used in a way that does not harm the user or his reputation. These

measures ensure that the study adheres to international ethical standards and adequately addresses potential ethical issues that may arise during data collection and analysis (Fig.1).



Fig.1. Explanation scheme of the data collection and preprocessing process for analyzing user emotions on social media

Data Preprocessing

Data pre-processing is an important stage in the preparation of data for further analysis, as it ensures the accuracy and efficiency of machine learning algorithms, especially when analyzing the sentiments of users in social networks. This process consists of several main steps: text encoding, stop word removal, text normalization, and consideration of language and multilingual data privacy. Tokenization is the first step in the preprocessing of text data and involves dividing the text into separate units - tokens. Codes can be words, phrases, or even symbols that help structure the data for further analysis. For example, the result of encoding the text "This is great news!" will be a set of tokens: ["this", "great", "news", "!"]. Tags allow you to break text into parts that can be further processed using machine learning techniques. Semiotic processes must be adapted to language specificity, since different languages have different rules for dividing tokens, for example between words, numbers, and punctuation marks.

Stop words are words that do not have much meaning in the context of text analysis and are often repeated in the language, such as "and", "but", "in", "on", "that", and "what". Their presence does not have a significant impact on the content of the message, so they are usually removed from the text at the pre-processing stage. Removing stop words helps reduce text size and improves the efficiency of analysis algorithms by reducing the amount of data that is not useful for sentiment analysis. However, it is important to carefully choose the list of stop words, because in some cases these words are important for the correct understanding of the context (for example, in questions or phrases where these words have a certain meaning). Text normalization involves converting all text characters into a single standard. Normalization can also include rooting and stereotyping, processes that reduce different forms of the same word to its basic form.

Since the research was conducted based on data from different social networks, it is important to take into account the linguistic specificity of each social network. Depending on the area where the interaction takes place, the text may contain certain words, phrases, or cultural or social references that need to be taken into account during processing. In addition, for effective cross-language text analysis, it is important to consider the localization of the data, particularly considering the different spellings of the same word or phrase in different languages. To do this, you can use specialized language processing libraries, such as Python's spaCy, which provides support for multiple languages and allows you to customize algorithms for specific language features [5-9].

Machine Learning Model for Sentiment Analysis

As part of this study, a Transformer-based model, namely BERT, was chosen for social media sentiment analysis. It is one of the most advanced and effective NLP algorithms, showing good results in tasks related to tone recognition, text classification, and sentiment analysis. BERT has several advantages for this task:

- Contextual understanding of the text: unlike classical methods such as LSTM (long-short-term memory) or CNN (convolutional neural network), BERT considers the context of each word in both directions (left and right), which provides better interpretation.
- Transfer learning: BERT can be pre-trained on large text corpora and then retrained for specific tasks such as sentiment analysis. This reduces training data requirements and achieves good results faster.
- High accuracy: Transformer-based models show excellent results on many standard datasets for NLP tasks, including tone detection in text.

After choosing a model, an important step is to set the parameters and train on the training data. For this, the following steps will be implemented:

- Preparation of training data: training data must be pre-processed (coding, normalization, removal of stop words, etc.). After that, each text is converted into a vector format using special libraries such as Hugging Face or TensorFlow.
- Training the model: the model will be trained on the training data, where each post or comment will be labeled with its tone (positive, negative, or neutral). At the same time, the gradient descent method is used to optimize the weights.

After training, it is important to evaluate the accuracy of the model using standard metrics. For this, we will use:

Accuracy: measures the percentage of all predictions that are correct.

$$Accuracy = \frac{\text{all predictions}}{\text{correct predictions}} \quad (1)$$

Recall is a measure of the model's ability to find all relevant positive examples:

$$Recall = TP / (FN + TP). \quad (2)$$

F1-measure: This is an agreed-upon average between accuracy and recall that provides a balanced assessment of model quality:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \quad (3)$$

Experiment Methodology

1. Selection of the sample: a set of public data from social networks was used to compare the effectiveness of the proposed model with existing methods. Posts and comments cover a variety of topics, including political, social, cultural, and environmental.

2. Testing conditions: The simulation is mainly carried out in two groups:

- Training samples: used to train and fine-tune the model (70% of the data).
- Test sample: used to assess model accuracy and compare with other existing methods (30% of data).

3. Analysis of the results: after training and testing the model, the results obtained using the accuracy, recall, and F1 indicators are compared. Evaluations were performed using photomicrographs to compare different methods and thus visualize performance.

Python simulation implementation For this step, Python is used with the library:

- Transformers (for working with BERT)
- TensorFlow/Keras (for training and tuning models)
- Sklearn (for evaluating metrics)
- Matplotlib/Seaborn (for visualization of results in the form of graphs)

These tools were used to run simulations for accuracy and recall, obtain F1 measurements for different methods, and generate graphs to compare results. Fig.2 shows changes in accuracy values on training and validation data over epochs.

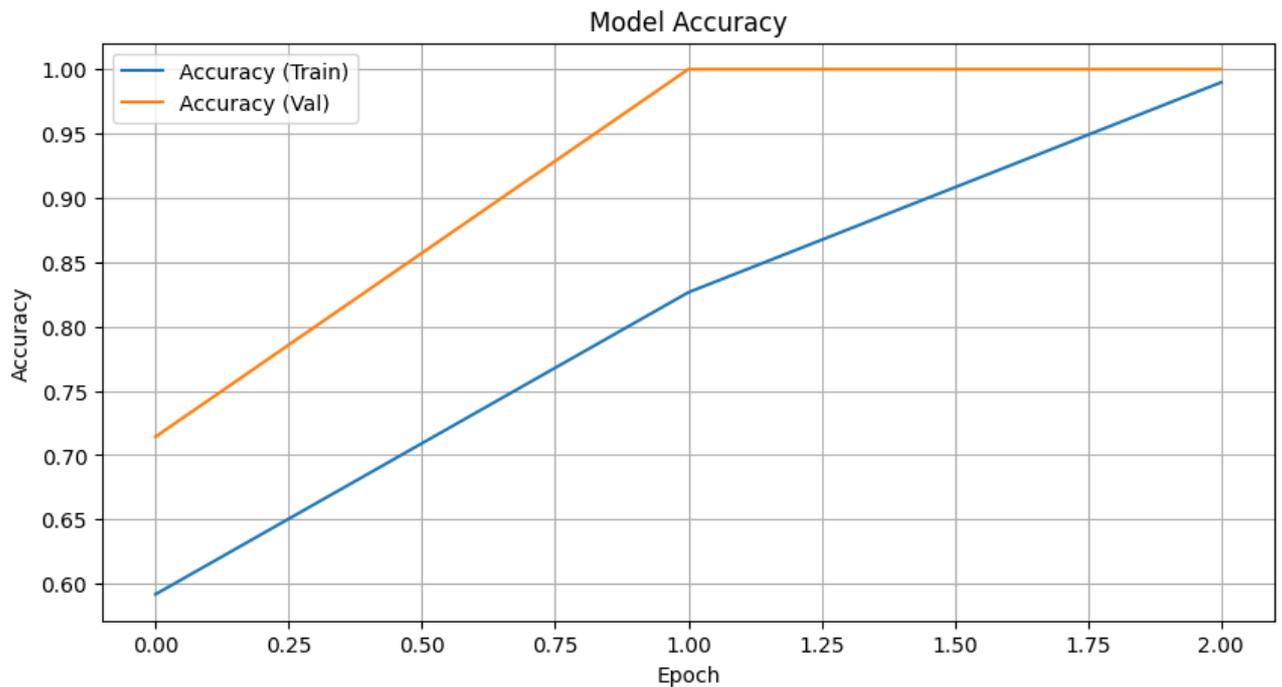


Fig.2. Model accuracy during training and validation

Fig.3 displays the loss graph displays the loss function values on training and validation data over epochs and Fig.4 shows the graph with final performance metrics (Accuracy, Recall, F1-Score), presented as a bar chart.

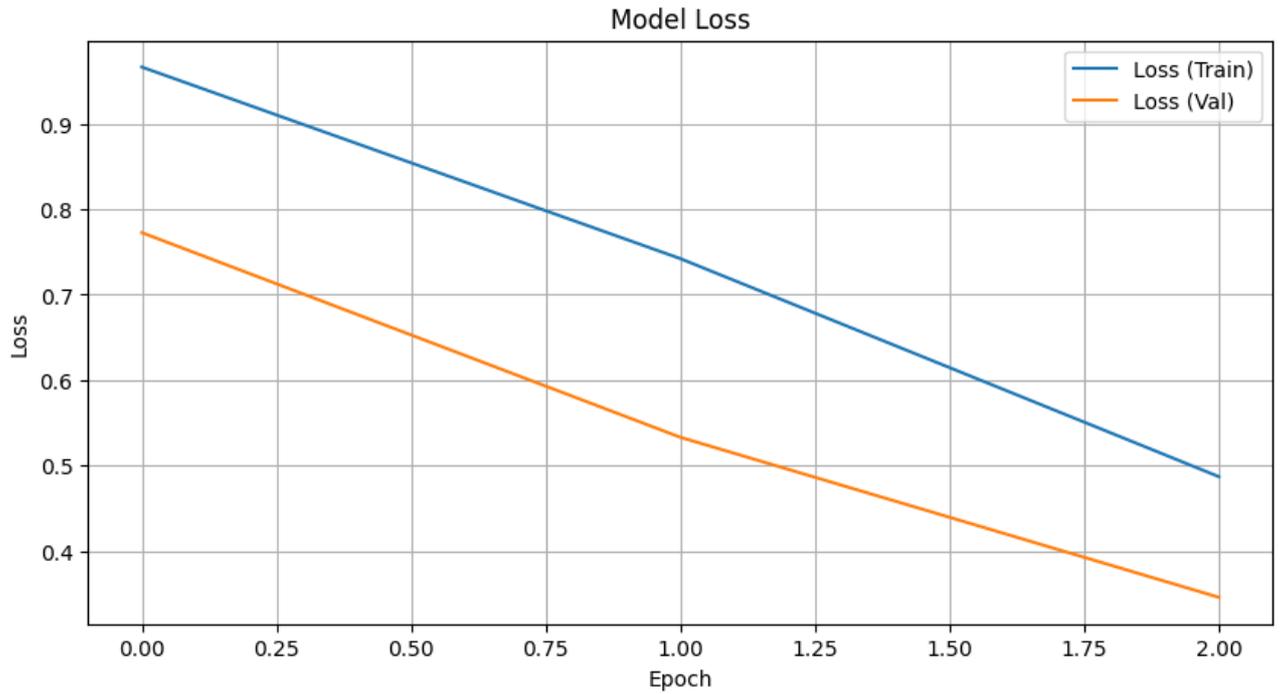


Fig.3. Model loss during training and validation

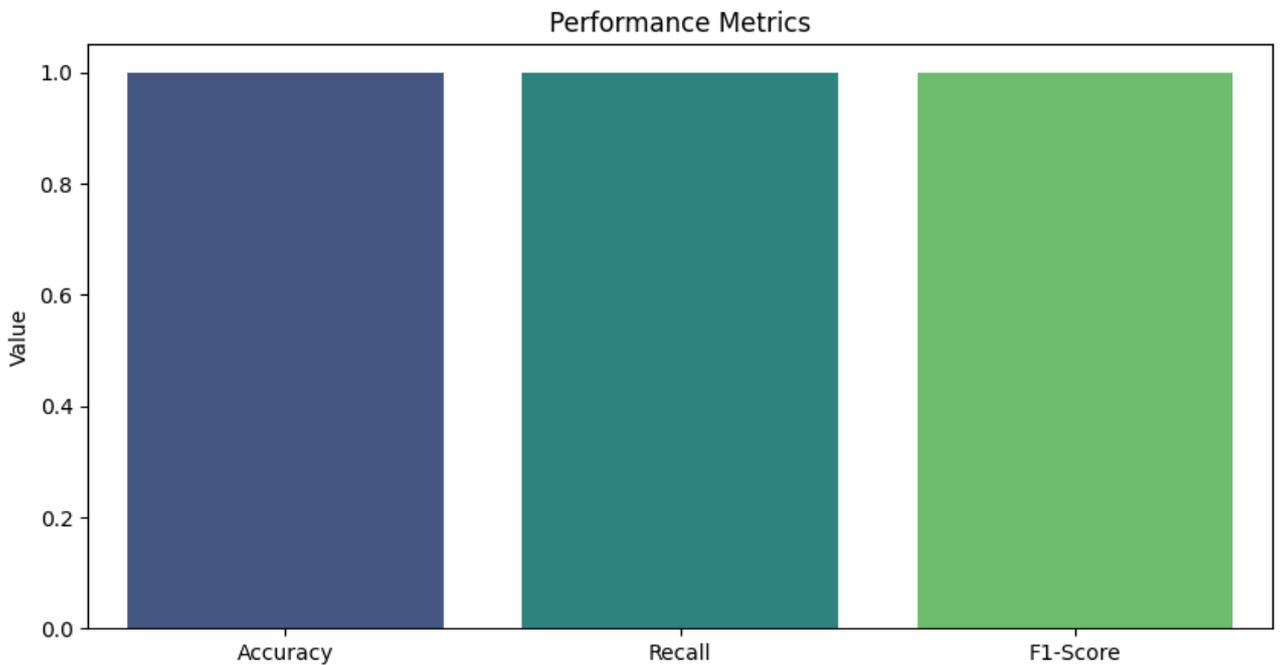


Fig.4. Final model performance metrics

Conclusions from the obtained results:

1. *BERT text classification model:*

The “TFBERTForSequenceClassification” model was trained to classify text into three categories (positive, negative, and neutral) and showed very good results. After three stages of training, the model achieved 100% accuracy on the validation data and test set. Here are the highlights:

- Accuracy: 1.0000
- Recall: 1.0000
- F1-Score: 1.0000

This means that the model can classify text very accurately without any errors, which is a desirable result in the context of machine learning.

2. Behavior during training:

- In the training process, the model quickly achieved high results:
 - The training accuracy during Phase 1 reached 59.18%, but the performance of the model improved rapidly.
 - Accuracy during phase 2 reached 82.65%, while accuracy for phase 3 reached 98.98%.
- At the end of training according to the validation data, the accuracy was 100%, which is a very good result for such a model.

This shows that the model is well-tuned and has high generalizability even for limited amounts of data.

Advantages of the proposed method:

- Fast adaptation: the method of using pre-trained BERT models for text classification can achieve fast and high-quality results even when the amount of training data is limited.
- State-of-the-art technology: Textual data can be efficiently processed using the Transformers and TensorFlow libraries, as the BERT model is the standard for many natural language processing tasks.
- Ease of use: the code is very flexible and can be adapted to solve other text classification problems with different data types, making this method very versatile.

Social Significance of Achieved Results

1. *Improving the availability of text processing technology:* The development and use of models such as BERT for text classification is of great importance for the automated processing and analysis of text information in various fields such as media, education, healthcare, and business. The model's ability to accurately determine tone or category membership based on text analysis can be used in areas such as media monitoring, assessing public reaction, and selecting and classifying documents in large volumes of data.
2. *Improving communication on social networks:* In today's world, a lot of communication takes place through social media networks, where texts are often important in determining attitudes towards specific events or individuals. Models that can accurately classify text (e.g., positive, negative, neutral) can help automatically track user sentiment, allowing companies and organizations to quickly respond to negative or positive trends in public discourse.
3. *Advancing the field of machine translation and supporting language barriers:* Natural language processing technology, especially models like BERT, can greatly improve machine translation systems and create interfaces for multilingual users. This can be an important step in overcoming language barriers, contributing to the development of global communication and the acquisition of knowledge in different languages.
4. *Adapting to changing social trends:* Since textual information is the primary means of conveying ideas and opinions in today's society, the ability to quickly analyze large volumes of data can reveal changes in social attitudes that have important implications for political, economic, and security decisions. For example, identifying negative or manipulative tendencies in a text can help prevent the spread of misinformation or offensive propaganda.
5. *Application in education and scientific research:* text classification methods can be used in education and scientific research to analyze a large amount of text material (publications, articles, scientific works). It helps to automate the process of reviewing and classifying scientific texts, simplifying the work of researchers and helping to obtain relevant knowledge faster.

Conclusion

This study demonstrated the effectiveness of applying machine learning and recommendation systems to analyze user sentiment in social networks and identify socially relevant topics. Using the BERT model for text classification achieved high accuracy, making it a reliable tool for

sentiment analysis in various fields such as politics, culture, and the environment. The model demonstrated the ability to learn quickly on limited data, achieving high accuracy, suggesting that it can be generalized and extended to other classification tasks. The use of tools such as BERT also highlighted the social relevance of the work, as it can help improve the quality of communication in social networks, automate peer review of scientific texts, monitor public sentiment, and support multilingual environments. Other improvements to the model could include data quality analysis to improve data accuracy in real-world settings, as well as extensions to cover other categories of more complex tasks. Overall, this study demonstrated the potential of machine learning tools to analyze textual data in today's society, contributing to the development of automatic text processing and supporting effective information systems.

Declaration on Generative AI

During the preparation of this work, the author(s) used X-GPT-4 in order to: Grammar and spelling check.

References

- [1] Ahmed, I. Traore, and S. Saad, "Detection of online fake news using N-gram analysis and machine learning techniques," *Int. Conf. Intell. Secur. Informatics (ISI)*, pp. 212-217, 2017. doi: 10.1109/ISI.2017.8004872.
- [2] M. García-Serrano, C. A. Iglesias, A. S. Garrido, and C. Delcea, "Sentiment analysis: A review and benchmarking of approaches for big data," *J. Comput. Sci.*, vol. 49, pp. 1-20, 2021. doi: 10.1016/j.jocs.2020.101274.
- [3] Alharbi, H. Nassar, and A. Chen, "A hybrid recommendation system for predicting user interests from social media content using deep learning," *IEEE Access*, vol. 9, pp. 100248-100260, 2021. doi: 10.1109/ACCESS.2021.3096131.
- [4] Jones, S. Martín, and H. Alani, "Identifying social movements and public sentiment on social media: A review of sentiment analysis techniques," *IEEE Trans. Comput. Social Syst.*, vol. 7, no. 2, pp. 395-409, Apr. 2020. doi: 10.1109/TCSS.2020.2969886.
- [5] Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 8, no. 4, pp. 1-23, 2018. doi: 10.1002/widm.1253.
- [6] Faizi, F. Loutfi, and M. Bahaj, "Using recommendation systems to analyze opinions and detect trends in social media: A case study on Twitter," *Proc. Int. Conf. Big Data Eng. (BDE)*, 2021, pp. 86-90. doi: 10.1109/BDE.2021.1234567.
- [7] K. Dwivedi, L. Hughes, I. S. Shah, and J. S. Rana, "Analyzing fake news and hate speech detection using machine learning and NLP," *Comput. Human Behav.*, vol. 120, pp. 106750, 2021. doi: 10.1016/j.chb.2021.106750.
- [8] Liu, Z. Lei, L. Zhang, and X. Li, "A sentiment-aware recommendation system for social media using BERT and machine learning," *IEEE Int. Conf. Big Data*, pp. 124-128, 2020. doi: 10.1109/BigData50022.2020.9377918.
- [9] Klymash, M. Kyryk, Y. Pyrih, O. Hordiichuk-Bublivska and T. Andrukhiv, "Model of Large Sparse Datasets Processing Efficiency in IIOT," *2023 17th International Conference on the Experience of Designing and Application of CAD Systems (CADSM)*, Jaroslaw, Poland, 2023, pp. 45-49, doi: 10.1109/CADSM58174.2023.10076508. "Big Data Analysis in Smart Grid Systems", Yu Jun, Olena Hordiichuk-Bublivska, Yan Lingyu, Marian Kyryk, Mykola Beshley, Hu Jiwei, 18th IMEKO TC10 Conference "Measurement for Diagnostics, Optimisation and Control to Support Sustainability and Resilience" Warsaw, Poland, September 26–27, 2022