INTELLIGENCE INFORMATION SYSTEM FOR TRANSFORMER-BASED SENTIMENT ANALYSIS

Mazurets Oleksandr
Ph.D. in Engineering Science, Associate Professor
Kuzmak Kiril
Bachelor Student
Rovinsky Andriy
Information Technology Design Specialist
Kadynska Veronika
Bachelor Student
Computer Science Department
Khmelnytskyi National University, Ukraine

The rapid development of artificial intelligence technologies has significantly changed the ways in which textual data are processed, interpreted, and used in decision-making systems. Natural language processing has become one of the most important areas of modern information technology, as it provides methods for the automated analysis of large volumes of unstructured text. Among the most widely used NLP tasks is sentiment analysis [1], which aims to determine the emotional polarity of a text message and classify it as positive, negative, or neutral depending on the selected model and task formulation.

The problem of gender bias [2] in transformer-based sentiment analysis is especially relevant because modern NLP systems are increasingly used in socially sensitive domains. Sentiment analysis tools may be applied in recruitment analytics, social media monitoring, educational platforms, customer feedback analysis, recommendation systems, and sociological research. If the model assigns different sentiment scores to semantically identical sentences that differ only in gender-related words, the system cannot be considered fully objective [3]. Therefore, the development

of intelligent information systems capable of detecting such differences is an important scientific and practical task.

In the context of the present research, the use of natural language processing methods [4, 5] is more appropriate than computer vision or convolutional neural networks, because the object of analysis is not an image, but a text message [6]. Textual data contain linguistic, semantic, and contextual features that require specialized methods for tokenization [7], contextual representation [8], and semantic interpretation [9]. NLP methods make it possible to analyze the structure of natural language, identify meaningful patterns, process gender-marked expressions, and evaluate how these expressions influence the output of a neural model [10].

Sentiment analysis occupies a special place among NLP tasks because it connects formal text processing with the interpretation of subjective meaning. A sentiment analysis model does not simply identify words, but evaluates the emotional orientation of an entire sentence or message [11]. In transformer-based models, this evaluation depends on the contextual representation of the input text. As a result, even small changes in wording may influence the output of the model. This is particularly important for gender bias detection, because the replacement of a male name with a female name, or a male pronoun with a female pronoun, should not change the sentiment score if the semantic content of the sentence remains the same.

Transformer architectures are effective for this task because they use attention mechanisms to represent the relationships between tokens in a sentence [12]. Unlike earlier neural network architectures, such as recurrent neural networks, transformers process the entire sequence more efficiently and capture long-range dependencies between words [13, 14]. This provides high quality in sentiment classification tasks. At the same time, the same contextual sensitivity may become a source of bias if the model has learned stereotypical associations from training data [15]. For example, the model may assign a more positive or negative sentiment to a sentence depending only on whether it contains a male or female name, although the content of the message remains unchanged [16].

At the same time, the detection of gender bias in sentiment analysis cannot be reduced only to the evaluation of general model accuracy. A model may demonstrate acceptable classification performance on standard test datasets and still produce unstable or unfair outputs for specific demographic markers. Therefore, additional auditing procedures are required to analyze how the model behaves under controlled changes in the input text. In this context, an information system that combines sentiment analysis with counterfactual comparison provides a practical mechanism for identifying hidden deviations in model behavior and supports a more transparent evaluation of transformer-based NLP systems [17].

The aim of the work is to develop an intelligent information system for transformer-based sentiment analysis that enables the detection of gender bias in text messages by comparing sentiment classification results for counterfactual variants of the same sentence.

To achieve this aim, the following tasks are addressed: analysis of the problem of gender bias in NLP systems; selection of a transformer-based model for sentiment

analysis; development of a procedure for generating counterfactual text pairs with male and female gender markers; design of the architecture of the information system; implementation of the software prototype; and experimental evaluation of the proposed approach on a set of text messages.

The proposed intelligent information system is designed to detect gender bias in transformer-based sentiment analysis by comparing the model outputs for semantically equivalent text messages that differ only in gender markers. The general logic of the system is based on counterfactual text transformation. For each input sentence, the system generates two variants: a female-marked version and a male-marked version. These two variants are then processed by the same transformer-based sentiment analysis model, which makes it possible to isolate the influence of gender markers on the final sentiment score.

The architecture of the system includes several interconnected modules. The first module performs input data processing and text validation. The second module generates counterfactual text pairs using a predefined dictionary of gender substitutions, including pronouns, gender-marked nouns, and personal names. The third module performs sentiment classification using a transformer-based model. The fourth module calculates the difference between the sentiment scores obtained for male and female variants of the same sentence. The fifth module aggregates the results and provides a summary interpretation of the detected bias.

The software implementation of the proposed system was carried out using the Python programming language. The sentiment analysis component was implemented on the basis of the DistilBERT transformer model, which is a compact version of BERT and is suitable for text classification tasks due to its lower computational complexity. The model returns a sentiment label and a confidence score for each analyzed sentence. To make the comparison possible, the obtained model output is converted into a unified numerical sentiment score in the range from −1 to 1, where negative values correspond to negative sentiment and positive values correspond to positive sentiment.

The key indicator used in the system is the difference between the sentiment score of the male version of the sentence and the sentiment score of the female version. This difference is interpreted as a quantitative indicator of gender-related deviation in the model output. If the absolute value of this deviation exceeds the selected sensitivity threshold, the sentence is considered biased. In the implemented prototype, the threshold value was set to 0.0001, which makes it possible to filter out insignificant numerical fluctuations and focus on meaningful changes in the model behavior.

The system also determines the direction of the detected bias. If the male-marked version of the sentence receives a higher sentiment score, the result is interpreted as Favoring Man. If the female-marked version receives a higher sentiment score, the result is interpreted as Favoring Woman. If the difference between the two scores does not exceed the threshold, the sentence is treated as Neutral or fair with respect to the tested gender marker. This approach allows the system not only to detect the presence of bias, but also to identify its direction.

For experimental evaluation, the Equity Evaluation Corpus was used as the source of text data. This dataset is appropriate for testing bias in sentiment analysis systems

because it contains text constructions that can be used to compare model behavior with respect to different demographic markers. In the experiment, a sample of 70 sentences was processed. For each sentence, the system generated a male and a female counterfactual version; therefore, the model performed sentiment analysis for 140 text variants in total.

The experimental results showed that the DistilBERT-based sentiment analysis model demonstrated gender-related differences in a considerable part of the analyzed cases. Out of 70 tested sentences, 38 sentences were classified as biased, which corresponds to a bias rate of 54.29%. The remaining 32 sentences did not demonstrate a deviation exceeding the selected threshold and were classified as neutral with respect to the tested gender marker.

The dominant direction of the detected bias was identified as Favoring Woman. This means that, within the analyzed sample, the female-marked variants of the same sentences more often received a higher sentiment score than their male-marked counterparts. Such a result confirms that the transformer-based model may react not only to the semantic content of the message, but also to gender-related lexical substitutions.

An additional observation concerns the type of gender marker used in the sentence. The experiment showed that sentences containing pronouns produced smaller deviations, whereas sentences containing personal names resulted in more noticeable differences in sentiment scores. This indicates that the model's reaction to gender bias may depend not only on the presence of a gender marker, but also on its linguistic form and contextual role in the sentence.

The reliability of the implemented software modules was additionally checked using unit testing. A total of 10 unit tests were conducted successfully. The tests covered the correctness of sentiment score conversion, validation of the output score range from −1 to 1, detection of the bias direction, and processing of empty input values. The successful completion of all tests confirms that the core components of the information system work consistently and can be used for further experimental analysis.

It should also be noted that the obtained results should be interpreted as an experimental evaluation of model behavior rather than as a final conclusion about all transformer-based sentiment analysis systems. The sample of 70 sentences allows demonstrating the functionality of the proposed information system and identifying gender-related deviations in the selected DistilBERT model, but further research requires testing on larger datasets, additional transformer architectures, and different types of gender-marked linguistic constructions. This makes it possible to extend the proposed approach from a prototype-level solution to a more general framework for fairness auditing in NLP-based information systems.

Thus, the proposed intelligent information system provides a structured and quantitatively interpretable approach to detecting gender bias in transformer-based sentiment analysis. Unlike ordinary sentiment analysis systems, the developed system evaluates not only the emotional polarity of a text, but also the stability of model behavior under controlled gender substitutions. This makes it possible to use the system

as a tool for ethical auditing of NLP models and for identifying hidden bias in automated text analysis systems.

References

1. Yeganegi, M. R., Hassani, H., & Komendantova, N. (2025). Identifying and mitigating gender bias in social media sentiment analysis: A post-training approach on example of the 2023 Morocco earthquake. Information, 16(8), 679.
2. Nemani, P., Joel, Y. D., Vijay, P., & Liza, F. F. (2024). Gender bias in transformers: A comprehensive review of detection and mitigation strategies. Natural Language Processing Journal, 6, 100047.
3. Murava, V., Zalutska, O., Didur, V., & Mazurets, O. (2025). Software architecture of information system for exchanging LLM thematic prompts. In Proceedings of the IV International Scientific and Practical Conference "Global Trends in the Development of Information Technology and Science" (pp. 121–127). Stockholm, Sweden.
4. Tymofiiev, I., Mazurets, O., Hardysh, D., & Molchanova, M. (2024). Neural network dual architecture for depression detection using cloud services. In Proceedings of the XLVI International Scientific and Practical Conference "Scientific Research in the Era of Digital Technologies: Challenges and Opportunities" (pp. 84–88). Barcelona, Spain.
5. Molchanova, M., Didur, V., Sobko, O., & Mazurets, O. (2025). Detection of web propaganda patterns by transformer neural networks: Improving efficiency via dataset balancing. CEUR Workshop Proceedings, 3988, 112–126.
6. Hladun, O., Mazurets, O., Molchanova, M., & Sobko, O. (2024). Real time detection the person emotion state using neural network. In Proceedings of the 2nd International Scientific and Practical Conference "Scientific Research: Modern Innovations and Future Perspectives" (pp. 119–123). Montreal, Canada.
7. Blazhuk, V., Mazurets, O., & Zalutska, O. (2024). An approach to using the mBERT deep learning neural network model for identifying emotional components and communication intentions. In Proceedings of the XLIV International Scientific and Practical Conference "The Impact of Scientific Research on the Development of the Modern World" (pp. 79–84). Dubrovnik, Croatia.
8. Mazurets, O., Molchanova, M., Klimenko, V., & Prosvitliuk, M. (2024). Practice implementation of neural network model BART-Large-CNN for text annotation. In Proceedings of the XXVII International Scientific and Practical Conference "Prospects of Scientific Research in the Conditions of the Modern World" (pp. 97–102). Rotterdam, Netherlands.
9. Yurchenko, D., Mazurets, O., Didur, V., & Molchanova, M. (2024). Approach to using cloud services for visual analytics of neural network analysis of texts emotional tonality. In Proceedings of the XLVII International Scientific and Practical Conference "The Future of Scientific Discoveries: New Trends and Technologies" (pp. 108–113). Marseille, France.
10. Mazurets, O., Tymofiiev, I., & Dydo, R. (2024). Approach for using neural network BERT-GPT2 dual transformer architecture for detecting persons depressive

state. In Ricerche scientifiche e metodi della loro realizzazione: esperienza mondiale e realtà domestiche. Proceedings of the VI International Scientific and Practical Conference (pp. 147–151). Bologna, Italy.
11. Mazurets, O., Vit, R., Molchanova, M., Tymofiiev, I., & Sobko, O. (2025). Context-enriched approach to students depression monitoring in education using BERT-GPT hybrid model. CEUR Workshop Proceedings, 4096, 167–176.
12. Molchanova, M., Didur, V., Sobko, O., & Mazurets, O. (2025). Detection of web propaganda patterns by transformer neural networks: Improving efficiency via dataset balancing. CEUR Workshop Proceedings, 3988, 112–126.
13. Mazurets, O., Sobko, O., Vit, R., & Pasternak, V. (2024). Practical approach for detection by deep learning of target objects of subject area based on semantic connectivity indicators in audio database. In Proceedings of the XXIV International Scientific and Practical Conference "Modern Scientific Challenges Are the Driving Force of the Development of Scientific Research" (pp. 91–96). Bruges, Belgium.
14. Mazurets, O. V., Sobko, O. V., Molchanova, M. O., Zalutska, O. O., & Yurchak, A. V. (2024). Practical implementation of neural network method for stress features detection by social internet networks posts. In Proceedings of the II International Scientific and Theoretical Conference "Scientific Review of the Actual Events, Achievements and Problems" (pp. 160–167). Berlin, Germany.
15. Sobko, O., Mazurets, O., Didur, V., & Chervonchuk, I. (2024). Recurrent neural network model architecture for detecting a tendency to atypical behavior of individuals by text posts. In Proceedings of the XXVI International Scientific and Practical Conference "Theoretical and Practical Aspects of Modern Research" (pp. 113–117). Ottawa, Canada.
16. Shevchuk, P., Molchanova, M., & Mazurets, O. (2024). Software for text messages reliability analysis based on the machine learning models ensemble. In Proceedings of the IV International Scientific and Practical Conference "Innovative Research and Perspectives of the Development of Science and Technology" (pp. 347–354). Stockholm, Sweden.
17. Mazurets, O., Vit, R., Molchanova, M., Sobko, O., Wierzbicki, A., & Chumachenko, D. (2025). Neural network detection of digital fatigue and burnout with interpretable thematic segmentation. CEUR Workshop Proceedings, 4141, 28–37.