Journal of Electrical Technology UMY (JET-UMY), Vol. 9, No. 1, June 2025
ISSN 2550-1186 e-ISSN 2580-6823

Stance Detection of Controversial Articles Using TF-IDF and BERT

Eka Parima Saragih*, Anggraini Dyah Ayu Sekarlangit, Faqih Al Suman

Master Program of Science in Information Technology (MSIT), Faculty of Computer Science, President University, Jababeka Education Park, Cikarang, Bekasi 17550, Indonesia
*Corresponding author, e-mail: eka.saragih@student.president.ac.id

Abstract – *Online misinformation and polarized discussions require better methods for automatically detecting a text's stance. As digital content increases, identifying whether a news article supports, opposes, or is neutral towards its headline is crucial for fighting the spread of false information. This study presents a hybrid model designed for this task. We combine lexical features from Term Frequency-Inverse Document Frequency (TF-IDF), which captures word-level patterns, with contextual semantic information from a pretrained BERT model (bert-base-uncased). The features from both TF-IDF and BERT's [CLS] token were concatenated and used to train a logistic regression classifier. The model was trained and tested on a filtered version of the Fake News Challenge (FNC-1) dataset, with "unrelated" pairs removed to focus on more nuanced stance classification. The final evaluation of this model achieved 83% accuracy with a macro F1-score of 0.68. This model evaluates best in the Neutral stance (F1-score 0.91), but has some difficulty detecting the stance in the Oppositional class (with an F1-score 0.39). The results of this evaluation show that surface level lexical features combined with deep contextual understanding can improve the performance of stance detection.*

Keywords: *BERT, Fake News Challenge, Hybrid Model, Stance Detection, TF-IDF*

I. Introduction

The huge proliferation of online content has made Information Retrieval (IR) systems more important for helping people find the information they need. However, the internet today also brings major challenges. Trusted news outlets are struggling, and online spaces tend to group people with similar views, which can lead to stronger polarization [1]. Search engines generally work well for simple, factual questions but often fail when people are searching for balanced opinions on controversial topics. In some cases, these systems even worsen bias by ranking certain viewpoints higher [7]. These problems show how human behavior plays a big role in spreading information. For example, false news often spreads quickly because people like to share surprising stories, not just because of bots [2]. A person's mood, such as being happy or frustrated, can also affect how they search for information and interact online [3]. Detecting fake news or misleading opinions online is still hard. It requires not only understanding the text but also the social context [6] and the back-and-forth conversations that often happen on social media [8].

In recent years, stance detection has advanced with the help of transformer-based models like BERT, which use contextual embeddings to classify stances in different types of text, from long news articles to short social media conversations [8, 9, 11]. Stance detection plays a key role in larger tasks like checking information credibility, detecting rumors, and identifying fake news [6, 19]. Some research has tried to improve stance detection by exploring the link between stance and sentiment. Sentiment features have been used in studies about polarized political issues like the Catalan independence referendum and climate change [13]. Other studies show that hybrid approaches, which combine text-based and numerical features—such as TF-IDF with ensemble classifiers—can significantly improve detection accuracy compared

Manuscript received May 2025, revised June 2025

to basic models [11].

However, these models still face a number of limitations. The complexity of political language and the limited availability of labeled data on a large scale are still major obstacles [11]. The rapid development of the internet also makes it easier to spread misinformation, so the urgency to verify the veracity of online content is increasing [9]. Furthermore, fake news often imitates the language style of real news so that it is difficult to distinguish, and existing models often suffer from training data bias so that they tend to overfit certain language patterns and fail to classify new examples accurately [11], [18]. Another challenge is to differentiate search engine results with recommendation algorithms to avoid filter bubbles and maintain information fairness [5].

In other words, stance detection in controversial online articles still faces major challenges, especially in capturing nuance, bias and diversity of viewpoints. Most existing models tend to rely only on simple lexical approaches or single contextual representations, so they are often less capable of detecting more subtle and complex meanings in text. This deficiency emphasizes the existence of a research gap in developing methods that are able to combine the power of surface analysis of words with deeper semantic understanding in a balanced, practical and efficient manner.

A number of recent studies have shown that both simple lexical approaches and transformer-based models can improve stance detection performance. For example, Karande et al. [12] used BERT embedding for information credibility analysis on social media and achieved high accuracy, while Bourgonje et al. [24] as well as Ghosh et al. [25] shows how stance detection in online news effectively supports the detection of clickbait and issues on social media. More sophisticated transformer models, such as RoBERTa and AraBERT, have also proven superior in certain languages and domains [26]-[28].

On the other hand, some studies try to combine lexical features with deep models. Kausar et al. [29] combined TF-IDF and deep learning models for fake news detection and showed significant performance improvements, while Essa et al. [30] proposed a combination of BERT with LightGBM, and Aljrees et al. [31] developed a special CNN-LSTM architecture for stance detection. Other research has even explored parallel BERT architectures [32], graph-based approaches [33], to hybrid TF-IDF methods with deep networks [34]. These results consistently show an increase in accuracy, especially in the neutral and supportive classes, but are often still weak in detecting oppositional stances.

Despite showing promising results, previous studies still face a number of important limitations. Transformer models tend to require large computational resources, making them difficult to implement widely in real Information Retrieval systems [11], [27]. Meanwhile, simple lexical approaches such as TF-IDF are more efficient, but are unable to capture complex semantic nuances, biases, and diversity of viewpoints in political language and controversial discourse [24], [25]. Some hybrid studies do try to combine word surface approaches with contextual models, but they still tend to be architecturally heavy or not explicitly enough to balance the two [29], [30].

In addition, many existing models still have difficulty generalizing across topics and domains due to dataset bias, and often the sentiment-based features used do not represent actual attitudes [11]. Thus, there is a clear research gap: there is still limited research that is able to integrate simple lexical methods such as TF-IDF with deep contextual models such as BERT in a practical, balanced and efficient manner to improve stance detection in controversial articles. This research attempts to close this gap by proposing a hybrid framework that not only balances the simplicity of TF-IDF and the semantic power of BERT, but also maintains interpretability and computational efficiency, making it more relevant for real applications in IR systems dealing with issues of bias, linguistic nuances, and diversity of perspectives.

Based on the research gap, the proposed problem statement is how to improve the performance of stance detection for controversial articles by combining lexical and contextual information, in order to overcome the limitations of current models in detecting bias, nuance and diversity of views in information retrieval systems.

This research aims to improve the accuracy of stance detection for controversial online articles by developing a hybrid model that combines both textual and numerical features. The model will integrate simple lexical approaches like TF-IDF with more advanced contextual models, such as pre-trained language models, to capture both surface-level and deep semantic features.

As a contribution, this research offers a new approach in the form of a hybrid model that combines traditional lexical methods with advanced contextual representation. Specifically, this research

combines two feature extraction techniques: TF-IDF and a pre-trained language model BERT. TF-IDF is effective for identifying important word patterns based on frequency and relevance, while BERT provides rich contextual representation [15], able to capture the meaning and nuances of sentences in depth. By integrating the two, it is hoped that the resulting model can utilize the advantages of each approach, so that stance detection can be carried out more accurately, efficiently, and has good scalability on various large-scale datasets.

This research focuses only on improving stance detection in textual data, especially for controversial news articles and social media content. This research does not discuss stance detection in other media such as images or videos. In addition, this research relies more on large pre-existing datasets, which may result in biased results. The use of complex models will be a limitation in this research, especially in the area of computational costs, which means large-scale implementation and extensive customization is beyond the scope of this work.

II. Method

II.1. *Research Design*

This study applies a hybrid feature-based approach (lexical and semantic) for stance detection. The main aim is to classify the stance expressed in the body of the article against the headline of the related article. To this end, this research utilizes two complementary feature extraction strategies, namely TF-IDF, which captures surface-level lexical features, and BERT, which extracts deep contextual embeddings. By combining these methods, it is hoped that the model can better understand the importance of frequency and nuances of meaning and semantic context contained in each text or sentence.

This hybrid approach is used to adapt to the limitations of stance detection which can usually focus more on shallow lexical features or perhaps only with complex and complex contextual models, especially in detecting less pronounced stance differences on controversial topics. In addition, this combination also gives the model the opportunity to take advantage of both simplicity and depth, which offers a more balanced and interpretable stance detection framework.

As depicted in the flowchart in *Figure 1*, this system follows the framework of supervised machine learning. Initially, the dataset will be organized and then pre-processed to make it clean and good for the training and evaluation process.

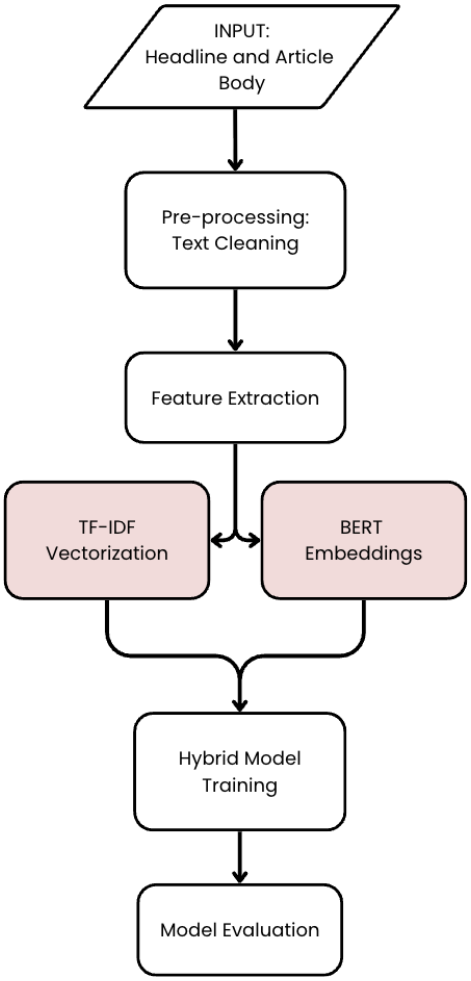

Fig. 1. Method Overview

Then the processed data will be continued with feature extraction with TF-IDF and BERT, then the vector will be given to the logistic regression model, which was chosen as the baseline classifier because it is simple and has interpretability. This choice allows research to clearly assess the effectiveness of feature fusion without introducing the complexity of deep neural networks for classification, thereby focusing the analysis on feature quality rather than model architecture. In the future, this framework can be extended by experimenting with more complex classifiers once the contribution of each feature type has been properly evaluated.

Then for the final results, this model will be evaluated by dividing the data for train and test, and ensuring that the stance categories have been divided appropriately and proportionally. The final

evaluation method is to look at accuracy using metrics such as precision, recall, and F1-score. This metrics report will then provide an explanation of the understanding of the model's performance for each class. The framework of this research allows us to focus on how combined TF-IDF and BERT embeddings can improve stance classification. The aim is additionally to verify whether these combined features are sensitive enough to differentiate supportive, oppositional, and neutral nuances of language when analyzing controversial topics.

In this research, TF-IDF (Term Frequency–Inverse Document Frequency) and BERT (Bidirectional Encoder Representations from Transformers) are not used as classification models, but rather as the main tools for feature extraction which are then combined. TF-IDF functions to capture lexical features at a surface level, assigning numerical weights to words based on their frequency in documents and across the corpus. Formally, the TF-IDF weight for a word t in the document d to the corpus D is defined as:

$$w_{t,d} = TF(t,d) \times IDF(t,D) \quad (1)$$

$$TF(t,d) = \frac{f_{t,d}}{\sum t' \in d\, f_{t',d}} \quad (2)$$

$$IDF(t,d) = log\left(\frac{N}{1+|\{d \in D : t \in d\}|}\right) \quad (3)$$

where $f_{t,d}$ is the frequency of word t in document D, N is the total number of documents, and the denominator $|\{d \in D : t \in d\}|$ is the number of documents containing word t. This formula is effective in identifying important keywords and lexical patterns that are often indicators of attitudes. On the other hand, BERT is used to extract deep contextual embeddings. In contrast to TF-IDF which only looks at word frequency, BERT understands the meaning of words based on the context of the sentence by utilizing a self-attention mechanism. which allows each token in the text to see other tokens to understand the global context. From the BERT output, the embedding extracted is a representation of the [CLS] token (h_{CLS}), which is used as a vector summarizing the overall meaning of the combined text (title and body of the article):

$$h_{CLS} = f_{BERT}(X) \quad (4)$$

The main task of these two methods is to create a richer and more comprehensive data representation. The sparse feature vector from TF-IDF (X_{TF-IDF}) and the dense embedding from BERT (h_{CLS}) are then combined horizontally to form one hybrid feature matrix:

$$X_{hybrid} = [X_{TF-IDF} \,||\, h_{CLS}] \quad (5)$$

This hybrid approach exploits the synergy between the two methods: TF-IDF anchors the model on important keywords, while BERT provides contextual understanding that allows the model to generalize to semantically rich claims. This combination aims to address the weaknesses of each method individually, ultimately resulting in a more balanced and accurate performance for the attitude detection task.

II.2. *Data Collection*

This research discusses stance detection in controversial articles, which means the model needs to capture differences in opinions and positions in complex and often conflicting texts. For this reason, this model will also be trained on the Fake News Challenge Stage 1 (FNC-1) dataset, which will be the primary data source for evaluating this model. The selection of this dataset is based on its connection to the domain discussed in this topic, namely fake news detection, which has a relationship with controversial content.

The features in the FNC-1 dataset display several types of stances such as agree, disagree, discuss, and unrelated which are very suitable for characterizing positions in controversial discourse. Moreover, fake news inherently involves controversial claims and polarized viewpoints, the dataset reflects the dynamics of controversial topics, making it a suitable tool for detecting stances in this broader domain.

By utilizing the FNC-1 dataset, this research focuses more on stance classification, not just filtering relevant content from irrelevant content. This opens up research to utilize well-established and well-annotated datasets that reflect the challenges and complexity of controversial articles, thereby enabling more rigorous and generalizable evaluation of TF-IDF and transformer-based hybrid models.

II.3. *Data Pre-processing*

Before inputting text into the TF-IDF and BERT

models, the dataset will first be cleaned to increase consistency and reduce noise. Both the article title and its contents will be processed independently and this pre-processing will involve converting all text to lowercase and removing URLs in the text, extra punctuation, and non-alphanumeric characters. The goal is to minimize noise and help statistical and neural models better generalize from the input data.

Once the title and content of the article are cleaned, they will be combined to form an integrated representation of the stance context. It is this combination that will capture the interaction between the title (which often contains opinion or summary language) and the body of the article (which provides factual or expanded content), thus providing a complete picture model for classification.

To improve the classification task, stances with the label “unrelated” will be removed from the data set to make the model more focused on the finer differences between stances related to the body and the title. The remaining stances are neutral, supportive (agree), opposition (disagree). This also prevents model performance from being affected by the large number of "unrelated" samples which are generally easier to identify. Finally, the three remaining stance labels will be converted into numerical format to prepare them for the next machine learning model.

II.4. *Data Processing*

The text from the dataset that has been previously pre-processed is converted into two different sets of numerical features, namely TF-IDF vectors and BERT embeddings.

First, it will start with TF-IDF applied to the combined headline and body text of the article. The technique then generates a sparse feature vector for each article, which assigns a numerical weight (vector) to each word based on its frequency and importance. To prevent the model from overfitting and to ensure the model remains focused on the most significant terms, the size of the vocabulary is limited to the top 5,000 features.

Afterwards, contextual embeddings are generated using a pre-trained base-uncased bert model, the aim of which is to capture deeper semantic information. Each text sample is tokenized and filled with a maximum of 128 tokens. The embeddings associated with [CLS] tokens are then extracted from the final layer of the model, serving as a consolidated numerical representation of the entire set. Then extraction is carried out per batch for computational efficiency, especially when using a GPU.

In the last step even the two features are finally combined, the sparse TF-IDF vector and the dense BERT embedding are combined horizontally to create a single hybrid feature matrix. This approach provides a surface-level lexical pattern classification model from TF-IDF and deep contextual understanding from BERT, whose goal is to improve predictive performance.

II.5. *Data Analysis*

After creating the feature vector, the dataset will then be divided into 80% training set and 20% test set. This division is carried out using stratified sampling, which ensures that the proportion of each stance (supportive, oppositional, neutral) is the same for train and test, so it will produce a fair final evaluation.

For classification, this research uses Logistic Regression classifier. The classifier will be configured with a high iteration limit (max_iter=1000) to handle the large number of features in the hybrid dataset. The model is trained with data from X_train and y_train and then used to produce predictions on X_test data.

At the end, to screen the model performance, a classification_report from scikit-learn is created. This report will show the results of precision, recall, F1-score for each of the three move classes, which will show how effectively the model identifies each class. The analysis focuses on how a hybrid approach that combines these different types of features can improve the results of the model. The main idea is that TF-IDF helps capture important word patterns from the given or input data, while the lever of BERT is to provide additional information about the context that cannot be covered by TF-IDF. Finding the right balance between these features is critical to model performance.

II.6. *Data Visualization*

To gain deeper insights into the stance classification data and model behavior, several visualizations were employed.

Figure 2 shows a visualization of t-SNE (t-distributed Stochastic Neighbor Embedding) applied to 1,000 examples of BERT-embedded sentences to show how they cluster based on meaning. This method reduces the high-dimensional

feature space to two dimensions to make it easier to visualize.

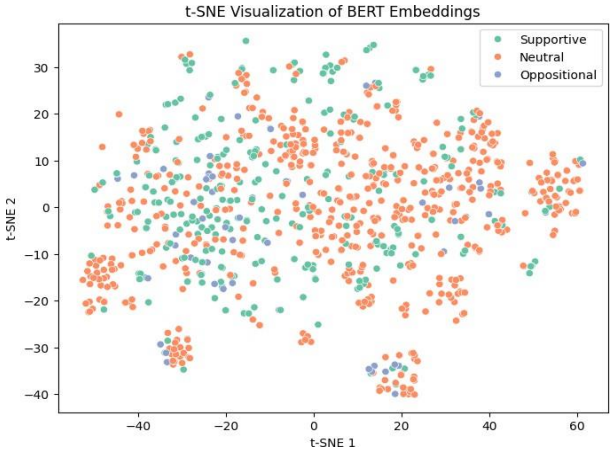

Fig. 2. Semantic Clustering of BERT Embeddings Visualized by t-SNE

The scatter plot shows how the categories of stance or class (supportive, neutral, oppositional) are distributed. Some small clusters can be seen, especially in the neutral class, but there is clearly visible overlap between the groups. This overlap reflects the semantic ambiguity and similarities between classes that often occur in social media conversations.

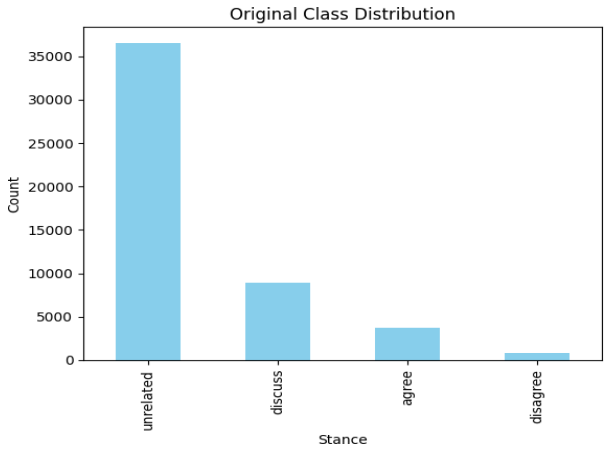

Fig. 3. Original Class Distribution

The original class distribution of the dataset is depicted in *Figure 3*, which clearly illustrates the dominance of the "unrelated" category, followed by a much smaller volume of samples labeled as "discuss", "agree", and "disagree". This significant imbalance can lead to biased model training.

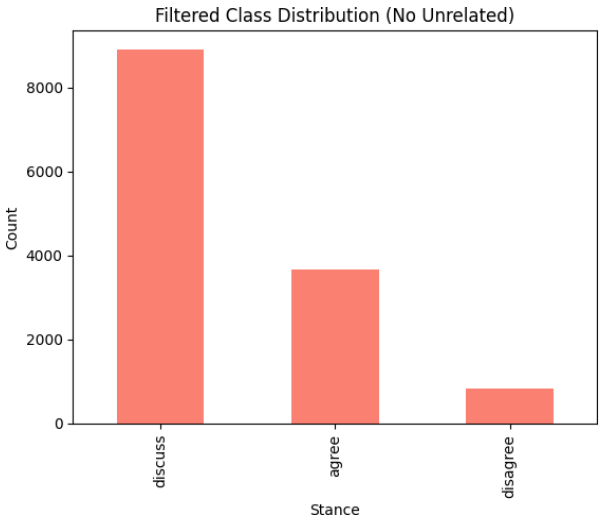

Fig. 4. Filtered Class Distribution (No Unrelated)

To address this, unrelated instances were filtered out, and the resulting distribution is visualized in *Figure 4*. The updated class distribution still shows some inequality where the 'discussion' class remains the most common, followed by 'agree' and 'disagree' which have the smallest samples. This imbalance is also taken into account during model training and evaluation to ensure fair representation of minority classes.

II.7. *Data Validity*

This research uses datasets from FNC-1 which is a benchmark that has been widely used for stance detection. The use of this dataset helps maintain consistency of the study with previous research and provides a solid basis for experimentation. To focus the task on more refined attitude classification, the dataset was filtered to remove 'unrelated' samples to narrow the task to differentiate between the 'agree', 'disagree' and 'discussion' classes, which are semantically more similar and more difficult to separate. The filtered class distribution has been visualized in *Figure 4*, the point is to check the pre-processing steps and see their effect on the class balance.

Data consistency is maintained by applying the same text pre-processing steps, then combining stance data by the body article and its headline with the same ID, and carefully converting text labels into numeric values. This is to reduce interference from irrelevant factors and ensure that input data is clean and well prepared. The dataset will also be divided into two as explained in the Data Analysis

section, which is important for a fair evaluation, especially considering the class imbalance.

To improve the generalization ability of the model, this research exploits the semantic power of pre-trained BERT embeddings. Even without domain-specific adjustments, these embeddings provide a competent understanding of the data context. This feature engineering strategy is a hybrid that combines deep semantic insights from BERT with lexical frequency data from TF-IDF. By combining these features, the results can achieve a richer data representation that includes the content and contextual meaning of the text. Although performance can be improved with domain-specific adjustments, current methods can be prioritized due to their high reproducibility and resistance to overfitting. That way, this can become a reliable basis for further research.

III. Results

By combining shallow lexical features from TF-IDF with deep semantic embeddings from a pre-trained BERT model, this hybrid approach to stance detection achieves quite strong performance on the Fake News Challenge (FNC-1) dataset. By excluding the 'unrelated' class, the model achieved an overall accuracy of 83.36% and a macro average F1-score of 0.68, which proves its capacity to capture the subtle stance between the title and body of the related article. The robustness of this model can be seen from the high F1-score for neutral (0.91) and supportive (0.74), which shows that the model is successful in utilizing surface level information and deep context.

The methodology starts with pre-processing and merging each pair of title and article body. Feature extraction then proceeds in two parts, namely TF-IDF which identifies the top 5,000 lexical features based on term relevance while the semantic vector is obtained from the [CLS] BERT token. These two vectors are then combined horizontally to form a single input for the Logistic Regression classifier. This classifier was trained on 2,886 filtered samples which were then categorized into stance labels namely Oppositional, Neutral, and Supportive (mapped from disagree → 0, discuss → 1, agree → 2).

Baseline Model Comparison

To isolate each feature set, we compared our hybrid model with two simpler versions. one only uses TF-IDF and the other only BERT.

A notable finding was the TF-IDF model's high accuracy of 82.13%, underscoring the value of simple keyword relevance. This impressive number, however, doesn't capture a crucial failure in which the model cannot interpret the underlying meaning of the text. For stances where context is more important than keywords (such as those in the form of paraphrased sarcasm or disagreement), TF-IDF still fails. This limitation is reflected in the lower F1 macro score of 0.64.

This BERT only model is intended to solve this semantic problem but the overall score is lower (accuracy 79.11% and macro F1-score 0.62). Despite understanding the context, they appear to struggle with class separation without the explicit term frequency signals that TF-IDF provides. This is especially evident in the difficulty of identifying the 'Opposition' class.

The success of this hybrid model lies in its ability to overcome the limitations of each component. By integrating appropriate TF-IDF stance with contextual understanding from BERT, the model achieves a superior macro F1-score of 0.68. More importantly, the model shows more consistent and balanced performance across all categories ultimately outperforming the TF-IDF-only and BERT-only models.

TABLE I
BASELINE MODEL COMPARISON

Model	Accuracy	Macro F-1 Score
TF-IDF Only	82.13%	0.64
BERT Only	79.11%	0.62
Hybrid (TF-DF + BERT)	83.36%	0.68

These results validate the initial hypothesis that the combination of shallow and deep linguistic features is complementary and highly effective. The success of the hybrid model lies in this synergy: BERT's contextual awareness allows it to generalize across semantically rich claims, while TF-IDF's lexical features anchor the model on important keywords. This dual approach overcomes critical failures of simpler models, for example TF-IDF's deceptively high accuracy because it cannot interpret the true meaning which is a requirement for robust and real applications.

Performance metrics in Table I provide a detailed picture showing strong predictive power with an overall accuracy of 83%, superior in the neutral and supportive categories but the main limitation is in detecting opposition stance. The low recall of 0.33 for this class also highlights a consistent tendency to

classify genuine disagreements as neutral discussions.

This model was created by first combining pairs of titles and contents and creating a combined feature vector by stacking 5,000 TF-IDF features horizontally with CLS token embeddings from BERT which were then used to train it on a Logistic Regression classifier.

TABLE II
MODEL PERFORMANCE METRICS

Stance Class	Precision	Recall	F1-Score	Support
Oppositional	0.47	0.33	0.39	168
Neutral	0.89	0.33	0.91	1782
Supportive	0.74	0.73	0.74	736
Accuracy			0.83	2686
Macro Avg	0.70	0.66	0.68	2686
Weighted Avg	0.83	0.83	0.83	2686

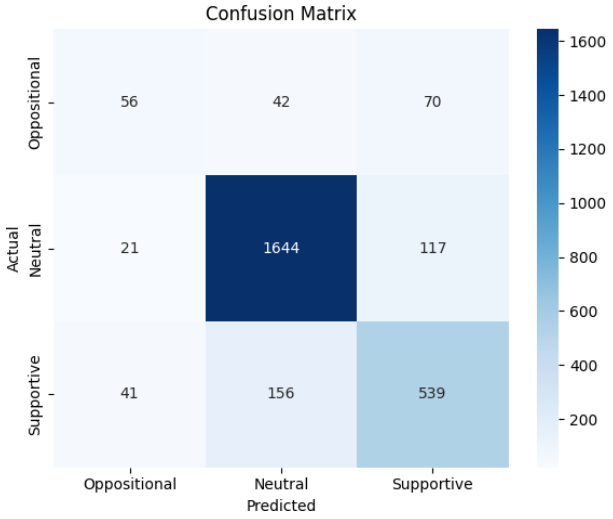

Fig. 5. Confusion Matrix Heatmap

The metrics of this model have shown a clear pattern, although it is very effective in the neutral and support stance, this model seems more difficult to apply in the opposition stance and one of the reasons is the lack of data in the disagree or opposition class. This is clearly visible from the low recall in the opposition group (0.33), which shows that these examples are often misinterpreted as neutral or supportive statements. And as has been said, this has to do with data imbalance and differences of opinion. However, even with these challenges, the macro's overall F1-score of 0.68 indicates a fairly balanced classifier.

So to better understand this behavior, the confusion matrix is checked (Fig. 5), which shows that the model is strongest in the neutral class (1,644 true positives) but still often confuses oppositional and neutral stances. This finding that it may be due to word similarity in subjective phrases confirms the trend seen in recall scores. This semantic overlap is also visible in the t-SNE visualization of BERT embeddings (Fig. 2). While neutral and supporting groups form close-knit groups, opposition groups are more dispersed and better mixed with other groups.

Analysis of terms or keywords in TF-IDF also increasingly supports this finding. Key words for supporting classes are for example "confirmation", "support", clear and expected. On the other hand, even though the opposition class uses stronger language, the lexical signal is less clear because the sample is smaller. This imbalance of the initial dataset is the reason for filtering the dataset as shown in the distribution graph in *Figure 3* and *Figure 4*. This figure illustrates the need to remove dominant "unrelated" classes before training.

Fig. 6. Model Performance Metrics by Stance Class

Figure 6 as shown details the model performance in each stance, and the results provide a clear picture. As previously explained, this model is superior in neutral and supportive stances, and has some difficulty detecting opposition stances. It performs best in the neutral class with an F1-score of 0.91, and shows solid and balanced results for the support class with an F1-score of 0.74, possibly because this class has more training data and clearer linguistic patterns.

The main challenge is the opposition class whose performance has dropped to an F1-score of 0.39. This very low recall of 0.33 confirms that the model

frequently fails to identify disagreements, often mislabeling them as neutral stances. This struggle can be traced to two problems: the class is highly underrepresented in the dataset with only 168 samples, and is often expressed subtly through sarcasm or indirect language that is inherently difficult for a model to capture.

The effectiveness of combining TF-IDF's keyword focus with BERT's contextual understanding is confirmed by the overall strong performance of the model. Because even with its limitations, it has achieved an accuracy of 83% and a macro F1-score of 0.68. The particular challenges of detecting conflict highlight clear paths for future improvement, such as implementing contrastive learning techniques or balancing classes. However, current models are proving to be valuable tools for real-world tasks such as analyzing information that does not match the claims of the headlines.

IV. Conclusion and Future Work

This study successfully demonstrated the effectiveness of a hybrid stance detection model that combines shallow lexical features from TF-IDF with deep semantic embeddings from a pre-trained BERT model. When tested on the Fake News Challenge (FNC-1) dataset, after excluding the dominant 'unrelated' class, our model achieved a strong overall accuracy of 83.36% and a macro-averaged F1-score of 0.68. The model's strong performance supports the core idea that integrating keyword detection with contextual understanding is a superior strategy. This hybrid approach demonstrates greater robustness than utilizing either method independently.

This model shows excellent performance in Neutral and Supportive stances, with respective F1-scores of 0.91 and 0.74. This success confirms the power of the hybrid approach, which can utilize both TF-IDF for keyword recognition and BERT for contextual understanding. The model's ability to perform well on these different categories is what strongly indicates that it has learned a balanced feature representation to identify agreement and neutral discussion.

Although this model shows good performance overall, there are still major limitations seen in its handling of oppositional attitudes. The F1-score for the opposition or disagree category is only 0.39, even the low recall is only 0.33. In practice this means the model often ignores genuine disputes and incorrectly classifies them as neutral discussions. This tendency has also been confirmed by the confusion matrix in *Figure 5*, which shows the overlap between the two classes which is further supported by the t-SNE visualization in *Figure 2*, where the opposition group appears more spread out and less clear than the others.

The model's difficulties in fighting the opposition class can be traced back to two fundamental issues identified in the analysis carried out in this research. These unsatisfactory results stem from two problems. The first is the sharp data imbalance (only 168 Opposition samples) making the model lack learning opportunities, especially for those with an opposition stance. Additionally, the complexity of meaning in conflict sentences, which often rely on indirect language such as sarcasm instead of clear keywords, makes it a significant challenge.

These limitations directly inform and open up future research directions. Data imbalance can be mitigated through methods such as data augmentation or class weighting. To capture complex linguistic cues, more sophisticated architectures, including fine-tuned transformers or contrastive learning, should be explored. And ultimately, the true value of this model and its ability to generalize remains to be verified by testing it against a wider range of social media and news or online content data thereby strengthening the potential use of this model in fighting misinformation especially in controversial political areas.

Acknowledgements

The author would like to thank President University for the academic guidance that helped develop this research. The author would also like to express special thanks to the Faculty of Computer Science for the support provided during the model development and evaluation process.

Apart from that, the author also wants to appreciate the Fake News Challenge Stage 1 (FNC-1) dataset, which is a fundamental benchmark for this stance detection task.

References

[1] D. M. J. Lazer et al., "The science of fake news," Science, vol. 359, no. 6380, pp. 1094–1096, 2018, doi: 10.1126/science.aao2998.

[2] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," Science, vol. 359, no. 6380, pp. 1146–1151,

2018, doi: 10.1126/science.aap9559.

[3] M. Yari Zanganeh and N. Hariri, "The role of emotional aspects in the information retrieval from the web," Online Information Review, vol. 42, no. 4, pp. 520–534, 2018, doi: 10.1108/OIR-04-2016-0121.

[4] N. K. Negied et al., "Academic assistance chatbot—a comprehensive NLP and deep learning-based approaches," Indonesian Journal of Electrical Engineering and Computer Science, vol. 33, no. 2, pp. 1042–1056, 2024, doi: 10.11591/ijeecs.v33.i2.pp1042-1056.

[5] H. Wu et al., "Result diversification in search and recommendation: A survey," IEEE Transactions on Knowledge and Data Engineering, vol. 36, no. 10, pp. 5354–5373, 2024, doi: 10.1109/TKDE.2024.3382262.

[6] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22–36, 2017, doi: 10.1145/3137597.3137600.

[7] Y. Ajjour, "Addressing controversial topics in search engines," Ph.D. dissertation, Bauhaus-Universität Weimar, 2023, doi: 10.25643/BAUHAUS-UNIVERSITAET.6403.

[8] P. Khandelwal, P. Singh, R. Kaur, and R. Chakraborty, "Stance detection in Twitter conversations using reply support classification," in Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods, 2025, pp. 235–242, doi: 10.5220/0013129800003905.

[9] L. Mascarell et al., "Stance detection in German news articles," ETH Zürich, 2021, doi: 10.3929/ETHZ-B-000523833.

[10] Y. Zhang et al., "Stance-level sarcasm detection with BERT and stance-centered graph attention networks," ACM Transactions on Internet Technology, vol. 23, no. 2, pp. 1–21, 2023, doi: 10.1145/3533430.

[11] S. Ng et al., "Stance classification: A comparative study and use case on Australian parliamentary debates," Journal of Computational Social Science, vol. 8, no. 2, p. 43, 2025, doi: 10.1007/s42001-025-00366-y.

[12] H. Karande et al., "Stance detection with BERT embeddings for credibility analysis of information on social media," PeerJ Computer Science, vol. 7, p. e467, 2021, doi: 10.7717/peerj-cs.467.

[13] Z. Elena, "Automatic stance detection on political discourse in Twitter," M.S. thesis, University of the Basque Country.

[14] J. Mina, "Evaluation of text transformers for classifying sentiment of reviews by using TF-IDF, BERT (word embedding), SBERT (sentence embedding) with support vector machine evaluation," M.S. dissertation, Technological University Dublin, 2023.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint, 2018, doi: 10.48550/ARXIV.1810.04805.

[16] J. Ramos, "Using TF-IDF to determine word relevance in document queries," in First Instructional Conference on Machine Learning, Rutgers University, 2003.

[17] S. Tadesse Guda, "Political stance detection on Amharic text using machine learning," M.S. thesis, St. Mary's University

[18] V. V. V. R. Gurram, "Automated detection of fake news in natural language processing: A comparative study of TF-IDF and lexical-based stance detection with logistic regression," B.S. thesis, Blekinge Institute of Technology.

[19] B. Schiller, J. Daxenberger, and I. Gurevych, "Stance detection benchmark: How robust is your stance detection?" KI - Künstliche Intelligenz, vol. 35, no. 3–4, pp. 329–341, 2021, doi: 10.1007/s13218-021-00714-w.

[20] I. Alsmadi, I. Alazzam, M. Al-Ramahi, and M. Zarour, "Stance detection in the context of fake news—A new approach," Future Internet, vol. 16, no. 10, p. 364, 2024, doi: 10.3390/fi16100364.

[21] B. Zhang et al., "A survey of stance detection on social media: New directions and perspectives," arXiv preprint arXiv:2409.15690, 2024, doi: 10.48550/arXiv.2409.15690.

[22] M. I. Alfarizi, L. Syafaah, and M. Lestandy, "Emotional text classification using TF-IDF (term frequency-inverse document frequency) and LSTM (long short-term memory)," JUITA: Jurnal Informatika, vol. 10, no. 2, p. 225, 2022, doi: 10.30595/juita.v10i2.13262.

[23] S. Pathiyan Cherumanal, D. Spina, F. Scholer, and W. B. Croft, "Evaluating fairness in argument retrieval," in Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 3363–3367, doi: 10.1145/3459637.3482099.

[24] P. Bourgonje, J. M. Schneider, and G. Rehm, "From clickbait to fake news detection: An approach based on detecting the stance of headlines to articles," in Natural Language Processing Meets Journalism, 2017, pp. 84–89.

[25] S. Ghosh, P. Singhania, S. Singh, K. Rudra, and S. Ghosh, "Stance detection in web and social media: A comparative study," in Lecture Notes in Computer Science, vol. 11696, 2019, pp. 75–87, doi: 10.1007/978-3-030-28577-7_4.

[26] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," arXiv preprint, 2021, doi: 10.48550/arXiv.2003.00104.

[27] V. Slovikovskaya, "Transfer learning from transformers to fake news challenge stance detection (FNC-1) task," arXiv preprint arXiv:1910.14353, 2019, doi: 10.48550/arXiv.1910.14353.

[28] C. Dulhanty, J. L. Deglint, I. B. Daya, and A. Wong, "Taking a stance on fake news: Towards automatic disinformation assessment via deep bidirectional transformer language models for stance detection," arXiv preprint arXiv:1911.11951, 2019, doi: 10.48550/arXiv.1911.11951.

[29] N. Kausar, A. AliKhan, and M. Sattar, "Towards better representation learning using hybrid deep learning model for fake news detection," Social Network Analysis and Mining, vol. 12, no. 1, p. 165, 2022, doi: 10.1007/s13278-022-00986-6.

[30] E. Essa, K. Omar, and A. Alqahtani, "Fake news detection based on a hybrid BERT and LightGBM models," Complex & Intelligent Systems, vol. 9, no. 6, pp. 6581–6592, 2023, doi: 10.1007/s40747-023-01098-0.

[31] T. Aljrees et al., "Fake news stance detection using

selective features and FakeNET," PLOS ONE, vol. 18, no. 7, p. e0287298, 2023, doi: 10.1371/journal.pone.0287298.

[32] M. Farokhian, V. Rafe, and H. Veisi, "Fake news detection using parallel BERT deep neural networks," Multimedia Tools and Applications, vol. 83, no. 15, pp. 43831–43848, 2023, doi: 10.1007/s11042-023-17115-w.

[33] S. Gong et al., "Fake news detection through graph-based neural networks: A survey," 2023, doi: 10.21203/rs.3.rs-3252100/v1.

[34] A. K. Yadav et al., "Fake news detection using hybrid deep learning method," TechRxiv, 2022, doi: 10.36227/techrxiv.19689844.

Authors' information

Eka Parima Saragih was born in Medan, Indonesia, on July 2, 2003. She received her Bachelor's degree in Information System from President University, Indonesia, in 2024, and is currently pursuing her Master's degree in Information Technology at President University, Indonesia. Her major field of study is natural language processing and machine learning.
She has previously worked on sentiment analysis and information retrieval. Her current research interests include stance detection, transformers, cross-lingual natural language processing, zero shot and hybrid semantic-lexical approaches.

Anggraini Dyah Ayu Sekarlangit was born in Malang, Indonesia, on April 13, 2003. She received her Bachelor's degree in Information Technology from President University, Indonesia, in 2024, and is currently pursuing her Master's degree in Information Technology at President University, Indonesia.

Faqih Al Suman was born in Baubau, Indonesia, on April 5, 2003. He received her Bachelor's degree in Information Technology from President University, Indonesia, in 2024, and is currently pursuing his Master's degree in Information Technology at President University, Indonesia. He is pursuing a professional career in the Pharmaceutical Manufacturing industry.