Transformers vs. LSTMs: Body-Headline Stance Classification with DeBERTav3

Abstract— **In this study, we investigate how combining traditional machine learning techniques with advanced deep learning models improves stance detection in the Fake News Challenge (FNC-1) dataset. Specifically, we explore the effectiveness of integrating TF-IDF features with a Support Vector Classifier (SVC) and compare it to the performance of deep learning approaches like DistilBERT embeddings, Long Short-Term Memory (LSTM) networks, and DeBERTav3. Our aim is to find the best method for accurately classifying the relevance of an article body to its headline. This research seeks to identify a model that effectively balances the strengths of both machine learning and deep learning to enhance accuracy in detecting the stance conveyed by the text.**

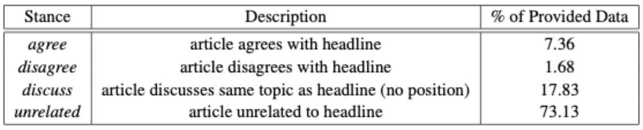
Keywords—FNC-1, TF-IDF, DistilBERT, SVC, LSTM, DeBERTav3

I. Introduction

Launched in 2017, the Fake News Challenge (FNC-1) [1] seeks to harness AI to identify the stance of a body text in relation to a headline, a pivotal issue in NLP due to the rise of misinformation [2]. This report adopts ML and DL methods for binary and multi-class classification within the FNC-1 dataset, comparing traditional ML algorithms and advanced DL models for effective stance detection. Additionally, it encompasses end-to-end testing, combining the optimal binary model with the multi-class model, benchmarking against the competition test set.

II. Problem Statement

Our study tackles stance detection, aiming to classify the relationship between a headline and body text into four categories: Agrees, Disagrees, Discusses, or Unrelated. Given the dataset's challenge, where 70% of articles are Unrelated, we approach this by dividing the task into Binary and Multi-Class Classification. Initially, we determine if an article is Related or Unrelated to the headline to address the data imbalance. Subsequently, for Related articles, we further classify them into Agrees, Disagrees, or Discusses. This two-step approach allows for more effective handling of the dataset's skewness and nuanced analysis of article stances.



Stance	Description	% of Provided Data
agree	article agrees with headline	7.36
disagree	article disagrees with headline	1.68
discuss	article discusses same topic as headline (no position)	17.83
unrelated	article unrelated to headline	73.13

Table 1. Distribution of Stances among Headline-Article Pairs.

III. Proposed Solution

After thorough research, I discovered that using an SVC with an RBF kernel on TF-IDF vectors achieved a 97.12% accuracy for the initial stage of classifying articles as related or unrelated [3][4]. Curiosity led to the implementation of DistilBERT embeddings into the SVC(RBF), significantly boosting the score to 99% in binary classification. Despite word embeddings being slower than TF-IDF features, their robustness as a feature extraction method is highlighted [6], underscoring the value of advanced techniques despite increased computational time.

DistilBERT utilizes a smaller, faster version of the BERT model, retaining most of its performance. It works by distilling the knowledge from BERT during training, essentially teaching DistilBERT to replicate BERT's word embeddings. This process involves training DistilBERT to predict the outputs of a pre-trained BERT model, allowing it to learn similar representations but with less computational overhead. The resulting word embeddings capture the context of words within a sentence, enabling effective language understanding with reduced model size and faster operation [5].

The selection of DistilBERT for embeddings was strategic, prioritizing efficiency without compromising the model's learning capacity, benefiting from knowledge distillation [5][17]. This decision aligns with the preference for BERT-like models in embedding selection [5][6][7][8][9], emphasizing the balance between performance and computational efficiency.

(1)
$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

(2)
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

(3)
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

(4)
$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$
$$h_t = o_t * \tanh(C_t)$$

(5)
$$\text{Dense output} = \text{softmax}(W_d \cdot h_T + b_d)$$

For binary classification, an LSTM architecture was chosen, recognized for its ability to capture temporal dependencies in text. In an LSTM network, the forget gate (1) controls the extent to which previous state information is forgotten. The input gate (2) decides which values will be updated in the current cell state. The new cell state (3) is a combination of the previous state and new candidate values, allowing the network to retain or discard information over sequences. The output gate (4) then utilizes the updated cell state to generate the output vector for the current timestep. For classification, the LSTM's last output **h_T** is fed into a Dense layer with a softmax activation function (5), producing a probability distribution over classes. The LSTM structure, characterized by its gates and cell state, is adept at capturing long-term dependencies within sequence data (see figure 1) [3],[11],[12].

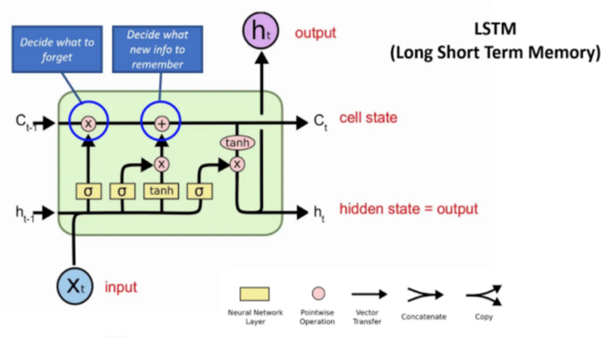
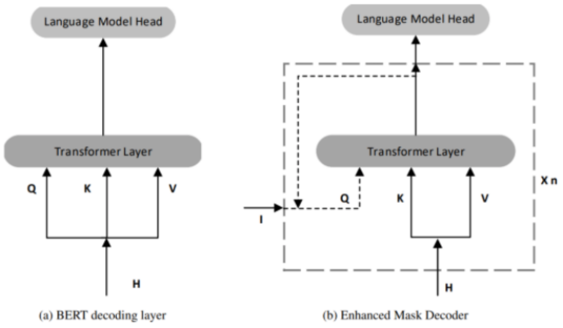


Figure 1. Long Short Term Memory Cell

Initially inclined towards employing a Bidirectional CE LSTM with Bidirectional Global Attention (BiCE LSTM BiGA) [3][11][12][21], we shifted our focus towards Transformer models after examining the efficiencies and capabilities of BERT and its variants [4][5][6][7][8][9][10][21], notably DeBERTa v3 [8]. This pivot was driven by their superior ability to process context and implement attention mechanisms. Specifically, DistilBERT [5][17] offered an appealing balance of performance and computational speed, cementing our selection of Transformer-based models for text classification due to their advanced architectural advantages.

DeBERTa improves upon BERT's architecture by introducing disentangled attention [8]. This mechanism separates the processing of word content and word position, enhancing the model's ability to understand the order and context of words in a sentence. Additionally, DeBERTa incorporates an 'Enhanced Mask Decoder' that considers both the relative and absolute positions of words, a notable advancement from previous models which considered these positions separately. These innovations allow DeBERTa, even with similar model sizes, to achieve state-of-the-art (SOTA) performance efficiently, outperforming larger models like T5, which have significantly more parameters. The diagram illustrates the traditional BERT decoding layer versus DeBERTa's Enhanced Mask Decoder [8] (see figure 2).



(a) BERT decoding layer

(b) Enhanced Mask Decoder

Figure 2. (a) BERT decoding layer, (b) Enhanced Mask Decoder

Consequently, DeBERTav3 was selected for its architectural advancements, making it particularly suited for the FNC-1 challenge [4][8]. This choice was motivated by the model's ability to exceed the performance benchmarks set by previous models, including RoBERTa's 93% accuracy, showcasing the continuous evolution in NLP model development for complex classification tasks [9].

IV. Analysis of Results

Model & Features	**Accuracy**	**Recall**	**Precision**	**F1-score**
SVC(RBF) & TF-IDF	97.76%	99%	97%	98%
SVC(RBF) & DistilBERT	99.27%	99%	99%	99%
LSTM & TF-IDF	97.56%	96%	97%	97%
LSTM + DistilBERT	73.07%	50%	54%	62%
DeBERTav3 (multi)	98.89%	98%	97%	97%

Table 2. Model and Features Accuracy

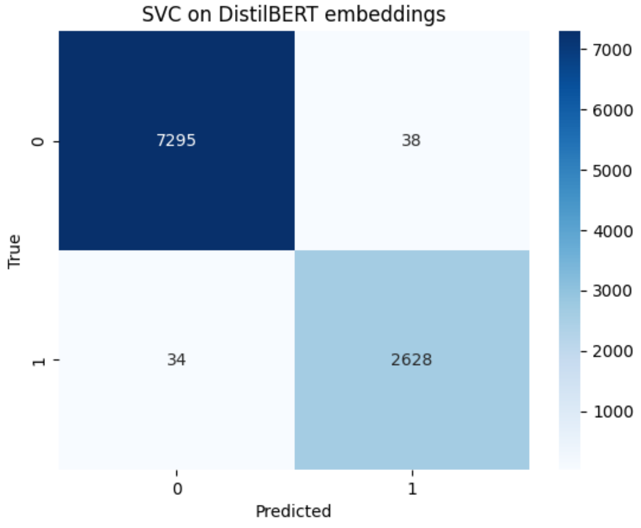


Figure 3.SVC on DistilBERT embeddings

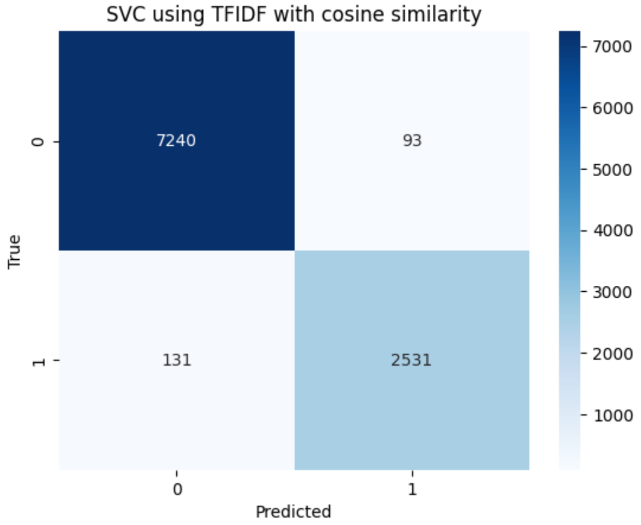


Figure 4. SVC using TFIDF with cosine similarity

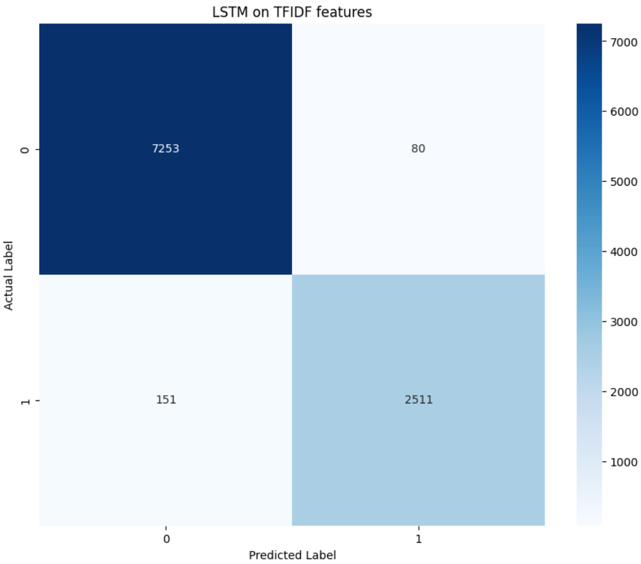


Figure 5. LSTM on TFIDF features

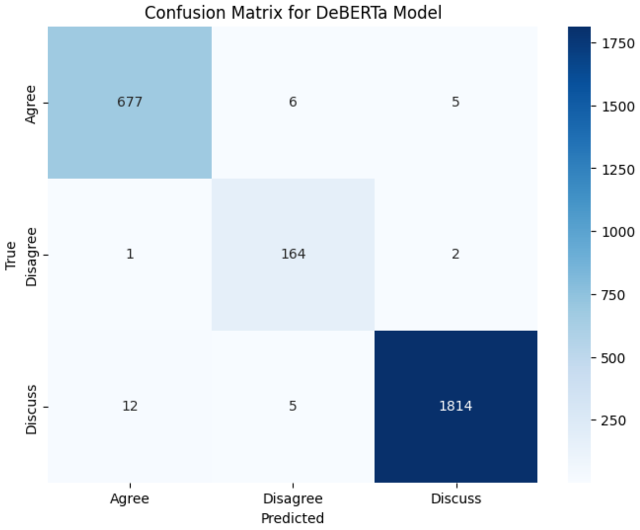


Figure 6. DeBERTav3 Multi-Classification

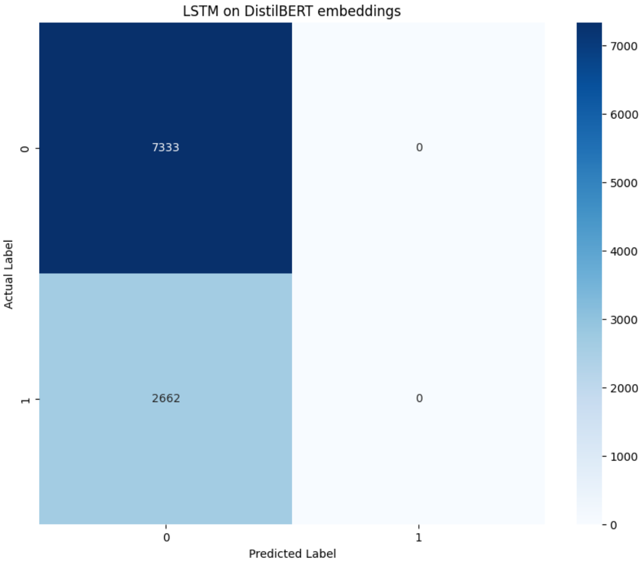


Figure 6. LSTM on DistilBERT embeddings

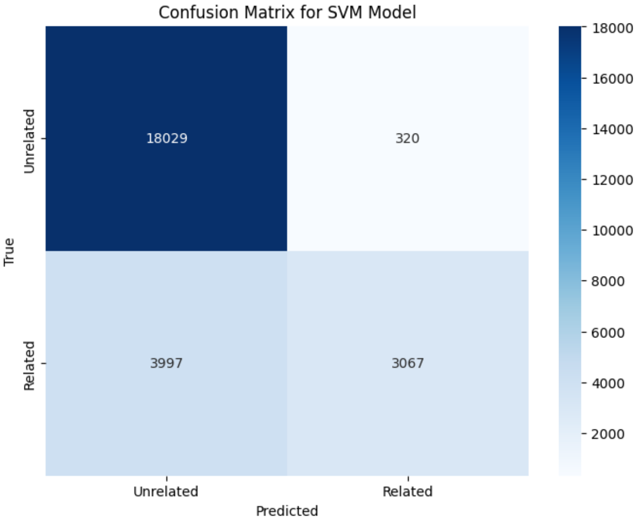


Figure 8. SVC + DistilBERT embeddings in end to end testing

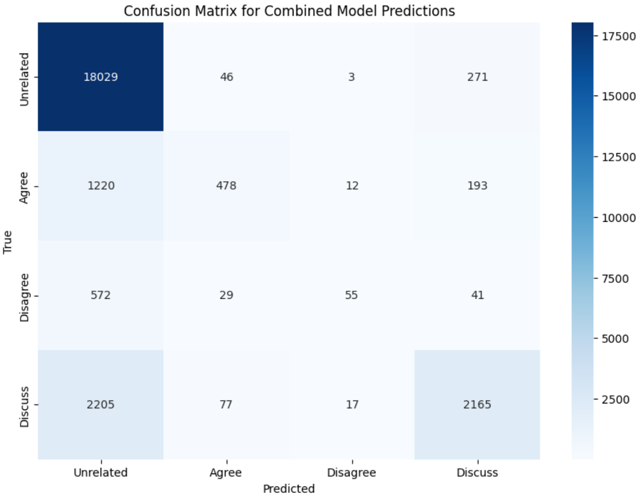


Figure 9. End to end test predictions

Figures 3 to 7 in our analysis indicate a general trend where models effectively identify the majority 'unrelated' class within a 2% margin, affirming their capability to capture the prevalent stance. In our analysis, the best performing model was the SVC with Transformer embeddings, while the LSTM with Transformer embeddings lagged. This disparity points towards the model's inherent limitations or possibly suboptimal hyperparameters. Despite attempts to balance the dataset by downsampling the majority class, the LSTM's performance did not improve, hinting that the issue might lie in the embeddings or the model's structure. Fine-tuning the LSTM could potentially elevate its performance; however, time constraints prevented such optimization. This scenario underscores the critical role of model selection and hyperparameter adjustment in text classification tasks.

Contrastingly, Figure 7 shows high accuracy for the multi-class model, likely a result of well-tuned hyperparameters. Important settings like a lower learning rate (6e-06) ensure gradual learning without overfitting, while the 'epoch' based evaluation strategy and a moderate number of epochs (5) balance between underfitting and overtraining. The use of AdamW optimizer with a linear scheduler and early warmup steps (50) provides a stable and effective learning process, further enhanced by mixed precision training (fp16), which increases computational efficiency on compatible GPUs [13] (see table 3 below).

Hyperparameter	**Value**
learning_rate	6e-06
per_device_train_batch_size	8
per_device_eval_batch_size	8
num_train_epochs	5
seed	42
weight_decay	0.01
evaluation_strategy	"epoch"
logging_strategy	"epoch"
optimiser	"adamw_torch"
lr_scheduler_type	"linear"
warmup_steps	50
fp16	True (if CUDA is available)

Table 3. List of Hyperparameters for DeBERTav3

V. Discussion

Addressing the data imbalance and the LSTM's simplicity in our study was crucial. A more balanced approach, through feature engineering or sampling techniques like down sampling the overrepresented "unrelated" class, might have enhanced the model's sensitivity to the "related" category. Increasing the LSTM's complexity, coupled with a reduced learning rate, could have potentially improved its ability to capture complex patterns within the "related" embeddings. While conditional attention encoding offers a promising avenue, it was deemed infeasible due to time and computational constraints.

If not for these limitations, exploring advanced models like Selective State Space Models, such as Mamba [19], or Retentive Networks [20], could be a strategic move. These alternatives promise reduced inference costs on GPUs and the potential to maintain or even boost accuracy. Such developments could provide a more resource-efficient and equally accurate approach to addressing the challenges presented by the FNC-1 dataset.

Future research into DeBERTa v3's applications for the FNC-1 challenge appears promising. Its advanced architecture and disentangled attention mechanism could be pivotal in deciphering complex textual relationships, potentially outperforming current models. Further testing could involve fine-tuning DeBERTa v3 with a more balanced dataset, potentially enhancing its ability to discriminate between subtle nuances in the data.

VI. Ethical Implications

When utilizing this proposed solution, ensuring that AI models like BERT do not inaccurately categorize unrelated texts due to biases in the training data, such as racial profiling, becomes paramount [14][16]. The data's extensive coverage of propaganda, controversial, and contextually sensitive topics amplifies the risk of misclassification, leading to potentially severe consequences in text classification tasks. This underscores the necessity for human oversight to verify that these models do not mistakenly link articles to sensitive categories like race or gender, thereby preventing harm to specific groups or the propagation of harmful ideologies [2][17]. The ethical implications and the possibility of future misuse of such technologies highlight the critical need for careful and responsible model training, evaluation, and deployment, ensuring AI systems are both effective and equitable.

VII. Conclusion

Our study implemented ML and DL models to classify the stance of articles in the Fake News Challenge, utilizing the novel application of DeBERTa v3. This advanced model, previously untested in the FNC-1 context, showed promising results. Our findings lay the groundwork for future research into DeBERTa's potential within fake news classification, signifying a pivotal step towards more sophisticated and reliable detection methods in the ongoing effort to combat misinformation.

References

[1] Jawad, Zainab A., and Ahmed J. Obaid. "A systemic literature overview of Fake News Challenge (FNC-1) dataset and its use in fake news detection schemes."

[2] Lazer, David MJ, et al. "The science of fake news." *Science* 359.6380 (2018): 1094-1096.

[3] Chopra, Sahil, Saachi Jain, and John Merriman Sholar. "Towards automatic identification of fake news: Headline-article stance detection with LSTM attention models." *Proc. Stanford CS224d Deep Learn. NLP Final Project* (2017): 1-15.

[4] Ghani Khan, Muhammad Usman Ghani, et al. "Fake News Classification: Past, Current, and Future." *Computers, Materials & Continua* 77.2 (2023).

[5] Sanh, Victor, et al. "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv 2019." *arXiv preprint arXiv:1910.01108* (2019).

[6] Morris, John X., et al. "Text embeddings reveal (almost) as much as text." *arXiv preprint arXiv:2310.06816* (2023).

[7] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[8] He, Pengcheng, Jianfeng Gao, and Weizhu Chen. "Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing." *arXiv preprint arXiv:2111.09543* (2021).

[9] Slovikovskaya, Valeriya. "Transfer learning from transformers to fake news challenge stance detection (FNC-1) task." *arXiv preprint arXiv:1910.14353* (2019).

[10] Babu, Raveen Narendra, Chung-Horng Lung, and Marzia Zaman. "Performance Evaluation of Transformer-based NLP Models on Fake News Detection Datasets." *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, 2023.

[11] Yu, Yong, et al. "A review of recurrent neural networks: LSTM cells and network architectures." *Neural computation* 31.7 (2019): 1235-1270.

[12] Staudemeyer, Ralf C., and Eric Rothstein Morris. "Understanding LSTM--a tutorial into long short-term memory recurrent neural networks." *arXiv preprint arXiv:1909.09586* (2019).

[13] Micikevicius, Paulius, et al. "Mixed precision training." *arXiv preprint arXiv:1710.03740* (2017).

[14] Torralba, Antonio, and Alexei A. Efros. "Unbiased look at dataset bias." *CVPR 2011*. IEEE, 2011.

[15] Hagendorff, Thilo. "The ethics of AI ethics: An evaluation of guidelines." *Minds and machines* 30.1 (2020): 99-120.

[16] Schramowski, Patrick, et al. "BERT has a Moral Compass: Improvements of ethical and moral values of machines." *arXiv preprint arXiv:1912.05238* (2019).

[17] Allein, Liesbeth, Marie-Francine Moens, and Domenico Perrotta. "Preventing profiling for ethical fake news detection." *Information Processing & Management* 60.2 (2023): 103206.

[18] Rana, Vineet, et al. "Compact BERT-Based Multi-Models for Efficient Fake News Detection." *2023 3rd International Conference on Intelligent Technologies (CONIT)*. IEEE, 2023. – BERT

[19] Gu, Albert, and Tri Dao. "Mamba: Linear-time sequence modeling with selective state spaces." *arXiv preprint arXiv:2312.00752* (2023).

[20] Sun, Yutao, et al. "Retentive network: A successor to transformer for large language models (2023)." *URL http://arxiv. org/abs/2307.08621 v1*.

[21] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).