computers

MDPI

Article

Unifying Sentence Transformer Embedding and Softmax Voting Ensemble for Accurate News Category Prediction

Saima Khosa [1,2], Arif Mehmood [1] and Muhammad Rizwan [2,*]

[1] Department of Information Security, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan; saimakhosa@yahoo.com (S.K.); arif.mehmood@iub.edu.pk (A.M.)
[2] Department of Information Technology, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan 64200, Pakistan
* Correspondence: rizwan2phd@gmail.com

Abstract: The study focuses on news category prediction and investigates the performance of sentence embedding of four transformer models (BERT, RoBERTa, MPNet, and T5) and their variants as feature vectors when combined with Softmax and Random Forest using two accessible news datasets from Kaggle. The data are stratified into train and test sets to ensure equal representation of each category. Word embeddings are generated using transformer models, with the last hidden layer selected as the embedding. Mean pooling calculates a single vector representation called sentence embedding, capturing the overall meaning of the news article. The performance of Softmax and Random Forest, as well as the soft voting of both, is evaluated using evaluation measures such as accuracy, F1 score, precision, and recall. The study also contributes by evaluating the performance of Softmax and Random Forest individually. The macro-average F1 score is calculated to compare the performance of different transformer embeddings in the same experimental settings. The experiments reveal that MPNet versions v1 and v3 achieve the highest F1 score of 97.7% when combined with Random Forest, while T5 Large embedding achieves the highest F1 score of 98.2% when used with Softmax regression. MPNet v1 performs exceptionally well when used in the voting classifier, obtaining an impressive F1 score of 98.6%. In conclusion, the experiments validate the superiority of certain transformer models, such as MPNet v1, MPNet v3, and DistilRoBERTa, when used to calculate sentence embeddings within the Random Forest framework. The results also highlight the promising performance of T5 Large and RoBERTa Large in voting of Softmax regression and Random Forest. The voting classifier, employing transformer embeddings and ensemble learning techniques, consistently outperforms other baselines and individual algorithms. These findings emphasize the effectiveness of the voting classifier with transformer embeddings in achieving accurate and reliable predictions for news category classification tasks.

Keywords: news category prediction; language models; soft voting; T5; RoBERTa; BERT; MPNet

Citation: Khosa, S.; Mehmood, A.; Rizwan, M. Unifying Sentence Transformer Embedding and Softmax Voting Ensemble for Accurate News Category Prediction. *Computers* **2023**, *12*, 137. https://doi.org/10.3390/computers12070137

Academic Editors: Phivos Mylonas, Katia Lida Kermanidis and Manolis Maragoudakis

Received: 5 June 2023
Revised: 6 July 2023
Accepted: 6 July 2023
Published: 8 July 2023

1. Introduction

Machine learning can help predict news categories because it can automate the process of putting news items into different categories such as politics, entertainment, sports, technology, etc. This can save a lot of time and work that would have been needed to put things into categories by hand. Different datasets of news articles such as BBC News and AG News can be used to train news category prediction models [1–3], which can then learn to spot patterns and traits that are unique to each category. Once these models have been trained [4], they can be used to automatically put new articles into categories as they are published. This makes it easier for news organizations to keep track of their material.

In addition, news category prediction can help users obtain news suggestions that are more relevant to them. By figuring out what the user likes and what they are interested in, the machine learning model can suggest articles from certain categories that are most useful

to the user. This can make the user's experience better and make them more interested in the news. Overall, machine learning can help news organizations be more efficient and effective by automating the process of putting news pieces into different categories and giving users personalized news recommendations.

Transformer language models such as RoBERTa, BERT, and T5 are state-of-the-art for encoding numerical representations of text in the form of transformer embedding [5]. Because of their usage of self-attention mechanisms, the transformer embedding features can capture deep contextualized information from the input sequence and give different parts of the sequence varied weights depending on their importance to the task at hand. The model can grasp the context in which each word appears thanks to the self-attention mechanism in transformers, which allows the model to record dependencies between distinct sections of the input sequence. Multi-head attention does this by splitting the input sequence into smaller pieces that are then processed independently.

By using attention heads, the model may zero down on certain subsequences within the input data and then combine the results from each attention head into a single output vector. The model is able to learn rich contextual information about each word in the input sequence thanks to this output vector, which contains data from all the attention heads. The output at each location in a transformer is also dependent on the complete input sequence, as the transformer uses a sequence-to-sequence design. By doing so, the model is able to pick up on dependencies over a large distance and comprehend the context of each word in connection to the full input sequence. Transformers are effective at a wide variety of NLP tasks, including text classification and question answering, because of their self-attention mechanism and sequence-to-sequence architecture.

The major purpose of this study is to evaluate the performance of alternative soft voting methods, specifically Softmax and Random Forest, by using a variety of sentence transformer embeddings as feature vectors. In more recent study, the use of soft voting with transfer embedding as a feature for news category prediction has been rarely studied. The purpose of this study is to determine which of the four sentence transformer embedding models [6], namely BERT, RoBERTa, MPNet, and T5 as well as their variants produces the best results when combined with the soft voting techniques of Softmax and Random Forest and applied to the two news datasets that are accessible on Kaggle. With the preceding hypothesis in mind, the purpose of this study is to seek the individual performance of Softmax and Random Forest as well as the soft voting of both by feeding sentence transformer embedding using evaluation measures such as accuracy, F1, precision, and recall. This study also makes a multifold contribution, which is to seek the individual performance of Softmax and Random Forest. Using two distinct news classification datasets, the performance of the multi-class classification system is evaluated and analyzed.

2. Related Work

In this section, we take a look at the contemporary research published on news or text classification with machine learning and deep learning. We also emphasize the central results, theories, and methods that have informed our perspective on this research with respect to the research gap related to news category prediction.

For example, [7] provides a quantitative evaluation of three Support Vector Machine (SVM) classifier variants on multi-category text classification with news data using different SVMs. These binary classifiers are extended by the authors to accommodate multi-category data via the One-Against-All strategy. They use simulations using benchmark UCI News datasets (including Reuters and 20 newsgroups) to evaluate the efficacy of each approach, with the feature set constructed using the Term Frequency-Inverse Document Frequency matrix. The outcomes demonstrate that LS-TWSVM is superior to the other two SVM versions in the context of news classification issues both in terms of accuracy and time complexity. The authors suggest that future work look into whether or not other SVM-based multi-classification techniques can benefit from further improvements and that the creation of new SVM-based classifiers for similar applications could significantly increase

the importance of SVM-based methods on text categorization problems. Furthermore, in [1], it provides 98% accuracy of BBC News datasets using CNN, RNN, SVM and a hybrid model. The sudty [8] provides 95 % accuracy on the BBC News dataset using using data augmentation and LSTM.

An innovative methodology [9] for the automatic categorization of Arabic news articles is discussed in this article. The technique uses machine learning classifiers such as Logistic Regression, Naive Bayes, Random Forest, XG Boosting, K-Nearest Neighbors, Decision Tree, Stochastic Gradient Descent, and Multi-Layer Perceptron to extract features from the dataset and make predictions. On Mendeley's Arabic news dataset, the proposed algorithm achieves an accuracy of above 95%. With certain tweaks, such as eliminating noise and integrating feature vectors, the paper's proposed methodology might be extended to additional datasets in multiple languages for text categorization, as emphasized in the article. What this article adds is a mechanism for automating the categorization of Arabic news articles that may be used with different datasets and languages. There are other example of Arabic news exist in literature, e.g., [9–11].

The application of word2vec Convolutional Neural Networks (CNNs) to the problem of topic classification in news articles and Twitter posts is explored in [12]. The research compares the effectiveness of CNNs trained with word2vec's Continuous Bag-of-Word (CBOW) and Skip-gram word-embedding techniques as well as CNNs trained without word2vec models. The results show that word2vec greatly enhances the classification model's precision. When compared to the Skip-gram model, the CBOW model consistently provides superior accuracy. When it comes to news stories, the CBOW model excels, but when it comes to tweets, the Skip-gram model triumphs. The research also indicates that varying algorithms according on the data type can improve results. There are other such examples of research articles which solve the news classification problem with similar techniques, e.g., [13–15]. The study's findings indicate that properly learnt word embedding improves CNN's ability to classify news articles and tweets, and that different word-embedding models should be used for different types of input.

Using fastText to produce text vectors and a Convolutional Neural Network (CNN) to automatically extract features, the research [16] provides a deep learning model for Amharic news document categorization. The suggested method outperformed popular machine learning algorithms including SVM, MLP, DT, XGB, and RF on a dataset consisting of six types of news articles, with a classification accuracy of 93.79 percent. The outcomes show promise for the suggested model's applicability in applications requiring the classification of Amharic documents. Additional classes and datasets will be investigated in forthcoming research. SVM, Logistic Regression, CNN, and others are developed and tested for Bangla news classification in the article [13,14,17]. The researchers created Potrika, a dataset of 664,880 Bangla news articles in eight categories from six famous Bangladeshi online news sites from 2014 to 2020. They developed Latent Dirichlet Allocation-based automatic labeling methods and tested single-label and multi-label article categorization approaches. GRU and FastText performed best for manually labeled data, whereas KNN and Doc2Vec performed best for automatically labeled data. The researchers propose to use hybrid deep learning algorithms with multiple word-embedding approaches to better NLP research in Bangla and other languages.

The general methodology regarding how machine learning can help social media health news analysis is proposed [18]. A methodology for collecting, modeling, and displaying health-related patterns across 13 Twitter news sources is proposed. SVM, Naive Bayes, Logistic Regression, and Decision Tree methods are used in a series of tests to find patterns in the data. The new approach performed better than earlier research and may help physicians, patients, and healthcare organizations make decisions.

The research [19] offers a 30-topic benchmark Vietnamese online news article dataset for multi-label text categorization. The authors alter the Vietnamese text classification pipeline by reducing feature vector dimension depending on phrase frequency across the corpus. They classify using neural network models without feature selection procedures

and perform well. The new dataset and neural network models can be used for Vietnamese news article classification.

Given the wide variety of approaches to categorization and presentation adopted by online Bangla news outlets [20], this research explores the difficulty of doing so on the basis of readers' preferences. The authors suggest an automated approach that employs machine learning models to address this issue, and they assess the efficacy of various models, including BiLSTM, Char-CNN, GRU-LSTM, LSTM, BERT, and Text-GCN, using a sample dataset. Text-GCN was found to have superior accuracy, precision, recall, and F1 score compared to the other models. Despite data scarcity, the study demonstrates Text-GCN's promise for categorizing Bangla news items [21,22]. Another example of text classification for depression intensity detection can be seen in [23].

In another study [24], the authors explore the use of taxonomies to structure online news databases. The authors address the need for data pre-processing in transforming unstructured data into a more manageable format prior to classification, using the removal of punctuation, stop words, stemming, lemmatizing words, and special characters as examples. To find the most effective news categorization module, they evaluate different classification models such as SVM, Naive Bayes and Logistic Regression. The research emphasizes the significance of data classification in making online news sources more user-friendly. The paper [25] examines the history and current state of text classification research, and it makes projections about where the discipline is headed in terms of both research priorities and new frontiers. Next, they build a text categorization model for network news using deep learning and explain how each component works. They automatically extract and identify information from news texts using a convolutional neural network and a dense Word2Vec word vector representation, which is an improvement over previous methods. The experimental results demonstrate the superiority of their suggested method over baseline text classification techniques, paving the way for more efficient news information management. The authors also offer a hybrid model that takes the best features from all three approaches while maintaining a manageable computational footprint. For the field of text classification as a whole, this research serves as a theoretical reference while also providing a workable model for improved news information services.

Deep learning-based methods for news article classification are discussed in the text [26]. The authors implement the Fasttext model for text classification using a Long-Short Term Memory (LSTM) neural network. In their analysis, they find that LSTM using the Fasttext model yields the best classification results compared to Word2vec and Doc2vec models. In light of the ever-increasing volume of online content, the study highlights the value of automatic text classification. The other similar examples can be found in [27,28]. Furthermore, the study [29] discusses the problem of online news websites wasting money on advertising by showing readers ads that are of little interest to them. Using a mix of word embedding and Convolutional Neural Network (CNN), the authors suggest a method for categorizing English news into four distinct groups. In comparison to all of the baseline approaches, the suggested method outperformed them all with high macro-f1 and micro-f1 scores of 0.90 and 0.89, respectively. The benefits to online news portals and advertisers are discussed along with the importance of integrating cutting-edge methodologies such as deep learning and NLP for news classification. Table 1 summarizes the latest existing problems from the literature.

Table 1. Summary of latest existing problems from the literature.

Study	Research Methodology	Dataset	Finding and Results
Jang et al. [12]	Topic classification in news article using CNN and Word2Vec	Twitter posts	Analyzed performance of different word embeddings with respect to news and tweets
Saigal et al. [7]	Multi-category news classification using different variants of SVM	Reuters and 20 newsgroups	Superior performance in terms of accuracy and time complexity

Table 1. *Cont.*

Study	Research Methodology	Dataset	Finding and Results
Endalie et al. [16]	Amharic news document categorization using CNN	Amharic news	Superior performance over traditional machine learning algorithms
Alfonse et al. [9]	Arabic news classification using Logistic Regression, Random Forest, KNN, etc.	Arabic news	Achieve better accuracy eliminating noise and using integrating feature vectors with 95 percent accuracy
Ugwuoke et al. [8]	BBC News classification using data augmentationa and LSTM	BBC News	95 percent accuracy
Karaman et al. [1]	BBC News classification using CNN, RNN, SVM and hybrid model	BBC News	98 percent accuracy

3. Methods

In this section, we will discuss the machine learning models and transformer language models that were utilized in this study.

3.1. Random Forest

Random Forest is an algorithm for machine learning that pertains to the family of ensemble learning. It combines the predictions of multiple decision trees to create predictions that are more accurate and robust. In a Random Forest, each decision tree is trained on a distinct subset of the training data, and the splitting procedure considers a random subset of features. This randomness increases tree diversity, preventing overfitting and enhancing generalization. Every decision tree in the forest independently predicts the input to make predictions, and the last prediction is determined by a majority vote or by averaging the classification of the individual trees. This method serves to reduce variance, increase stability, and effectively manage high-dimensional datasets. Random Forests are extensively used for classification and regression tasks due to their ability to handle complex relationships, missing data, and feature importance rankings.

3.2. Softmax Regression

The classification process known as Softmax regression, which is also referred to as multinomial logistic regression, is used to give categorical labels to the data that are input. It is an extension of the logistic regression model that handles cases with more than two different classes. In Softmax regression, a generalization of the sigmoid function called the Softmax function is used to turn a vector of real-valued scores into a probability distribution for the classes. This is accomplished by using the Softmax function. The scores are normalized by the Softmax function, which does so by exponentiating them and then dividing by the sum of the scores after they have been exponentiated. This ensures that the probabilities that are produced add up to one. During training, the algorithm learns the weights and biases associated with each class by minimizing the cross-entropy loss between the predicted probability and the genuine class labels. This is completed to ensure that the predicted probabilities are as close as possible to the actual class labels. Image identification, natural language processing, and sentiment analysis are all examples of applications that make extensive use of Softmax regression's multi-class classification capabilities.

3.3. Sentence Transformer Embedding Using Pooling

Sentence transformer embedding using pooling is a method for producing vector representations of variable-length sentences or paragraphs with fixed-length vectors. It entails applying a pre-trained transformer model to the input text, such as BERT or RoBERTa, to acquire a sequence of contextualized word embeddings [30,31].

Tokenization is the process of breaking a sentence down into its component parts, usually using a mechanism such as WordPiece or Byte Pair Encoding. This process separates the sentence into its component parts or tokens. The semantic meaning of each word is then

captured by changing each token into its word embedding. Embeddings for words can be given a random start or pre-trained with unsupervised techniques such as Word2Vec or GloVe. During training, the transformer model may acquire knowledge of the word embeddings on its own.

To compensate for the fact that transformers do not automatically preserve word order, positional encoding is performed on the input embeddings to reveal these data. The position of each word in the input sequence is encoded as a fixed-length vector appended to the word embedding. A set of stacked transformer layers is applied to the input embeddings that carry positional encodings. Each transformer layer is split into two parts: a feed-forward neural network and a multi-head self-attention mechanism. Words in an input sequence can have their context and dependency on one another captured by these levels.

The outputs of the transformer layers are pooled in order to obtain a fixed-length sentence embedding. Average embeddings, maximum or mean value over time steps, and attention-based pooling are all examples of common pooling algorithms. To arrive at the final fixed-length sentence embedding, the pooled representations are often processed further through a linear layer followed by a non-linear activation function.

The resulting pooled embeddings encapsulate the meaning of the input sentence in a dense, fixed-length vector that can be used for a variety of NLP tasks, including sentiment analysis, text classification, question answering, etc. [32]. Furthermore, the transformer models used in this study are listed in following subsections, and mean pooling is used in all experiments presented in this study.

3.3.1. Megatron Pre-Trained Network Language Model

The Megatron Pre-trained Network (MPNet) is a large-scale transformer-based language model optimized for distributed-system training efficiency. To put it simply, MPNet works by receiving a stream of input tokens (often words or subwords) and passing them through a hierarchy of multilayer transformer blocks. When predicting an output, the model can zero in on specific parts of the input sequence thanks to the attention techniques applied by these transformer blocks.

Masked language modeling is used to train MPNet; this technique involves masking out random values of input tokens and training the language model to predict the accurate tokens based on the context of the surrounding words. Furthermore, MPNet employs a cutting-edge training approach dubbed "Adaptive Prompting", which enables the model to automatically modify the complexity of the training instances in light of the model's current skill level. When it comes to text classification, question answering, and language translation, MPNet is a very advanced language model that is built to perform well.

In order to process sequential data, both MPNet and the transformer architecture rely on multilayered transformer blocks, which they have in common. Nonetheless, the two models diverge in significant ways. When compared to the original transformer, MPNet is a massive upgrade. Unlike the original transformer, which had 110 million parameters, MPNet can have up to 3.5 billion. MPNet can now capture more intricate linguistic connections and patterns as a result of its increased size. MPNet has a unique pre-training method in comparison to the traditional transformer. Masked language modeling, which is used to pre-train MPNet, involves masking out a random part of input tokens while training the model to predict the accurate tokens based on the context of the neighboring words. Conversely, the original transformer employs "autoencoding" as a pre-training approach, where the model is taught to recover the correct input sequence from a corrupted one. MPNet is intended to make training on distributed systems more effective. Several methods are employed to lessen the amount of time spent on training due to processor-to-processor communication.

MPNet employs a cutting-edge training approach dubbed "Adaptive Prompting", which allows the model to automatically modify the complexity of the training instances in light of the model's current skill level. This skill is missing from the classic transformer.

Although there are many similarities between MPNet and the transformer architecture, MPNet is a more complex model that was developed with distributed training in mind. It also features several novel training techniques that make it more robust and flexible than the original transformer. Four variants of MPNet are used in this study, i.e., MPNet v1 [33], MPNet v2 [34], MPNet v3 [35] and MPNet v4 [36].

3.3.2. BERT Language Model

Natural language processing (NLP) tasks were benefited by BERT (Bidirectional Encoder Representations from Transformers), which is a transformer-based model architecture. The use of a masked language modeling (MLM) target during pre-training allows it to extract bidirectional contextual information from text. Transformer encoders are a key component of the encoder stack that makes up BERT's architecture. A multi-head self-attention mechanism and a feed-forward neural network make up the two sub-layers of each transformer encoder. In order to process the input text, BERT employs a stack of these encoders.

During the pre-training phase, BERT trains on a huge unlabeled text corpus to learn contextual representations of words. It presents MLM (masked language modeling) and NSP (next sentence prediction) as training goals. In MLM, some of the input words are obscured at random, and the model is tasked with identifying them from their context. Predicting the co-occurrence of two sentences is a key part of NSP. Contextual knowledge and deep word representations are two things BERT can acquire from these pre-training activities. Once BERT has been pre-trained, it can be tweaked for use in a variety of NLP applications. In most cases, additional task-specific layers are added to the model on top of the BERT architecture that has already been trained. Text classification, namely entity identification, question answering, and other tasks are only few examples of how BERT's task-specific layers help it learn and adapt.

BERT's ability to self-attentively capture long-range dependencies in the input text is a result of the transformer architecture it employs. Each word in a sentence might "pay attention" to the others, taking in the context of what came before and what came after. When compared to models that depended just on a one-way transmission of data, this ability to recognize context from both ends is a major improvement. Overall, BERT's architecture has shown great success in capturing rich contextual representations of text, leading to significant advances in a number of NLP tasks thanks to its multilayer stack of transformer encoders, self-attention mechanism, and pre-training objectives. Two variants of BERT used in this study, i.e., DistilBert [37] and BERT2.0 [38], both generate a vector with a fixed dimension length of 768.

3.3.3. Robustly Optimized BERT Pre-Training Approach Language Model

In 2019, a group of researchers at Facebook AI introduced RoBERTa (A Robustly Optimised BERT Pre-training Approach), which is a transformer-based language model. RoBERTa is a variant of the BERT (Bidirectional Encoder Representations from Transformers) architecture that was trained with a different set of pre-training objectives and procedures in order to obtain better results on downstream NLP tasks. The enormous corpus of text data used to pre-train RoBERTa is quite similar to the data used to pre-train BERT with some key differences. Compared to BERT, RoBERTa's training involved more pre-training stages, longer sequences, and larger batches. Unlike BERT, which uses a static mask during training, RoBERTa was trained with a dynamic mask, which masks out various tokens at random throughout each round of training.

As an added bonus, RoBERTa was trained using a method termed "training with large amounts of unlabeled data across multiple domains" (TLM), which entails pre-training the model on different domains of text data before fine-tuning it on a single downstream NLP task. This method was demonstrated to boost the model's generalization performance across numerous NLP tasks [39]. To further enhance the stability and convergence of the training process, RoBERTa employs a variation of the optimizer employed by BERT dubbed AdamW, which combines the Adam optimizer with weight decay regularization. Across

multiple natural language processing (NLP) benchmarks, including sentiment analysis, text categorization, and question answering, RoBERTa has proven to be a very successful and robust language model.

Two variants of RoBERTa are used in this study for obtaining sentence transformer embedding. First, all-roberta-large-v1 [40] is used as it converts texts and paragraphs into a dense vector space with 1024 dimensions to capture the contextualized. Secondly, all-distilroberta-v1 [41] is used, and it converts text into a 768-dimension vector.

3.3.4. Text-to-Text Transfer Transformer Language Model

In 2019, a group of Google researchers unveiled T5 (Text-to-Text Transfer Transformer), which is a transformer-based language paradigm. T5 stands out from other transformer-based language models since it is intended to serve as a generic model for a wide range of NLP applications. T5 is pre-trained on a wide variety of text-to-text transfer tasks, unlike other models which are generally fine-tuned on a specific job. This means that it learns to map input text to output text in a wide range of scenarios.

T5 is built on a slightly modified version of the transformer design. T5 specifically ushers in a new "decoder-only" architecture, with no encoder layers and just decoder layers present in the transformer. In contrast, models such as BERT and GPT have separate layers for encoding and decoding. T5 is pre-trained with a collection of text-to-text transfer tasks, each of which requires a unique mapping of input text to output text. One task could be to convert one language into another, while another could be used to condense a lengthier text into a more concise one. T5 is able to generalize successfully to numerous NLP tasks because it has been pre-trained on a large variety of activities.

T5's adaptability is a major strength. For the most part, any natural language processing task may be tweaked to perfection by simply giving an appropriate input–output pair as part of training data, because the model is intended to be general purpose. This adaptability means the model can be employed in a variety of natural language processing tasks. In conclusion, T5 is a state-of-the-art language model that excels in a wide variety of NLP tasks thanks to its effectiveness and adaptability. It is a very flexible model that can be customized for a variety of NLP applications because of its novel text-to-text transfer pre-training approach and decoder-only design. Three variants of T5 are used in this study, i.e., T5 base [42], T5 Large [43] and T5 XL [44].

4. Experimental Setup

The complete workflow of this study can be seen in Figure 1. The data are divided into two sections: the train section and the test section. Five-fold cross-validation is used to ensure the reliable performance and results from the experiments conducted in this study. For the purposes of ensuring that each category in the news category categorization datasets is represented in an equal manner, the data are divided in a stratified method. The training part contains 85 percent of the total instances, and the test part contains 15 % of the total examples from each dataset. In the following stage, the embedding of each word in the news text is generated by making use of all of the transformer models that were discussed in the preceding section. The final, hidden layer of a transformer model, such as BERT or RoBERTa, is selected to serve as the embedding for each word in the news document. This is completed for embedding purposes. Here, 768 is the standard dimension that must be met in order for a vector to be collected from the final hidden layer of every transformer model in all the transformer models used in the study except for RoBERTa Large, which is 1024, and this is the one reason RoBERTa Large is able to obtain a deeper context of each word in the document. Then, a single vector representation known as sentence embedding is created as a result of mean pooling's calculation of the average of these word vectors. This representation accurately conveys the overarching content of the news article.

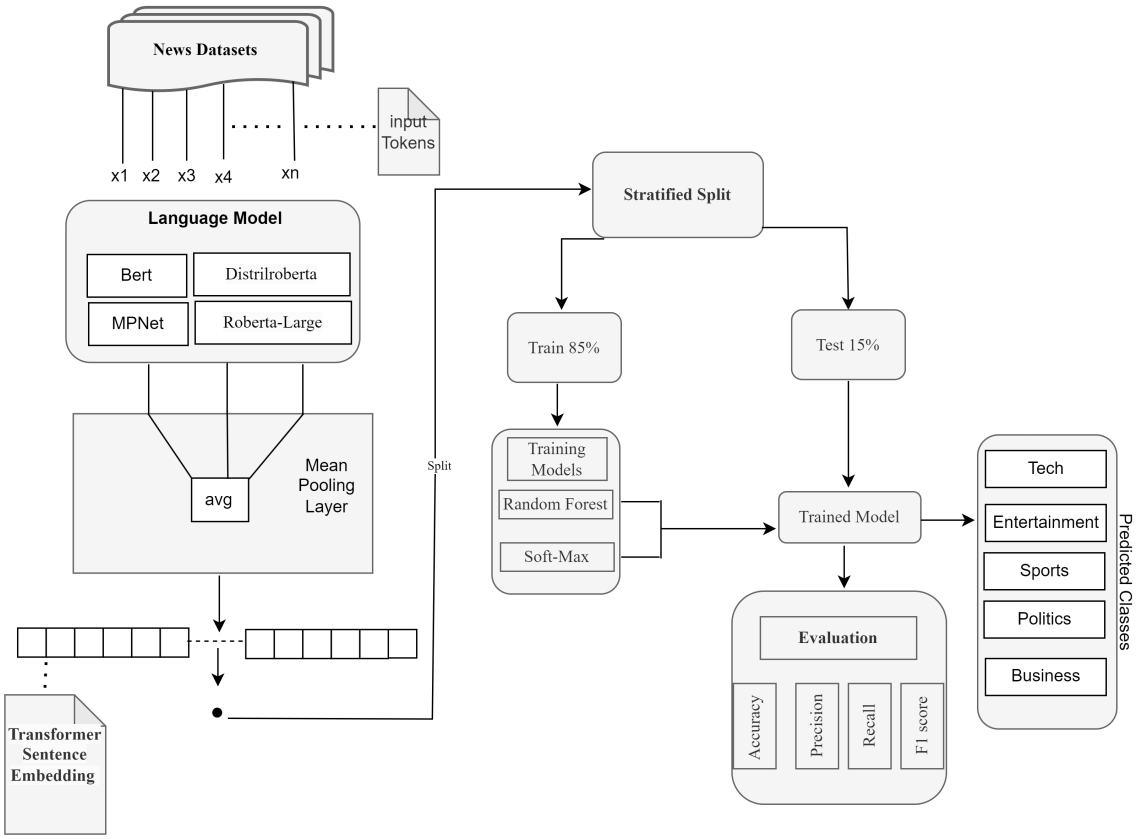

Figure 1. Workflow of the study.

The method is particularly helpful when working with tasks that need input of a constant length, such as text categorization, which is the situation with our investigation into the classification of news categories. After that, the feature vectors derived from each sentence embedding are input into two different algorithms, namely Softmax regression and Random Forest. Both algorithms has been trained on the training dataset and then tested to see how well they performed individually. It has been noted that the use of soft voting improves the performance of both Softmax regression and Random Forest, notably in terms of the F1 score. Calculating the macro-average of F1 over all classes allows for further comparison of the effectiveness of various transformer embeddings when applied to the same experimental conditions.

In a nutshell, the experimental setting included the stratified separation of data into training and test sets, which was then followed by the development of word embeddings through the utilization of transformer models. After that, the approach of mean pooling was used to construct sentence embeddings, which were then used to capture the overarching meaning of news documents. Both the Softmax regression and Random Forest methods were trained and evaluated on their own using these phrase embeddings as feature vectors. The effectiveness of various transformer embeddings was evaluated and compared based on the macro-average F1 score obtained from all of the classes. This methodical methodology guaranteed a unified and consistent experimental setting for the machine learning studies that were carried out in the realm of news category classification.

The Random Forest algorithm's hyperparameter of 50 trees is set initially, and the "gini" rule is used to split nodes. The trees can go as deep as they want, and a node needs at least two samples to split and one sample to be a leaf. The method takes into account all features when deciding how to split them, and the number of features is equal to the square root of the total. There is no cap on how many leaf nodes can be in a tree. The method does not use out-of-bag samples to estimate how accurate it is. Instead, it uses bootstrap samples. In the Logistic Regression algorithm's hyperparameter choices for multi-class classification,

L2 regularization with a penalty value of 1.0 is used. It uses the "lbfgs" solver method, which works well for multi-class problems and can converge in as few as 100 steps. The model fits an intercept term, but it does not scale the intercept. Since the class weights are not made clear, all classes are given the same amount of weight. The seed for the random number generator is not set, and the solver's level of detail is set to 0. The "auto" choice lets the data decide how to handle multiple classes.

4.1. Dataset

Two news datasets have been used in this study, which are presented in the following subsections.

4.1.1. BBC News

The BBC News dataset on Kaggle includes news stories from 2004 and 2005 that were originally published by the BBC [8,45]. The dataset contains a total of 2225 news articles. Each news article in the dataset has the following attributes: "Article Text", which contains the article's primary textual content, and "Category", which includes the category or topic to which the news article pertains. It may fall into one of the following five categories: business, entertainment, politics, sport, or technology. The representation of class counts can be observed in Figure 2.

The dataset contains a total of 2225 news articles. Each news article in the dataset has the following attributes: "Article Text", which contains the article's primary textual content, and "Category", which includes the category or topic to which the news article pertains. It may fall into one of the following five categories: business, entertainment, politics, sport, or technology.

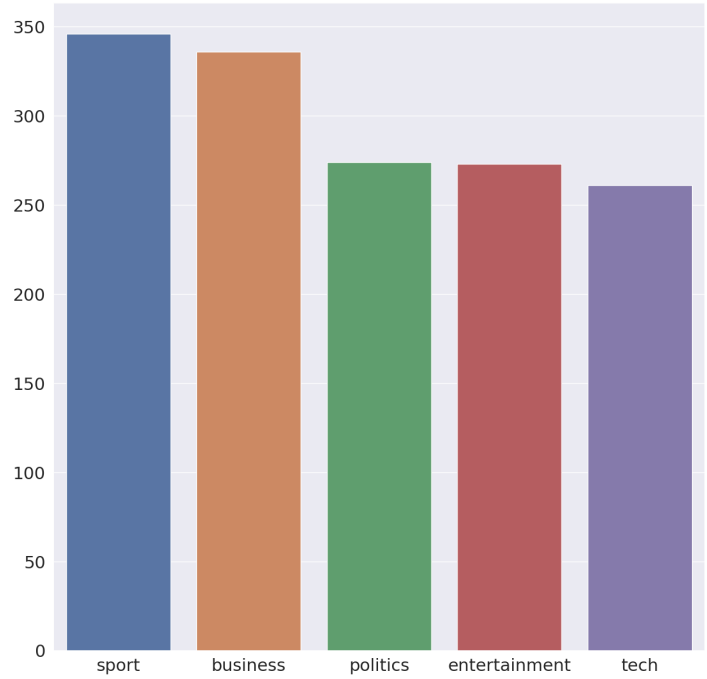

Figure 2. BBC news dataset class-wise number of instances.

There are many natural language processing activities that can benefit from this dataset. These include topic modeling, sentiment analysis, and text categorization. Since the articles are written in English and are contained in a Comma Separated Values (CSV) file, they can be read and processed by a wide range of programming languages and machine learning frameworks. Various machine learning algorithms and natural language processing tech-

niques have been evaluated and compared in academic studies and publications using the BBC News dataset.

4.1.2. AG News

The AG News dataset on Kaggle is a compilation of news stories on World News, Sports, Business, and Science Tech. There are a total of 1,200,000 articles in the dataset (300,000 in each of the 10 categories) [46,47]. The content was culled from the English-language version of Google News, which was published between 1999 and 2003. Attributes for each news instance in the dataset include "Title", which contains the headline, "Description", which contains a brief description or summary of the news item, "Text", which contains the article's primary text content, and "Category", which specifies the topical category to which the news item belongs; these categories are World, Sports, Business, and Science and Technology.

Text classification problems are complicated by the short length and high noise levels of the articles in the AG News dataset. However, the dataset has proven to be beneficial in assessing the efficacy of various machine learning algorithms and NLP methods. Examples of deep learning models that have been tested on this dataset include Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) as well as variants on these. Models trained on the AG News dataset and then fine-tuned on other datasets have been studied as part of the field of transfer learning. The representation of class counts can be observed in Figure 3.

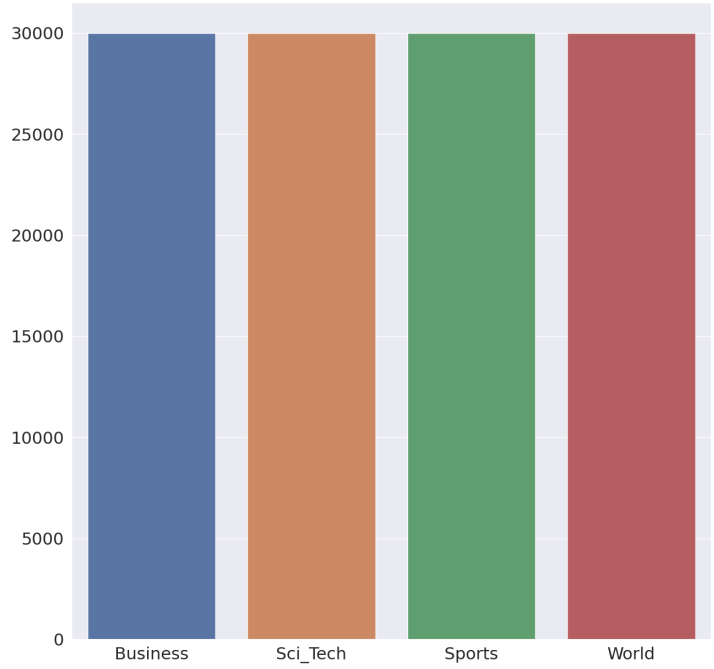

Figure 3. AG News dataset class-wise number of instances.

The AG News dataset has been utilized for both academic purposes, including automatic news categorization, content screening, and personalized news recommendation systems. The dataset can be downloaded by anybody and used for any purpose, including academic study. Researchers and data scientists with an interest in natural language processing and text classification will find the AG News dataset to be an invaluable resource. Its availability has allowed for the creation of a number of machine learning models for use in NLP. Table 2 gives the sample data from AG News dataset.

Table 2. Sample data rows from AG News dataset. 1 represents Business, 2 represents Sci_Tech, 3 represents Sports and 4 represents World.

Class Index	Title	Description
4	Tech 2004: Where the Candidates Stand	We uncover the candidates' views on Internet taxes, privacy, and other tech hot buttons.
1	Ethnic violence in China region	Martial law has been imposed in parts of the Chinese province of Henan after ethnic clashes in which witnesses say several people were killed.
3	Jamdat shares up 45 percent in first day of trade	Shares of wireless game maker Jamdat Mobile Inc. (JMDT.O: Quote, Profile, Research) rose as much as 45 percent on Wednesday in their first trading day, boosted by investors betting.
2	Rogers rejoices after gold rush	We interrupt our regularly scheduled programming to bring you something from October. Don't touch that dial. The Astros may remember their 88th victory more for what it wasn't than what it was.

4.2. Evaluation Metrics

Each experiment's performance for news category classification is tested and compared using evaluation criteria such as F1, accuracy, precision, and recall. Accuracy measures how well a model does in making predictions relative to how many predictions it makes. While it works well for evenly distributed datasets, its failure to account for class imbalance can lead to false conclusions. Precision is a measure of how often a prediction of positivity turns out to be right. Its primary goal is to reduce the number of false positives, making it an excellent choice when those results would be expensive or otherwise unwelcome. The less likelihood of false positives a model has, the more precise it is.

The F1 score is a statistic that averages the two separate values of precision and recall. It works especially well with unbalanced datasets. It is a balanced measure of a model's efficacy because it takes into account both accuracy and recall.

Alternatively called sensitivity or true positive rate, recall quantifies how often a positive outcome was properly predicted. It is helpful in situations where it is essential to detect all positive instances, even at the cost of false alarms, and its primary goal is to do so with as few false positives as possible. The higher the recall, the better the model performs. Overall correctness is measured by accuracy, whereas precision and recall indicate how well certain sorts of errors are handled (false positives and false negatives, respectively). The situation at hand and the weight attached to various kinds of error dictate which evaluative parameter should be used.

5. Results and Discussion

Experiments were run to examine the efficacy of using transformer sentence embeddings with soft voting of Softmax and Random Forest, with accuracy, precision, recall, and F1 score serving as the macro-average measure. The results of the experiments and the efficacy of several transformer types were thoroughly analyzed thanks to this exhaustive assessment. The experiment aimed to take a comprehensive look at the performance measures by using the macro-average measure to do so. For a complete picture of the experiments' findings, we calculated the precision, recall, and F1 score. Several studies were conducted to examine the effectiveness of using transformer sentence embedding in tandem with the soft voting of Softmax and Random Forest. The goal was to determine what setup will function most effectively and produce the best outcomes. The performance was evaluated fairly by evaluating the macro-average measure, which took into account many parameters simultaneously.

The results offer light on how various experiments and transformer sentence embeddings fared, which is helpful. After careful examination, it was clear that some permutations were superior in certain settings. Each experiment and transformer sentence embedding configuration was evaluated based on its accuracy, precision, recall, and F1 score. The experiment

gave a thorough evaluation of the available transformer models and embeddings by utilizing the soft voting of Softmax and Random Forest. In order to determine which setups worked best, we used the macro-average metric, which allowed for an objective comparison. Each experiment and transformer sentence embedding was evaluated based on its accuracy, precision, recall, and F1 score using the macro-average method. The most optimal settings were analyzed in depth by combining the soft voting of Softmax and Random Forest. These results provide a unified framework for making sense of the disparate data gathered from experiments, shedding light on how various transformer models and their embeddings perform.

The machine learning experiments focused on calculating language embeddings using versions of four primary transformer models. The goal was to compare how well each model did on the Random Forest benchmark using the BBC dataset. The result of all experiments on BBC dataset can be found in Table 3. Using the Random Forest framework, the phrase embeddings of MPNet versions v1 and v3 performed the best, achieving a remarkable F1 score of 97.7%. DistilRoBERTa's phrase embedding came in close second with an F1 score of 97.3%. These results demonstrate that when combined with Random Forest, these transformer models provide higher performance for the specified task.

Table 3. Performance of experiments on the BBC News dataset.

Model	Transformer Sentence Embedding	Accuracy	Precision	Recall	F1
Random Forest	RoBERTa Large	0.955	0.956	0.953	0.954
	DistilRoBERTa	0.973	0.974	0.973	0.973
	T5 Base	0.960	0.959	0.958	0.958
	T5 Large	0.960	0.960	0.958	0.959
	T5 XL	0.973	0.974	0.972	0.972
	MPNet V1	0.978	0.979	0.977	0.977
	MPNet V2	0.969	0.968	0.968	0.968
	MPNet V3	0.978	0.977	0.978	0.977
	MPNet V4	0.964	0.966	0.963	0.964
	Distil BERT	0.960	0.963	0.957	0.959
	BERT	0.835	0.850	0.822	0.828
Logistic Regression	RoBERTa Large	0.973	0.972	0.971	0.972
	DistilRoBERTa	0.973	0.975	0.971	0.973
	T5 Base	0.973	0.973	0.972	0.972
	T5 Large	0.982	0.982	0.982	0.982
	T5 XL	0.978	0.979	0.978	0.978
	MPNet V1	0.969	0.968	0.968	0.967
	MPNet V2	0.969	0.970	0.968	0.968
	MPNet V3	0.978	0.979	0.976	0.977
	MPNet V4	0.973	0.974	0.973	0.973
	Distil BERT	0.978	0.978	0.976	0.977
	BERT	0.915	0.912	0.913	0.912
Voting Classifier	RoBERTa Large	0.973	0.972	0.973	0.972
	DistilRoBERTa	0.964	0.966	0.963	0.964
	T5 Base	0.969	0.967	0.968	0.967
	T5 Large	0.973	0.972	0.973	0.972
	T5 XL	0.978	0.979	0.978	0.978
	MPNet V1	**0.987**	0.988	0.985	**0.986**
	MPNet V2	0.969	0.968	0.968	0.967
	MPNet V3	0.982	0.981	0.984	0.982
	MPNet V4	0.978	0.979	0.977	0.977
	Distil BERT	0.982	0.982	0.980	0.981
	BERT	0.929	0.926	0.925	0.925

Furthermore, the T5 big embedding had the highest F1 score (98.2%) when evaluating the individual performance of Softmax regression. T5 XL and Distil Bert were very close behind in second place, separated by barely 0.1% in terms of F1 scores. This provides support for the idea that these particular transformer models may be useful for making trustworthy predictions when used in Softmax regression. In addition, after analyzing the results of each vote classifier, it was clear that MPNet variation v1 was the most successful, with a remarkable F1 score of 98.6%. With only 0.5% separating their F1 scores, MPNet and

Distil Bert tied for second place in this scenario. These findings demonstrate the usefulness of MPNet version v1 in producing extremely accurate and comprehensive predictions when used in the voting classifier. The result of all experiments on the AG news dataset can be found in Table 4. Further the confusion matrices of all experiments can be seen in Appendix A.1.

Table 4. Performance of experiments on AG News dataset.

Model	Transformer Sentence Embedding	Accuracy	Precision	Recall	F1
Random Forest	RoBERTa Large	0.892	0.892	0.892	0.892
	DistilRoBERTa	0.885	0.885	0.885	0.885
	T5 Base	0.876	0.875	0.875	0.875
	T5 Large	0.876	0.875	0.875	0.875
	T5 XL	0.876	0.875	0.875	0.875
	MPNET V1	0.891	0.891	0.891	0.891
	MPNET V2	0.887	0.887	0.887	0.887
	MPNET V3	0.891	0.891	0.891	0.891
	MPNET V4	0.888	0.888	0.888	0.888
	Distil BERT	0.876	0.875	0.875	0.875
	BERT	0.876	0.875	0.875	0.875
Logistic Regression	RoBERTa Large	0.908	0.908	0.908	0.908
	DistilRoBERTa	0.904	0.904	0.904	0.904
	T5 Base	0.897	0.897	0.897	0.897
	T5 Large	0.897	0.897	0.897	0.897
	T5 XL	0.897	0.897	0.897	0.897
	MPNET V1	0.905	0.905	0.905	0.905
	MPNET V2	0.902	0.902	0.902	0.902
	MPNET V3	0.905	0.905	0.905	0.905
	MPNET V4	0.902	0.902	0.902	0.902
	Distil BERT	0.897	0.897	0.897	0.897
	BERT	0.897	0.897	0.897	0.897
Voting Classifier	RoBERTa Large	**0.911**	0.911	0.911	0.911
	DistilRoBERTa	0.906	0.906	0.906	0.906
	T5 Base	0.907	0.907	0.907	0.907
	T5 Large	0.907	0.907	0.907	0.907
	T5 XL	0.907	0.907	0.907	0.907
	MPNET V1	0.908	0.908	0.908	0.908
	MPNET V2	0.906	0.906	0.906	0.906
	MPNET V3	0.907	0.907	0.907	0.907
	MPNET V4	0.906	0.906	0.906	0.906
	Distil BERT	0.907	0.907	0.907	0.907
	BERT	0.907	0.907	0.907	0.907

When the study was extended to include the AG dataset, it was discovered that RoBERTa Large's phrase embedding performed best for Random Forest with an F1 score of 89.2%. With an F1 score of 89.1%, only 0.1% separated MPNet v1 from first place. RoBERTa Large's embedding achieved best in Softmax regression, with an F1 score of 90.8%, leaving MPNet v1 in the dust by a mere 0.3%. Comparing voting classifier results, MPNet version v1 came out on top with a score of 91.1% on the F1 test, while MPNet came in second with a score of 90.8%, which was a difference of only 0.3%. In conclusion, the results of these studies show that when utilized to calculate sentence embeddings within the Random Forest framework, some transformer models, such as MPNet v1, MPNet v3, and DistilRoBERTa, perform exceptionally well. As an added bonus, both T5 big and big RoBERTa perform admirably in Softmax regression and Random Forest. These results are consistent with one another, highlighting the benefits of these transformer models in different settings and shedding light on their unique performances and their potential to enhance machine learning tasks.

Based on all experiments, the voting classifier with MPNet v1 embedding obtained the best F1 score of 98.6% for classifying BBC News into news categories. In addition, the voting classifier with RoBERTa Large embedded received the highest F1 score of 91.1% for classifying AG News into the news group. These results show that when the voting predictor was used with the Softmax and Random Forest algorithms and transformer

embeddings, it always did better than other baselines and individual algorithms. These results give strong support to the idea that the voting classifier, which uses transformer embeddings and ensemble learning techniques, produces much better results than other methods. In particular, the combination of Softmax and Random Forest with transformer embeddings showed that it could make very accurate and reliable forecasts for news category classification tasks.

In a larger sense, it can be said that the MPNet transformer model always does better than other models when it comes to classifying news by category. Its high F1 scores show that it is a top-performing model for this job. In short, the tests showed that the voting classifier with MPNet v1 embedding was better than other methods for classifying BBC News categories, with an impressive F1 score of 98.6%. In the same way, the voting classifier with RoBERTa Large embedded did a great job of classifying AG News categories, obtaining the best F1 score of 91.1%. These results show that the vote classifier with transformer embeddings works well and show how much better MPNet is at classifying news categories.

6. Conclusions

In conclusion, this study reveals the effectiveness of transformer sentence embeddings and soft voting of Softmax and Random Forest algorithms for new catgory classification tasks. Through rigorous analysis and comparison of transformer models, performance measures were understood. The macro-average metric assessed the accuracy, precision, recall, and F1 score in the studies.

The T5 large embedding had the highest F1 score of 98.2% for Softmax regression alone. T5 XL and Distil Bert had F1 scores 0.1% apart. MPNet variant v1 won the voting classifier news category prediction with a 98.6% F1 score, and MPNet and Distil Bert finished second in F1 scoring by 0.5%. These results demonstrate MPNet version v1's impressive prediction power when used with the voting classifier of Softmax regression and Random Forest. The voting classifier with MPNet v1 embedding had the greatest F1 score of 98.6% for BBC News category classification, while the voting classifier with RoBERTa Large embedding had 91.1% for AG News category classification. This shows that Softmax and Random Forest algorithms with transformer embeddings outperform baselines and individual methods. These results show that the voting classifier, using transformer embeddings and ensemble learning, can increase news category classification accuracy. MPNet models consistently beat other transformer models in news category classification, as their strong F1 scores prove they are the top models for this assignment. These findings demonstrate the voting classifier's transformer embeddings and MPNet models' news category classification supremacy.

7. Future Work

Future work in this area can focus on several factors. Firstly, exploring additional transformer models, such as GPT, Transformer-XL, or ELECTRA, would provide a more comprehensive understanding of their performance in conjunction with soft voting of Softmax and Random Forest algorithms. It would be valuable to assess the generalization capabilities of the proposed approach by testing it on external datasets beyond the BBC and AG News datasets used in the current study. This would help validate the effectiveness of the transformer embeddings and voting classifier across a wider range of news datasets. Evaluating the performance on diverse datasets can provide insights into the robustness and reliability of the proposed approach in different news classification scenarios. In short, future research should expand the investigation by considering different transformer architectures, fine-tuning hyperparameters, assessing generalization on external datasets, exploring alternative ensemble techniques, and analyzing interpretability and explainability. These research directions might contribute to a deeper understanding of the performance, robustness, and interpretability of the proposed approach, ultimately advancing news category classification using transformer embeddings and ensemble learning techniques.

The following are some potential limitations of the news category prediction study. The BBC and AG news datasets from Kaggle are utilized for this research. Although these datasets may be adequate for preliminary testing, their limited scope may hinder the generalizability of the results. The findings may not necessarily apply to additional diverse and extensive news datasets. The focus of this study is on four transformer models (BERT, RoBERTa, MPNet, and T5) and their variants for generating sentence embeddings. However, numerous other transformer architectures, including GPT, Transformer-XL, and ELECTRA, have not been investigated. The limited variety of transformer models may impede a thorough comprehension of their performance when combined with ensemble techniques.

Author Contributions: Conceptualization, S.K. and A.M.; methodology, S.K. and M.R.; investigation, S.K. and A.M.; experimentation, S.K. and M.R.; validation, M.R. and A.M.; formal analysis, M.R. and S.K.; resources, A.M.; data curation, M.R. and S.K.; writing—original draft preparation, S.K.; writing—review and editing, M.R. and A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All data were presented in the main text.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1

This appendix section contains the confusion matrices of a voting classifier regarding different experiments with different transformer sentence embedding.

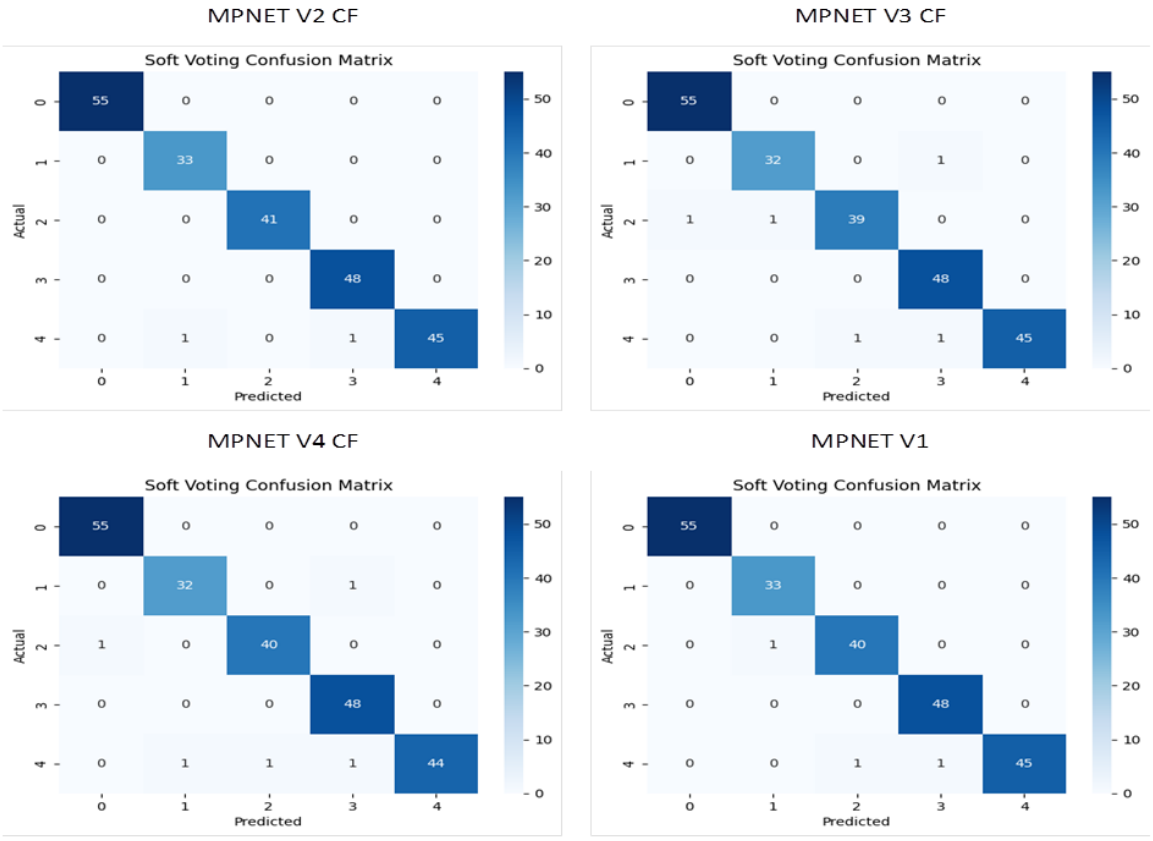

Figure A1. Voting classifier confusion matrices with multiple MPNet versions on BBC News dataset.

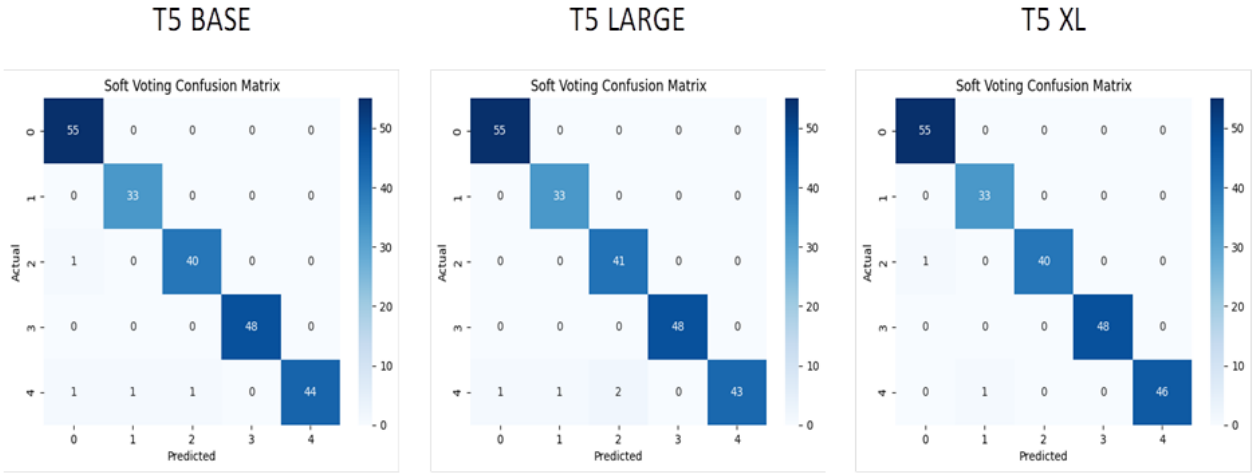

Figure A2. Voting classifier confusion matrices with multiple T5 versions on BBC News dataset.

Figure A3. Voting classifier confusion matrices with multiple BERT and RoBERTa versions on BBC News dataset.

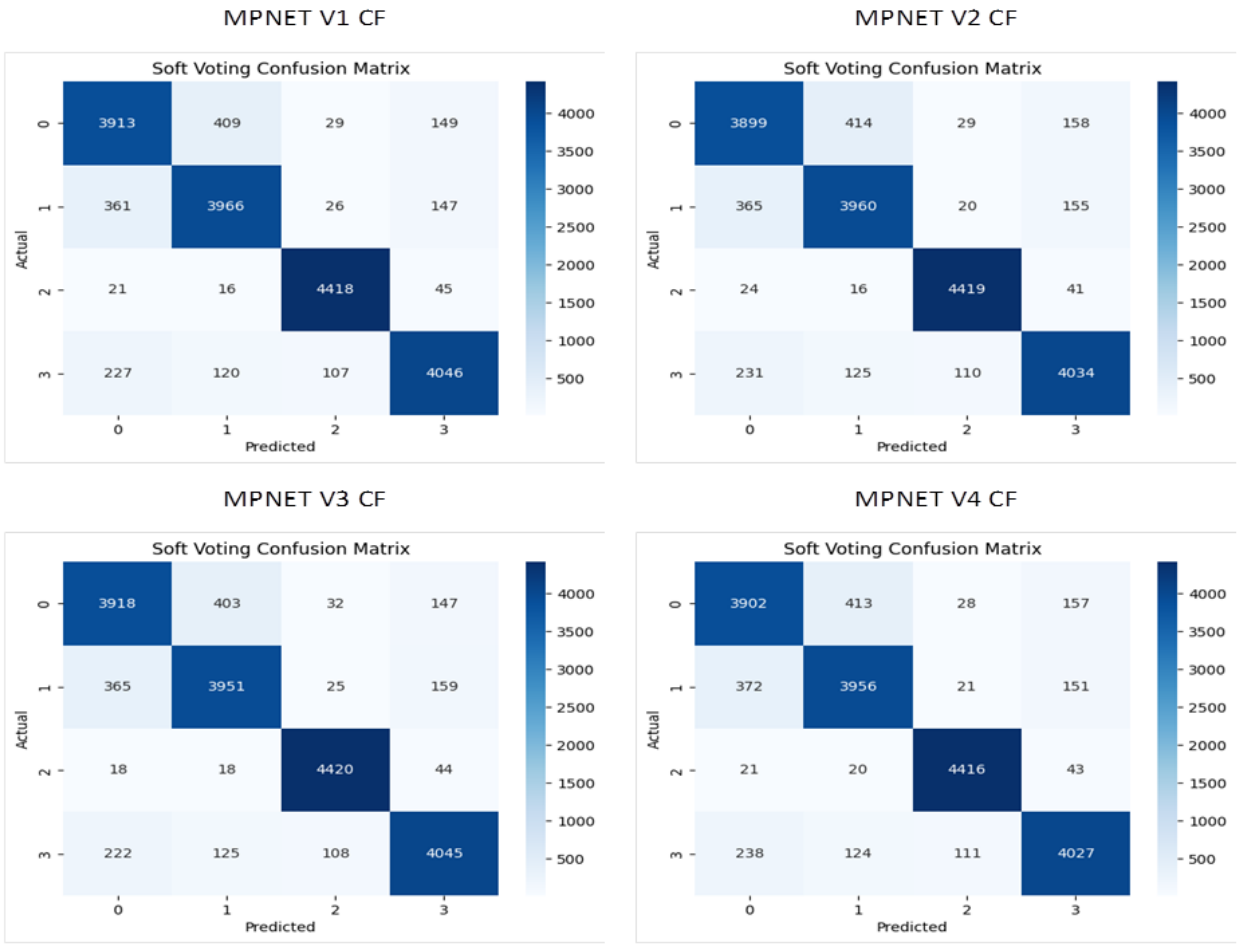

Figure A4. Voting classifier confusion matrices with multiple MPNet versions on AG News dataset.

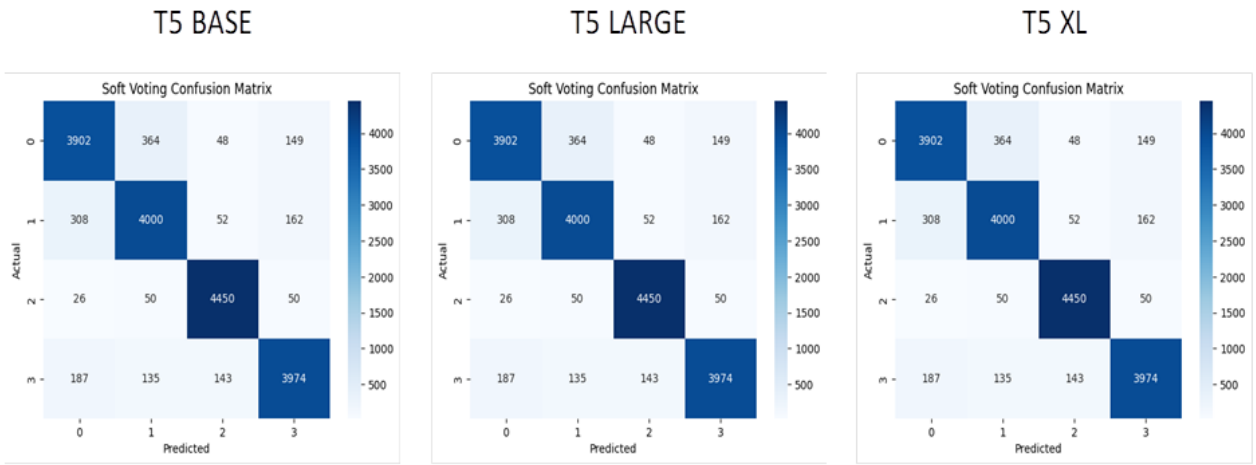

Figure A5. Voting classifier confusion matrices with multiple T5 versions on AG News dataset.

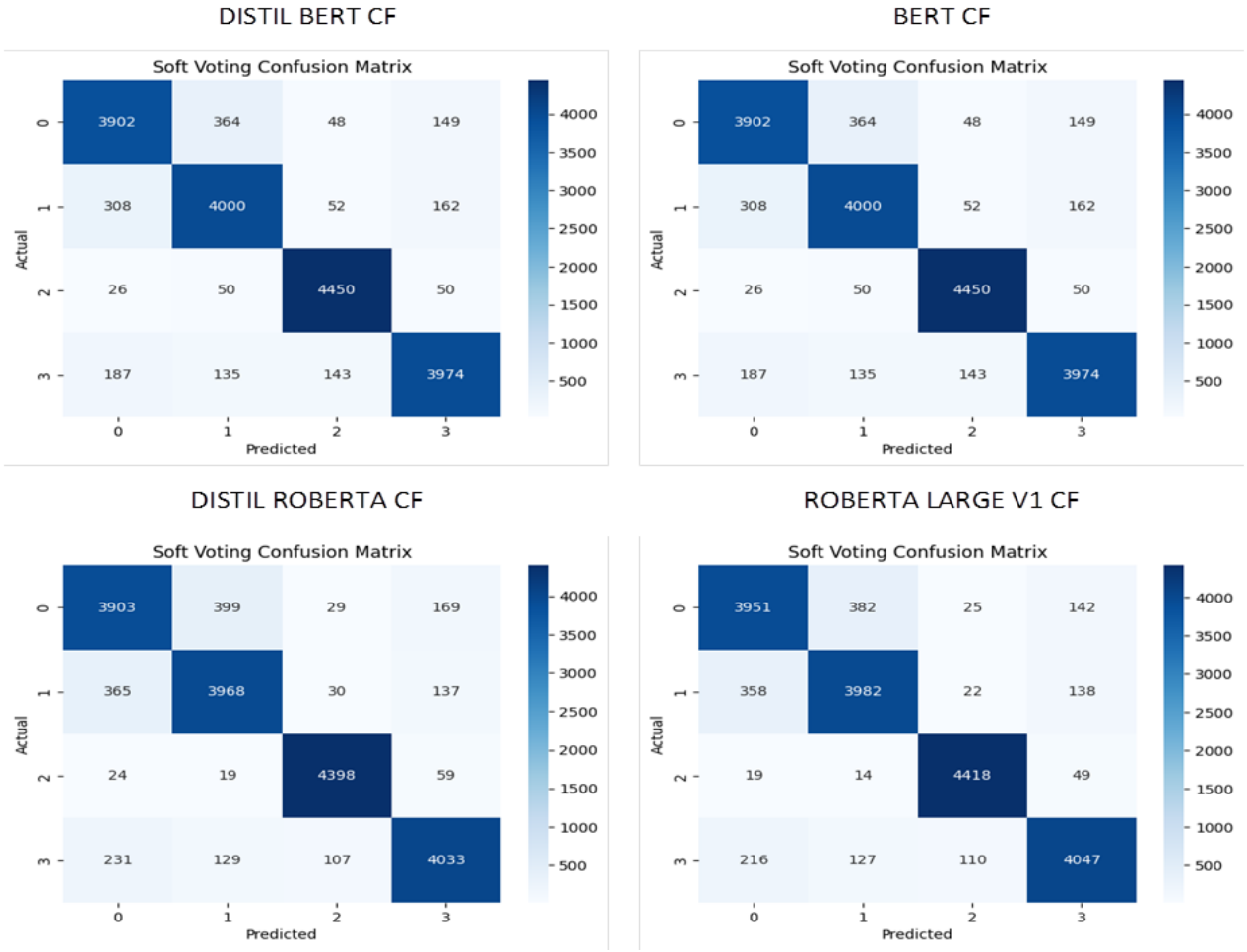

Figure A6. Voting classifier confusion matrices with multiple BERT and RoBERTa versions on AG News dataset.

References

1. Karaman, Y.; Akdeniz, F.; Savaş, B.K.; Becerikli, Y. A Comparative Analysis of SVM, LSTM and CNN-RNN Models for the BBC News Classification. In Proceedings of the 7th International Conference on Smart City Applications, Castelo Branco, Portugal, 19–21 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 473–483.
2. Gupta, A.; Chugh, D.; Anjum; Katarya, R. Automated news summarization using transformers. In *Sustainable Advanced Computing: Select Proceedings of ICSAC 2021*; Springer: Singapore, 2022; pp. 249–259.
3. Ding, H.; Yang, J.; Deng, Y.; Zhang, H.; Roth, D. Towards open-domain topic classification. *arXiv* **2023**, arXiv:2306.17290.
4. Nawaz, S.; Rizwan, M.; Rafiq, M. Recommendation of effectiveness of YouTube video contents by qualitative sentiment analysis of its comments and replies. *Pak. J. Sci.* **2019**, *71*, 91.
5. Choi, S.; Lee, H.; Park, E.; Choi, S. Deep learning for patent landscaping using transformer and graph embedding. *Technol. Forecast. Soc. Chang.* **2022**, *175*, 121413. [CrossRef]
6. Mars, M. From word embeddings to pre-trained language models: A state-of-the-art walkthrough. *Appl. Sci.* **2022**, *12*, 8805. [CrossRef]
7. Saigal, P.; Khanna, V. Multi-category news classification using Support Vector Machine based classifiers. *SN Appl. Sci.* **2020**, *2*, 458. [CrossRef]
8. Ugwuoke, U.C.; Aminu, E.F.; Ekundayo, A. *Performing Data Augmentation Experiment to Enhance Model Accuracy: A Case Study of BBC News' Data*; Elsevier: Amsterdam, The Netherlands, 2022.
9. Alfonse, M.; Gawich, M. A novel methodology for Arabic news classification. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2022**, *12*, e1440. [CrossRef]
10. Einea, O.; Elnagar, A.; Al Debsi, R. Sanad: Single-label arabic news articles dataset for automatic text categorization. *Data Brief* **2019**, *25*, 104076. [CrossRef]
11. Al-Laith, A.; Shahbaz, M. Tracking sentiment towards news entities from Arabic news on social media. *Future Gener. Comput. Syst.* **2021**, *118*, 467–484. [CrossRef]

12. Jang, B.; Kim, I.; Kim, J.W. Word2vec convolutional neural networks for classification of news articles and tweets. *PLoS ONE* **2019**, *14*, e0220976. [CrossRef]
13. Zhao, W.; Zhu, L.; Wang, M.; Zhang, X.; Zhang, J. WTL-CNN: A news text classification method of convolutional neural network based on weighted word embedding. *Connect. Sci.* **2022**, *34*, 2291–2312. [CrossRef]
14. Deng, L.; Ge, Q.; Zhang, J.; Li, Z.; Yu, Z.; Yin, T.; Zhu, H. News Text Classification Method Based on the GRU_CNN Model. *Int. Trans. Electr. Energy Syst.* **2022**, *2022*, 1197534. [CrossRef]
15. Liu, K.F.; Zhang, Y.; Zhang, Q.X.; Wang, Y.G.; Gao, K.L. Chinese News Text Classification and Its Application Based on Combined-Convolutional Neural Network. *J. Comput.* **2022**, *33*, 1–14.
16. Endalie, D.; Haile, G. Automated Amharic news categorization using deep learning models. *Comput. Intell. Neurosci.* **2021**, *2021*, 3774607. [CrossRef] [PubMed]
17. Ahmad, I.; AlQurashi, F.; Mehmood, R. Machine and Deep Learning Methods with Manual and Automatic Labelling for News Classification in Bangla Language. *arXiv* **2022**, arXiv:2210.10903.
18. Majeed, F.; Asif, M.W.; Hassan, M.A.; Abbas, S.A.; Lali, M.I. Social media news classification in healthcare communication. *J. Med. Imaging Health Inform.* **2019**, *9*, 1215–1223. [CrossRef]
19. Vinh, T.N.P.; Kha, H.H. Vietnamese News Articles Classification Using Neural Networks. *J. Adv. Inf. Technol. (JAIT)* **2021**, *12*, 363–369. [CrossRef]
20. Rahman, M.M.; Khan, M.A.Z.; Biswas, A.A. Bangla news classification using graph convolutional networks. In Proceedings of the 2021 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 27–29 January 2021; pp. 1–5.
21. Chowdhury, P.; Eumi, E.M.; Sarkar, O.; Ahamed, M.F. Bangla news classification using GloVe vectorization, LSTM, and CNN. In Proceedings of the International Conference on Big Data, IoT, and Machine Learning: BIM 2021, Cox's Bazar, Bangladesh, 23–25 September 2021; Springer: Singapore, 2022; pp. 723–731.
22. Amin, R.; Sworna, N.S.; Hossain, N. Multiclass classification for bangla news tags with parallel cnn using word level data augmentation. In Proceedings of the 2020 IEEE Region 10 Symposium (TENSYMP), Dhaka, Bangladesh, 5–7 June 2020; pp. 174–177.
23. Rizwan, M.; Mushtaq, M.F.; Akram, U.; Mehmood, A.; Ashraf, I.; Sahelices, B. Depression Classification From Tweets Using Small Deep Transfer Learning Language Models. *IEEE Access* **2022**, *10*, 129176–129189. [CrossRef]
24. Chandana, N.; Sreelekha, A.; Rasi, K.; Sreeja, J.; Prassanna, P.L. BCC NEWS Classification Comparison between Naïve Bayes, Support Vector Machine, Recurrent Neural Network. In Proceedings of the 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 4–6 February 2021; pp. 833–835.
25. Sun, N.; Du, C. News text classification method and simulation based on the hybrid deep learning model. *Complexity* **2021**, *2021*, 8064579. [CrossRef]
26. Nergız, G.; Safali, Y.; Avaroğlu, E.; Erdoğan, S. Classification of Turkish news content by deep learning based LSTM using Fasttext model. In Proceedings of the 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Turkey, 21–22 September 2019; pp. 1–6.
27. Dogru, H.B.; Tilki, S.; Jamil, A.; Hameed, A.A. Deep learning-based classification of news texts using doc2vec model. In Proceedings of the 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), Riyadh, Saudi Arabia, 6–7 April 2021; pp. 91–96.
28. Zhu, Y. Research on news text classification based on deep learning convolutional neural network. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 1508150. [CrossRef]
29. Ahmed, F.; Akther, N.; Hasan, M.; Chowdhury, K.; Mukta, M.S.H. Word embedding based news classification by using CNN. In Proceedings of the 2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM), Pekan, Malaysia, 24–26 August 2021; pp. 609–613.
30. Li, B.; Zhou, H.; He, J.; Wang, M.; Yang, Y.; Li, L. On the sentence embeddings from pre-trained language models. *arXiv* **2020**, arXiv:2011.05864.
31. Huang, J.; Tang, D.; Zhong, W.; Lu, S.; Shou, L.; Gong, M.; Jiang, D.; Duan, N. Whiteningbert: An easy unsupervised sentence embedding approach. *arXiv* **2021**, arXiv:2104.01767.
32. Jiang, T.; Jiao, J.; Huang, S.; Zhang, Z.; Wang, D.; Zhuang, F.; Wei, F.; Huang, H.; Deng, D.; Zhang, Q. Promptbert: Improving bert sentence embeddings with prompts. *arXiv* **2022**, arXiv:2201.04337.
33. Sentence-Transformers/All-Mpnet-Base-v1·Hugging Face—Huggingface.co. Available online: https://huggingface.co/sentence-transformers/all-mpnet-base-v1 (accessed on 24 May 2023).
34. Sentence-Transformers/All-Mpnet-Base-v2·Hugging Face—Huggingface.co. Available online: https://huggingface.co/sentence-transformers/all-mpnet-base-v2 (accessed on 24 May 2023).
35. Flax-Sentence-Embeddings/All-Datasets-v3-Mpnet-Base·Hugging Face—Huggingface.co. Available online: https://huggingface.co/flax-sentence-embeddings/all_datasets_v3_mpnet-base (accessed on 24 May 2023).
36. Flax-Sentence-Embeddings/All-Datasets-v4-Mpnet-Base·Hugging Face—Huggingface.co. Available online: https://huggingface.co/flax-sentence-embeddings/all_datasets_v4_mpnet-base (accessed on 24 May 2023).
37. Sentence-Transformers/Msmarco-Distilbert-Base-Tas-b·Hugging Face—Huggingface.co. Available online: https://huggingface.co/sentence-transformers/msmarco-distilbert-base-tas-b (accessed on 24 May 2023).
38. Bongsoo/Moco-SentencebertV2.0·Hugging Face—Huggingface.co. Available online: https://huggingface.co/bongsoo/moco-sentencebertV2.0 (accessed on 24 May 2023).

39. Briskilal, J.; Subalalitha, C. An ensemble model for classifying idioms and literal texts using BERT and RoBERTa. *Inf. Process. Manag.* **2022**, *59*, 102756. [CrossRef]
40. Sentence-Transformers/All-Roberta-Large-v1·Hugging Face—Huggingface.co. Available online: https://huggingface.co/sentence-transformers/all-roberta-large-v1 (accessed on 24 May 2023).
41. Sentence-Transformers/All-Distilroberta-v1·Hugging Face—Huggingface.co. Available online: https://huggingface.co/sentence-transformers/all-distilroberta-v1 (accessed on 24 May 2023).
42. Sentence-Transformers/gtr-t5-Base·Hugging Face—Huggingface.co. Available online: https://huggingface.co/sentence-transformers/gtr-t5-base (accessed on 24 May 2023).
43. Sentence-Transformers/gtr-t5-large·Hugging Face—Huggingface.co. Available online: https://huggingface.co/sentence-transformers/gtr-t5-large (accessed on 24 May 2023).
44. Sentence-Transformers/gtr-t5-xl·Hugging Face—Huggingface.co. Available online: https://huggingface.co/sentence-transformers/gtr-t5-xl (accessed on 24 May 2023).
45. Abhishek, K. News Article Classification using a Transfer Learning Approach. In Proceedings of the 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 13–14 October 2022; pp. 1–6.
46. Patil, S.; Lokesha, V.; Anuradha, S.G. Multi-Label News Category Text Classification. *J. Algebr. Stat.* **2022**, *13*, 5485–5498.
47. Ali, H.; Khan, M.S.; Al-Fuqaha, A.; Qadir, J. Tamp-X: Attacking explainable natural language classifiers through tampered activations. *Comput. Secur.* **2022**, *120*, 102791. [CrossRef]