RESEARCH ARTICLE

EMPOLITICON: NLP and ML Based Approach for Context and Emotion Classification of Political Speeches From Transcripts

AZHER AHMED EFAT, **ASIF ATIQ**, **ABRAR SHAHRIAR ABEED**, **ARMANUL MOMIN**,
AND MD. GOLAM RABIUL ALAM, (Member, IEEE)
Department of Computer Science and Engineering, BRAC University, Dhaka 1212, Bangladesh

Corresponding author: Azher Ahmed Efat (azher.ahmed.efat@g.bracu.ac.bd)

ABSTRACT Political speeches have played one of the most influential roles in shaping the world. Speeches of the written variety have been etched into history. These sorts of speeches have a great effect on the general people and their actions in the coming few days. Moreover, if left unchecked, political personnel or parties may cause major problems. In many cases, there may be a warning sign that the government needs to change its policies and also listen to the people. Understanding the emotion and context of a political speech is important, as they can be early indicators or warning signs for impending international crises, alignments, wars and future conflicts. In our research, we have focused on the presidents/prime ministers of China, Russia, the United Kingdom and the United States which are the permanent members of the United Nations Security Council and classified the speeches given by them based on the context and emotion of the speeches. The speeches were categorized into optimism, neutral, joy or upset in terms of emotion and five context categories, which are international affairs, nationalism, development, extremism and others. Here, optimism is a secondary emotion, whereas joy and upset are primary emotions. Apart from classifying the speeches based on context and emotion, one of the major works of our research is that we are introducing a dataset of political speeches that contains 2010 speeches labelled with emotion and context of the speech. The speeches we have worked on are large in word count. We propose EMPOLITICON-Context, a soft voting classifier ensemble learning model for context classification and EMPOLITICON-Emotion, a soft voting classifier ensemble learning model for emotion classification of political speeches. The proposed EMPOLITICON-Context model has achieved 73.13% accuracy in terms of context classification and the EMPOLITICON-Emotion model has achieved 53.07% accuracy in classifying the emotion of the political speeches.

INDEX TERMS Political speeches, emotion classification, context classification, ensemble learning, SMOTEN, longformer, EMPOLITICON.

I. INTRODUCTION

Politics is an intrinsic part of our society. It is a competition that goes on between people, which they usually do in groups with the goal of shaping decisions in their favor. There are some speeches that changed the world like the Quit India speech given by Mahatma Gandhi, which helped India gain

independence, ''I Have a Dream'' by Martin Luther King, Jr. and the 7th March speech of Sheikh Mujibur Rahman, which played a vital role in the independence of Bangladesh. In most cases, political speeches are controversial. The aftereffects of such speeches or opinions may be felt years after they were made. Politics and political speeches influence people's life directly or indirectly. From the economy to social or personal life, everything is influenced by politics. The person that we select as our leader leads the country and has a major impact

The associate editor coordinating the review of this manuscript and approving it for publication was Alberto Cano .

on the country's progress and on our lives. It is critical to know our leaders so that we can make informed decisions. Thus, it is important to understand what the country's leaders talk about. While looking at the previous works, we found that there are very few works in this domain. The works that we have seen are mostly based on Twitter data and usually related to elections. However, we have not seen any work that solely focused on the speeches given by the country leaders and tried to understand the context and emotion of the speeches. Which made us think that analyzing the speeches to understand what emotion they showed and what context they were made in is an intriguing field to conduct research on.

In this work, we classified the political speeches into optimism, neutral, joy and upset emotions and classified the political speeches on the basis of context of the speeches into development, nationalism, extremism, international affairs and others. The key contributions of this research are summarized as follows:

- We classified political speeches based on their context and emotion. To the best of our knowledge, there has been no previous work focusing on classifying the context and emotion of political speeches.
- There is no dataset containing the political speeches of the country leaders of the United Nations Security Council's permanent members. For the purpose of this research, we have created our own dataset. Through this research, we are introducing the dataset containing 2010 political speeches of the country leaders of China, Russia, United Kingdom and United States which are permanent members of the United Nations Security Council. The speeches are labelled with context and emotion.
- The works that we have seen related to political persons are usually done focusing on short tweets, whereas our work focused on long documents. In this research, we have used longformer for processing the long documents for emotion and context classification.
- We have developed two soft voting classifier ensemble learning models to classify political speeches based on their context and emotion. EMPOLITICON-Context has achieved 73.13% accuracy in terms of context classification of political speeches and an accuracy of 53.07% in classifying the emotion of political speeches has been achieved by EMPOLITICON-Emotion.

The rest of the paper is organized as follows, section II talks about the existing works. Section III describes the proposed methodology for the classification of emotion and context of political speeches. In section IV, the result analysis of the models is given, followed by section V, which presents a discussion of the results. Finally, section VI concludes the paper.

II. EXISTING WORKS
In this section, we discuss the existing works that has been done in the political field. We have also studied works related

to long document processing, which are going to be discussed in this section.

The paper [14] describes the combination of both LDA and SVM properties to make a new classifier with a huge margin that can be implemented with the current SVM model with an accuracy of SVM/LDA = 98.3%, SVM = 95.8%, LDA = 97.4% on the Twonorm dataset. On the downside, only linear decision boundaries can be implemented using the SVM/LDA formulation proposed in this study. The paper [6] presents the BoostingBERT model, which incorporates multi-class boosting into BERT. Their results demonstrate that BoostingBERT consistently beats bagging BERT. The findings of the experiment reveal that BoostingBERT beats the strong BERT baseline on all tests and that it is beneficial in a variety of NLP tasks. The paper [7] uses sentiment analysis and defines six Twitter user classes during the 2016 US presidential election to evaluate political homophily. The researchers discovered that negative users have the highest amount of homophily and form the most homogeneous societies. In the paper [9], the authors show that political party representatives can be characterized using data sciences along with their evolution with frequency analysis, which helped in extracting each party's distinct ideological bubble and then moved on to a more sophisticated analysis. It had a precision range of 71-75% and was also able to find left or right political alignment with a precision of 90%.

In the study [8], in order to predict election results, the researchers proposed a method to infer citizen's political alignment. They built a dictionary that characterizes a political leader's speech. Using the dictionary, they trained a SVM that can predict the user's political alignment with an accuracy of 87%. They have also designed a metric to quantify the political lining of a tweet. Tweets categorized with the metric reflected the election outcome with 98.72% accuracy. The research work [5] examines how the appearance of political hashtags on social media can influence people's reactions through a randomized control experiment. Finally, they came to know that when a hashtag is used, the particular news gets more attention, and that news becomes more partisan and controversial. The research work [2] proposes IAD sentiment analysis by using word embedding which is basically a feature extracting model by using four supervised learning methods. They said that negative samples LR and SVM classifiers give slightly better performance and vice versa relation between DT and NB. In the paper, [11] the authors focused on the sentiment that users showed toward the Facebook posts made at the time of the election campaign. They collected and analyzed 4128 Facebook posts and the reactions to these posts. The sentiment index they got for the political parties was 1.08 for PRI, 2.72 for MORENA, 3.46 for PRD and 1.06 for PAN. The authors [10] conducted polarity along with subjectivity sentiment analysis on tweets by taking time as the basis for the dimension of SA. This is done using different Natural Language Processing (NLP) techniques. Results obtained from TextBlob are 2447 (32.93%) Positive, 3971 (53.44%) neutral and 1012

(13.62%) negative respectively. Again, results obtained from SentiWordNet were 2916 (39.25%) Positive, 3085 (41.52%) neutral and 1429 (19.23%) negative respectively. In the paper [13], the authors incorporated sentiment analysis techniques in the domain of political news from columns of Turkish news sites. They have used SVM, Naïve Bayes, Maximum Entropy and character based N-gram language model for sentiment classification and compared their results. In all the approaches, they achieved accuracy ranged from 65% to 77%. In the paper [15], the authors built an emotion embedding model using CNN. They have worked with 8 emotions. Based on recall, the classification results show that Joy (22%), Sadness (19.88%), Fear (16.4%) and Anger (14.4%) are the best 4 emotions. In the paper [16], the authors proposed emotion classification in dialogue based on the semi-supervised word-level emotion embedding.

The paper [12] portraits a primitive approach on sentiment analysis on tweets from twitter. The data preprocessing takes advantage of two new different sets of dictionaries: the emoji dictionary and the acronym dictionary. The paper discusses the design of the kernel tree. In the paper, [4] the authors used hybrid n-gram models to counter the 'Zero Count problem'. They have applied laplace smoothing for the purpose of creating a classifier to achieve higher accuracy. Their research result also relates to the result of the 2008 US presidential election. A sentiment Tree Bank [1] is being made from scratch which is a required resource for the RNN model to be made. Ambiguity was overcome by in-context morphological analysis and disambiguation and other relevant machine learning.

One of the problems with BERT and other transformer-based models is that they can not process sequences longer than 512 tokens due to their self-attention operation which scales quadratically with the sequence length. This problem is addressed in the paper [17] and the authors introduced a new model Longformer. The attention mechanism of Longformer scales linearly unlike of the other transformer models with the sequence length. The attention mechanism combines a local window attention and a global attention. Longformer is able to process sequences up to 4096 tokens which makes it easier to work with long documents.

BERT and the variants of BERT models suffer when the length of the context is high. In the paper [20], the issues of context length and the model's size have been addressed. They have applied the sliding window concept and increased the maximum context length. After finetuning, they were able to achieve 0.881 accuracy, which was higher than the BERT base models. In the research [3], the authors incorporated recurrent attention learning framework in order to classify long documents. They trained a recurrent neural network based controller that can focus its attention on the discriminative parts and extracted glimpsed features by a short text level convolutional neural network (CNN) from the focused group of words. The controller uses the context information consisting of the coarse representation of the original
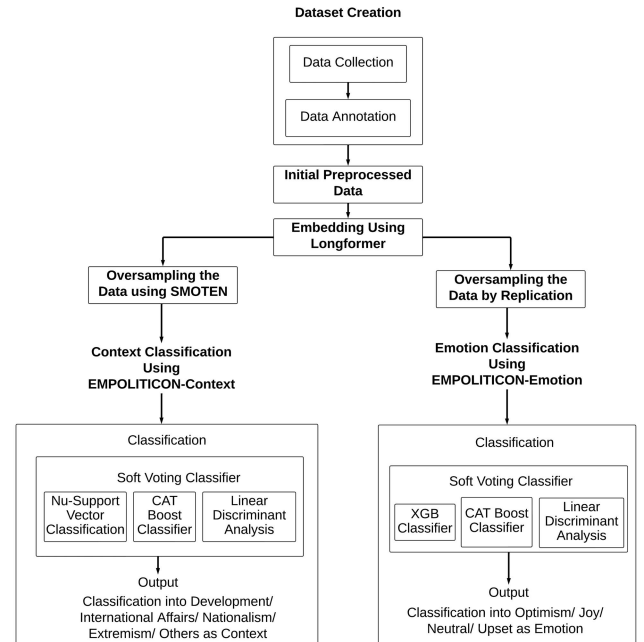

FIGURE 1. Top Level Overview of the Workflow.

document and the memorized glimpsed features to locate its attention. In the paper [21], a text truncation method has been proposed, in which the original text length is reduced to a predefined limit to improve performance while keeping the computational costs low.

III. PROPOSED METHODOLOGY FOR CLASSIFICATION OF EMOTION AND CONTEXT OF POLITICAL SPEECHES

The classification of political speeches into context and emotion work is divided into five distinct phases. (1) dataset creation, (2) initial data preprocessing, (3) generating embedding using Longformer, (4) oversampling the data, (5) classification of the political speeches. The dataset creation step is again divided into two steps. (1) data collection, (2) data annotation. For classification, we have built two soft voting classifier ensemble models called EMPOLITICON-Context and EMPOLITICON-Emotion. Figure 1 shows the top level overview of the research work.

A. DATA COLLECTION
We have built a custom dataset for the purpose of our research. We have built the dataset using the written transcripts of the speeches of China, Russia, United Kingdom and United States which are the permanent members of the United Nations Security Council. These countries play a vital role in global politics and decision making. As a result, any statements made by these countries are integral to understanding the global political dynamics. There are five permanent members of the United Nations Security Council out of which we have considered four of the countries. We did not consider

TABLE 1. Dataset Fields.

Field Name	Field Description
Country	Name of the country.
Date	Date on which the speech was given.
Speaker	Speaker of the speech.
Headline	Title of the transcript of the speech.
Text_of_Speech	The full transcript of the speech.
Designation	The designation of the Speaker.
Running President/PM	The President/PM of the country when the speech was given.
Speech Link	The link of the full transcript of the speech. The speech was collected from that link.
Emotion	The class assigned to that speech as emotion.
Context	The class assigned to that speech as context.

TABLE 2. Dataset Summary.

Country	Number Of Speeches	Year	Speaker	Main Website Link
Russia	557	2000-2021	Vladimir Putin, Dmitry Medvedev	http://en.kremlin.ru
US	989	2012-2021	Barack Obama, Donald Trump, Joe Biden	https://obamawhitehouse.archives.gov/ https://trumpwhitehouse.archives.gov/ https://www.whitehouse.gov
China	278	2003-2021	Wang Yi, Xi Jinping, Li Keqiang, Hu Jintao	https://www.fmprc.gov.cn
United Kingdom	186	2012-2019	Boris Johnson, Theresa May, David Cameron	https://www.gov.uk

France, which is the other permanent member of United Nations Security Council for our research. The reason for not considering France is that the speeches of France were available in French instead of English. In our dataset, we have collected a total of 2010 speeches. The speeches that we have collected were made by the presidents or prime ministers of the countries. However, in the case of China, we collected the speeches from the minister of foreign affairs as well as the prime minister and president, as there were not many speeches available from the president and prime minister. All the speeches were collected from the official websites of the government. While collecting the speeches, we went through each of the speeches before selecting the speech, and if there were multiple speakers in that speech, then the speech was rejected. We have only considered speeches that contain only one speaker so that there is no chance of multiple emotions or context from multiple people being present in the speech. In the dataset, each sample of speech has the speaker, date of speech, and link of the transcript associated with it so that the original source of the speech can be easily accessed. The fields that the dataset contains are given in Table 1.

To collect the speech transcripts of United States of America, Russia and China, we have gone through each of the government website and from there we have collected the speeches. However, the UK speeches were collected directly from the government website by using an automated web scraping program. The program collects the speeches from gov.uk which is an authentic source for the speeches of the prime minister of the UK. The program reads a list of the names of prime ministers and for each prime minister the speech collection process starts. At first, a search is made on the site with name of the prime minister with the appropriate date filters. Then it goes through the search results and filters out the ones that are not speeches. This was achieved by matching the url pattern: whether the url starts with ''government/speeches'' after the site domain. After this, each result was opened and the speech was extracted. After collecting the speeches of the UK using the scrapping program, we went through the speeches to check if there were any speeches with multiple speakers or from other speakers who were not the current prime minister. In the case of any speech found with multiple speakers or from other speakers who were not the

current prime minister, the speech was removed. We have chosen the speeches from various time ranges and from various speakers to make it diverse. For the USA, the speakers are Barack Obama, Donald Trump, Joe Biden and the selected time range is 2012-2021. For Russia, we have chosen the speeches of Vladimir Putin, Dmitry Medvedev from 2000-2021. The speakers from China are Wang Yi, Xi Jinping, Li Keqiang, Hu Jintao and the time range is 2003-2021. The speakers and time range of speeches of UK are Boris Johnson, Theresa May, David Cameron from 2012-2019. As we have collected these speeches based on our own judgement, there may be some biases as we have omitted some speeches that we felt were not relevant to our research. A summary of the collected speeches of our dataset is mentioned in the Table 2.

B. DATA ANNOTATION

After collecting the speeches, the next step was to label each speech. Each speech is labeled into two criteria. One is based on emotion and the other is based on context. For emotion there are four classes which are optimism, joy, neutral and upset. The joy and upset are primary emotions, whereas optimism is a secondary emotion. According to Plutchik [26], optimism is a primary-dyads emotion, formed by the combination of two primary emotions, joy and anticipation. On the other hand, the classes for context are development, international affairs, nationalism, extremism and others. For the annotation part, the speeches were given to three different annotators at first. The annotators read each individual speech and then labelled the speeches based on their interpretation of the speech. Once all the annotators labelled the speeches, the labels were compared and the label with the most votes for each individual speech was assigned to the speech. In case of a tie, the speeches were sent to further one person to review to get the final label. For data annotation validation, once all the speeches were labelled, all the annotators sat together and discussed whenever there was any conflict in labeling. Each individual speech was assigned with one emotion label and one context label by the annotators. One of the difficulties in labelling the speeches was that in different parts of the speeches, the speaker showed different emotions. Moreover, the speaker also talked on various contexts in different sub-texts of the speeches when the speeches were long. In that case, the emotion and the context that were most dominant

TABLE 3. Summary of Speech Label and Label Description.

Category	Class Name	Class Description	Speech Count
Context	Development	Focuses on infrastructure, economy, health and related context.	750
	International Affairs	Focuses on context related to international affairs.	518
	Nationalism	Focuses on nation's history, independence and related contexts.	139
	Extremism	Focuses on crime, violent and extremist activities and related context.	64
	Others	Focuses on context that does not fall under the other four context categories. For example: Social events, Election.	539
Emotion	Optimism	Optimism is known as a secondary emotion.	583
	Joy	Joy is known as a primary emotion.	525
	Neutral	No emotion was shown.	589
	Upset	Upset is known as a primary emotion.	313

were chosen in order to assign one emotion label and one context label for the full speech by the annotators. As we have done our labeling based on the annotator's interpretation, there might be some unwanted bias in our dataset. A summary of the labelled data and label description is given in the Table 3.

C. INITIAL DATA PREPROCESSING

Preprocessing is a very integral part of natural language processing task. In our work, the preprocessing has been done in two steps. Firstly, we have done some initial preprocessing of the political speeches. In the initial preprocessing step, preliminary tasks like removing stop words were carried out. Moreover, nouns, special characters, spaces, alphanumeric characters and slashes were also removed. We have also lemmatized the speeches and made all the letters lowercase. Additionally, numerical values were erased since the main content of the data set is speeches of large political figures from which we are trying to classify emotions and contexts where numerical values like dates are less likely to contribute any value. This also reduces the size of vector matrices and optimizes the feature extraction. As we are dealing with entire speeches, the token length of each speech is quite large. The highest token size of a speech after the initial preprocessing is 2736 tokens. While the average token size after initial preprocessing ranged from 600 to 1500 tokens.

D. GENERATING EMBEDDINGS USING LONGFORMER

For classifying the emotion and context of the political speeches, we needed contextual embedding of the speeches. As our speeches were long documents, we have used Longformer to generate the embeddings. The attention mechanism of longformer scales linearly with sequence length, which makes processing documents of thousands or longer tokens easy for it. Using multiple layers of attention, longformer builds contextual representations of the entire context,

reducing the need for task-specific architecture. The attention mechanism combines a global attention driven by end task and a windowed local-context self-attention. While the global attention is utilized to develop whole sequence representation for prediction, the local attention is used to build contextual representation. Longformer adds global attention to few pre-selected input locations and makes the attention operation symmetric. In longformer a token with global attention attends to all tokens across the sequence and all tokens in the sequence attend to it. Figure 2 [17] shows an example of a sliding window attention with global attention at a few tokens at custom locations. Longformer has various tokens, for example, <s> tokens for classification, question tokens for question answering tasks. As ours is a classification task, we set the global attention to the <s> and </s> tokens. We have used the pooler output of the model which is the last layer hidden-state of the first token of the sequence (classification token) further processed by a Linear layer and a Tanh activation function. For each document, the pooler output of the model is a tensor of dimension [1, 768]. After generating the embeddings, we split them into training and test data, where training data was 70% and test data was 30%.

E. OVERSAMPLING THE DATA

As the dataset was imbalanced, we oversampled the training data. Two different techniques were used to oversample the training data based on the classification task.

- Oversampling Using SMOTEN for Context Classification: As the dataset was imbalanced for context classification, in the next step, we balanced the training dataset using the SMOTEN oversampling technique. SMOTEN synthesizes new data from the minority class.
- Oversampling by Replication for Emotion Classification: As the dataset was imbalanced for emotion classification, we balanced the training dataset for emotion classification by oversampling the minority classes by replicating.

F. METHODS USED FOR EMPOLITICON-CONTEXT AND EMPOLITICON-EMOTION

For the classification of emotion and context of political speeches, we are proposing two soft voting classifiers. A voting classifier is an ensemble learning technique. Ensemble learning models use the decisions from different baseline models to improve the overall performance of ensemble models. A soft voting ensemble model predicts the class label based on the argmax of the sums of the predicted probabilities of the baseline models. For emotion classification, we built EMPOLITICON-Emotion which is a soft voting classifier with XGB Classifier, Cat Boost Classifier and Linear Discriminant Analysis as the baseline models. On the other hand, for context classification, we built EMPOLITICON-Context, a soft voting classifier with Nu-Support Vector Classification, Cat Boost Classifier and Linear Discriminant Analysis as the baseline models. At first, we discuss each of the models and
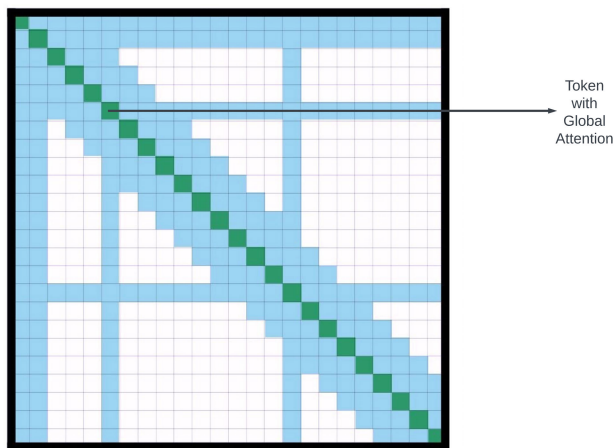
FIGURE 2. Example of Longformer's Sliding Window Attention With Global Attention.

later on we discuss the tuning of parameters that we have done to build our models.

- XGBoost Classifier: XGBoost [23] is a boosting ensemble learning model. In boosting ensemble learning, the model is built on a few weak learner models. The weak learner models are in sequential order, where each model tries to reduce the errors from the predecessor model. The boosting ensemble technique can be summarized in three simple steps.

1) To predict the target variable y, a baseline model F_0 is established. With this model, a residual will be connected $(y - F_0)$

2) The residuals from the previous model are fitted to a new model, h1.

3) Now F1, the enhanced form of F0, is created by combining F0 with h1. In comparison to F0, F1's mean squared error will be smaller.

$$F_1(x) < F_0(x) + h_1(x) \qquad (1)$$

Now, m number of iterations can be performed to minimize the residuals as much as possible.

$$F_m(x) < F_{m-1}(x) + h_m(x) \qquad (2)$$

XGBoost is an improvement of the gradient boosting algorithm that is designed to be highly efficient, flexible and portable. As XGBoost is a boosting algorithm, it combines a number of weak learners in order to form a strong learner. It trains a number of decision trees, each of which is trained on a portion of the data and whose predictions are then combined to provide the final predictions.

- CatBoost Classifier: Cat Boost classifier [22] is another boosting ensemble learning model based on gradient boosted decision tree. It builds a set of decision trees during training where each successive tree reduces the loss of the predecessor tree. One of the major points of the CatBoost model is that it is able to handle categorical

features. We have added the CatBoost classifier as a base model in our soft voting classifier.

- Linear Discriminant Analysis: The goal of a discriminant rule is to separate the data space into K distinct regions, each of which represents one class. With these regions, classification by discriminant analysis essentially implies that we assign x to class j if x is in region j. Linear Discriminant Analysis is a classifier with a linear decision boundary generated by fitting class conditional densities to the data and using Bayes' rule. Assuming that all classes have the same covariance matrix, the model fits a Gaussian density to each class.

$$\delta_k(x) = x^T \sum{}^{-1} \mu_k - \frac{1}{2} \mu_k^T \sum{}^{-1} \mu_k + log\pi_k \quad (3)$$

Assuming that the data comes from multivariate Gaussian distribution with the distribution of X can be characterized by its mean (μ) and covariance $(\sum)$, we can obtain the above discriminant function [18] which is linear. As a result, the decision boundary between any pair of classes is also a linear function in x.

- Nu-Support Vector Classifier: Support Vector Machine is an algorithm that plots each data item as a point in n-dimensional space with the value of each feature being the value of a particular coordinate. The goal is to find the optimal hyper-plane that differentiates the classes to perform classification. The Nu-Suppourt Vector Classifier [25] is similar to the SVC with the only difference that the Nu-SVC classifier has a nu parameter that can control the number of support vectors.

- Soft Voting Classifier: Voting classifiers combine the base models to come up with the final classification. We have built two soft voting classifiers: EMPOLITICON-Emotion and EMPOLITICON-Context. EMPOLITICON-Emotion is built with XGB Classifier, Cat Boost Classifier and Linear Discriminant Analysis as the baseline models. EMPOLITICON-Context has Nu-Support Vector Classification, Cat Boost Classifier and Linear Discriminant Analysis as the baseline models. In a soft voting ensemble classifier, the classifier combines the probabilities of each prediction in each model and picks the prediction with the highest total probability.

$$\hat{y} = argmax_i \sum_{j=1}^{m} w_j p_{ij} \qquad (4)$$

Here w_j is the weight assigned to the jth classifier and p is the prediction.

G. PROPOSED EMPOLITICON-EMOTION MODEL

The proposed EMPOLITICON-Emotion model is a soft voting classifier model with XGB Classifier, Cat Boost Classifier and Linear Discriminant Analysis as the baseline models. To create the soft voting classifier, we have tuned each baseline model to maximize the performance. The parameters of the baseline models that we have used are given on Table 4.

TABLE 4. Parameters of Baseline Models of EMPOLITICON-Emotion.

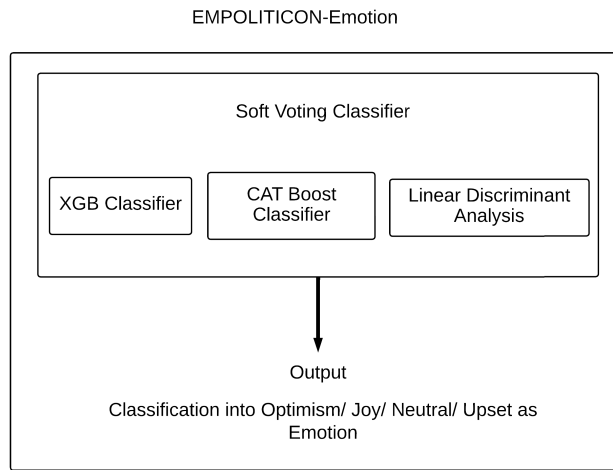| Baseline Model	Parameters
XGB Classifier	learning_rate=0.01, random_state=1
Cat Boost Classifier	n_estimators=1550, learning_rate=0.1, depth=5, task_type='CPU', random_state=1, verbose=False
Linear Discriminant Analysis	solver='lsqr', shrinkage='auto'

EMPOLITICON-Emotion

FIGURE 3. Structure of Proposed EMPOLITICON-Emotion Model.

The structure of the proposed EMPOLITICON-Emotion is shown in Figure 3.

H. PROPOSED EMPOLITICON-CONTEXT MODEL

The proposed EMPOLITICON-Context model is a soft voting classifier model with Nu-Support Vector Classification, Cat Boost Classifier and Linear Discriminant Analysis as the baseline models. To create the soft voting classifier, we have tuned each baseline model to maximize the performance. The parameters of the baseline models that we have used are given on Table 5.

The structure of the proposed EMPOLITICON-Context is shown in Figure 4.

I. CLASSIFICATION OF CONTEXT AND EMOTION

After oversampling, the training dataset was provided as input in our proposed EMPOLITICON-Context and EMPOLITICON-Emotion to train for context and emotion classification respectively and later on the models were tested on the test data. The output of the models are the predicted context and emotion.

IV. RESULTS ANALYSIS

In this section, we will talk about the results which we have obtained from testing our models on our political speeches
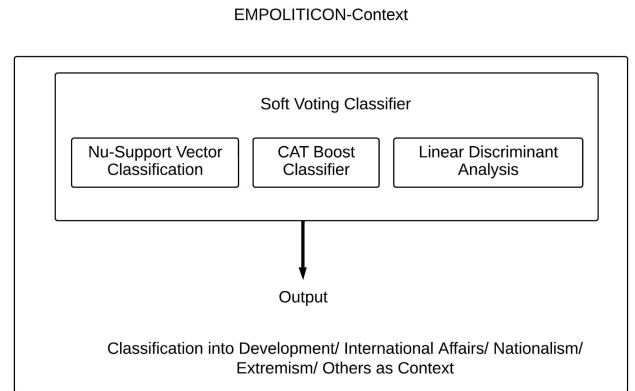
TABLE 5. Parameters of Baseline Models of EMPOLITICON-Context.

Baseline Model	Parameters
Nu-Support Vector Classification	decision_function_shape='ovr', break_ties=True, probability=True
Cat Boost Classifier	n_estimators=1550, learning_rate=0.1, depth=5, task_type='CPU', random_state=1, verbose=False
Linear Discriminant Analysis	solver='lsqr', shrinkage='auto'

EMPOLITICON-Context

FIGURE 4. Structure of Proposed EMPOLITICON-Context Model.

TABLE 6. Precision, Recall and F1-Score of Emotion Classification Using EMPOLITICON-Emotion.

	Precision	Recall	F1-Score
Optimism	0.48	0.42	0.45
Neutral	0.47	0.54	0.50
Joy	0.63	0.64	0.63
Upset	0.57	0.54	0.55

data set. Our evaluating factor was accuracy. Moreover, we have also measured the precision, recall and F1-score of our proposed model. In addition to accuracy, it is a necessity to evaluate and validate the performance of the models. To evaluate the performance of the models, we compared the AUC-ROC Curve of the models. With the help of AUC-ROC curve we can visualize the performance of our classifier models.

A. COMPARATIVE ANALYSIS OF MODELS FOR EMOTION CLASSIFICATION

After testing our EMPOLITICON-Emotion model on our test dataset, we found that our model was able to yield a score of 53.07%. The confusion matrix of EMPOLITICON-Emotion for emotion classification is shown in Figure 5. Moreover, from Table 6 we can see the precision, recall and F1-Score of emotion classification using EMPOLITICON-Emotion.

To compare the performance of our proposed model, we have also tested some other traditional classifier models on our dataset and found that our proposed model performs better than the other models. Moreover, we have also tested
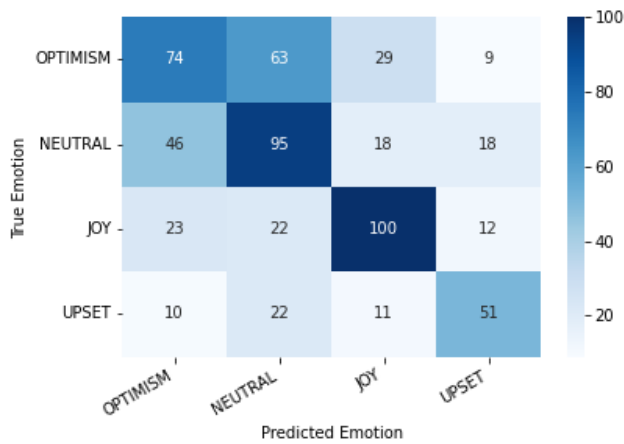
FIGURE 5. Confusion Matrix of EMPOLITICON-Emotion for Emotion Classification.

TABLE 7. Architecture of Bidirectional LSTM for Emotion Classification.

Layer	Output Shape
Input_layer (InputLayer)	(None, 1)
Elmo_Embedding (Lambda)	(None, None, 1024)
BiLSTM (Bidirectional)	(None, 256)
dense (Dense)	(None, 128)
dense_1 (Dense)	(None, 64)
dense_2 (Dense)	(None, 4)
Total params: 1,222,084	
Trainable params: 1,222,084	
Non-trainable params: 0	

TABLE 8. Training Parameters of Bidirectional LSTM for Emotion Classification.

Name of the Parameter	Value
optimizer	adam
loss	sparse_categorical_crossentropy
metrics	accuracy

the performance using deep learning models. We have trained a Bidirectional LSTM and a LSTM [24] model with ELMO [19] as an embedding layer without oversampling the data. However, the deep learning models did not perform well compared to our EMPOLITICON-Emotion model. For the deep learning models, the dataset was split into 80% training data, 10% validation data and 10% test data. We have run 20 epochs on the bidirectional LSTM model but found no significant improvement on the validation accuracy after some epochs and we have run 10 epochs on the LSTM model but found that the validation accuracy got flattened and did not change after a few epochs. The architecture of the bidirectional LSTM for emotion classification is given in Table 7.

Parameters for training the bidirectional LSTM for emotion classification are given in Table 8 and the training vs validation accuracy is shown in Figure 6.

The architecture of the LSTM model for emotion classification is given in Table 9.

Parameters for training the LSTM model are given in Table 10 and the training vs validation accuracy is shown in Figure 7.
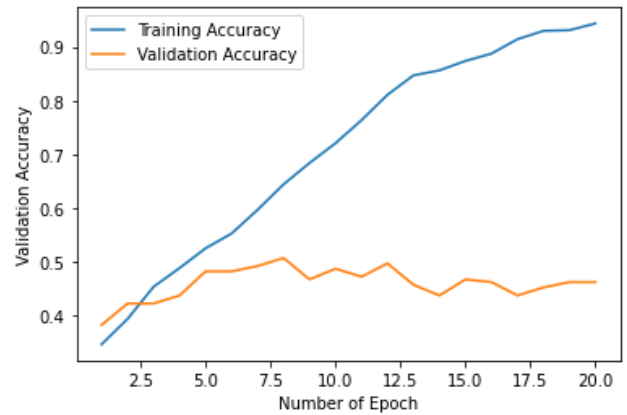

FIGURE 6. Training vs Validation Accuracy of Bidirectional LSTM for Emotion Classification.

TABLE 9. Architecture of LSTM Model for Emotion Classification.

Layer	Output Shape
Input_layer (InputLayer)	(None, 1)
Elmo_Embedding (Lambda)	(None, None, 1024)
LSTM0 (LSTM)	(None, None, 1024)
LSTM1 (LSTM)	(None, None, 512)
LSTM2 (LSTM)	(None, 128)
dense (Dense)	(None, 64)
dense_1 (Dense)	(None, 32)
dense_2 (Dense)	(None, 16)
dense_3 (Dense)	(None, 4)
Total params: 11,879,604	
Trainable params: 11,879,604	
Non-trainable params: 0	

TABLE 10. Training Parameters of LSTM for Emotion Classification.

Name of the Parameter	Value
optimizer	adam
loss	sparse_categorical_crossentropy
metrics	accuracy

The comparison between the models performance based on accuracy are shown in Table 11.

Figure 8-14 represents the AUC-ROC curve of the different models for emotion classification. Looking at the AUC-ROC curve of the models for emotion classification, we can see that the Cat Boost Classifier model performs the best in classifying optimism with our proposed EMPOLITICON-Emotion at the third. On the other hand, the EMPOLITICON-Emotion performs better than the other models in identifying the Neutral, Joy and Upset emotions. To draw an overall conclusion on the AUC-ROC curve, we can see that the EMPOLITICON-Emotion model performs better than the other models.

B. COMPARATIVE ANALYSIS OF MODELS FOR CONTEXT CLASSIFICATION

After testing EMPOLITICON-Context on our test data set, we see that the model was able to achieve an accuracy of
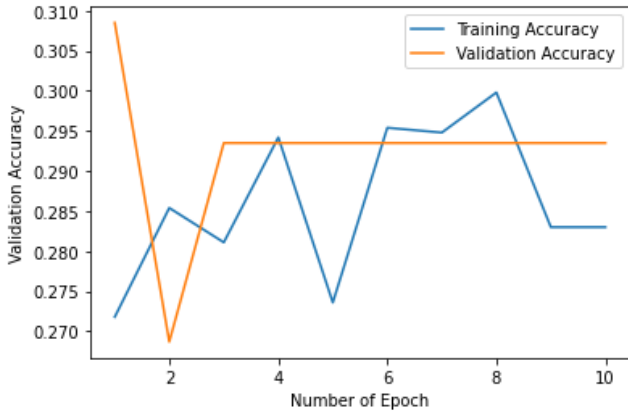
FIGURE 7. Training vs Validation Accuracy of LSTM for Emotion Classification.

TABLE 11. A Compararison of Models based on Accuracy for Emotion Classification.

Model	Accuracy
Gradient Boosting Classifier	51.91%
XGB Classifier	48.59%
Linear Discriminant Analysis	36.82%
Cat Boost Classifier	52.40%
LSTM	28.86%
Bidirectional LSTM	45.27%
EMPOLITICON-Emotion	53.07%
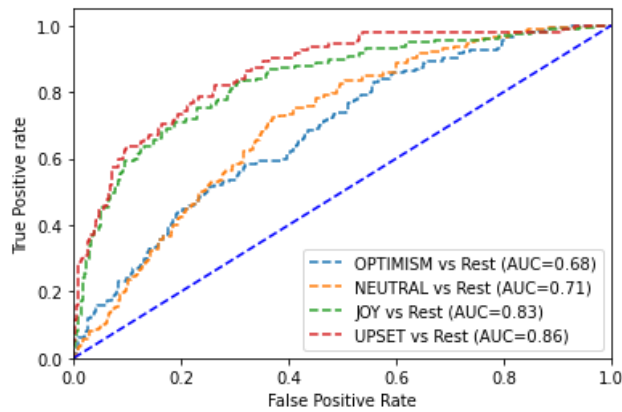
FIGURE 8. AUC-ROC Curve of EMPOLITICON-Emotion for Emotion Classification.

FIGURE 9. AUC-ROC Curve of XGB Classifier for Emotion Classification.
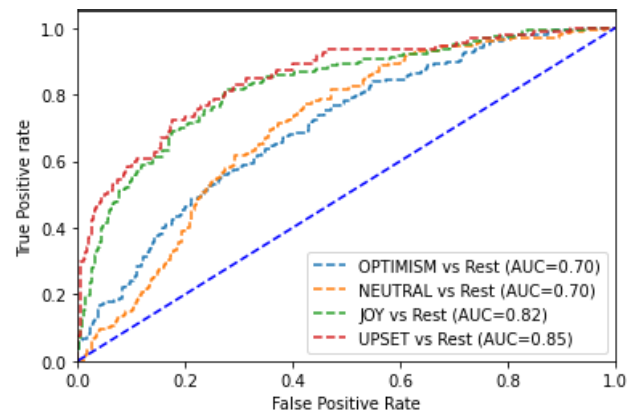
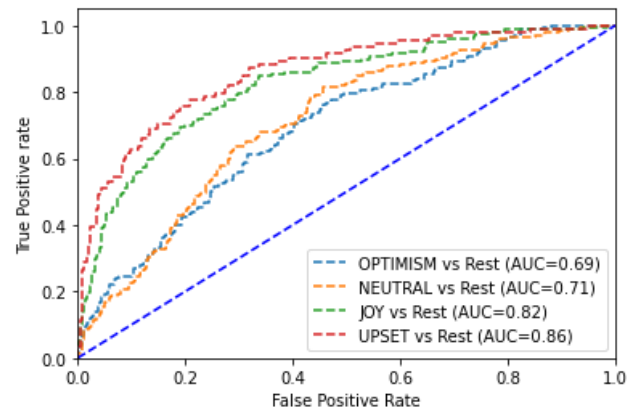**FIGURE 10.** AUC-ROC Curve of Cat Boost Classifier for Emotion Classification.

FIGURE 11. AUC-ROC Curve of Gradient Boosting Classifier for Emotion Classification.

73.13%. The confusion matrix of EMPOLITICON-Context is shown in Figure 15.

The precision, recall and F1-Score of context classification using EMPOLITICON-Context can be seen from Table 12.

To compare the performance of our proposed model, we have also tested some other traditional classifier models on our dataset and found that our proposed model performs better than the other models. Moreover, we have also tested the performance using deep learning models. We have trained a Bidirectional LSTM and a LSTM model with ELMO as

an embedding layer without oversampling the data. However, the deep learning models did not perform well compared to our proposed EMPOLITICON-Context. For the deep learning models, the dataset was split into 80% training data, 10% validation data and 10% test data. We have run 20 epochs on the bidirectional LSTM model but found no significant
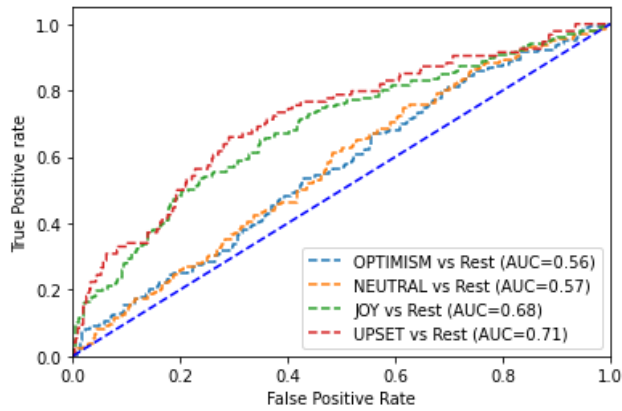
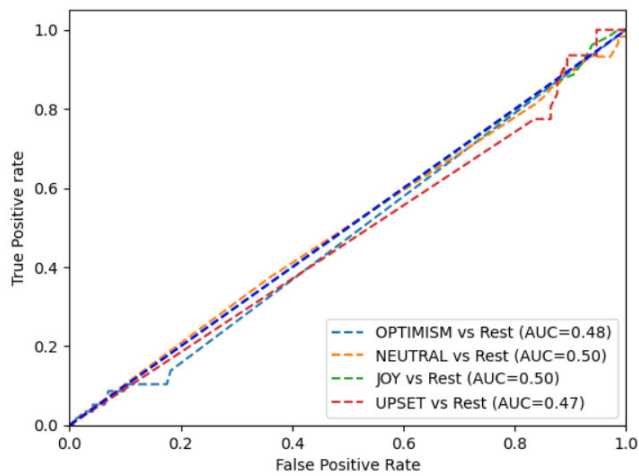**FIGURE 12.** AUC-ROC Curve of Linear Discriminant Analysis for Emotion Classification.

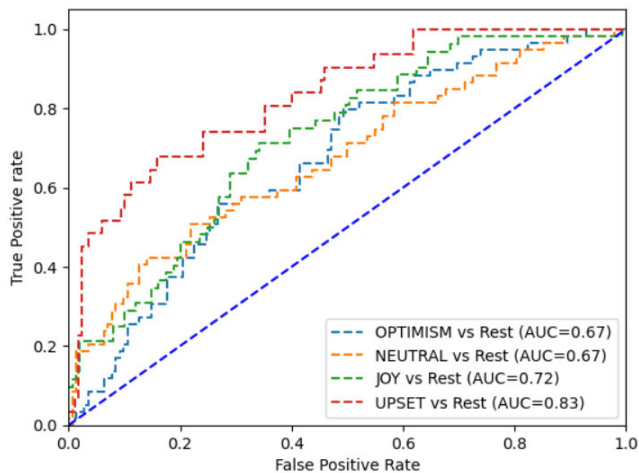**FIGURE 13.** AUC-ROC Curve of LSTM for Emotion Classification.

FIGURE 14. AUC-ROC Curve of Bidirectional LSTM for Emotion Classification.

improvement in the validation accuracy after some epochs and we have run 10 epochs on the LSTM model but found that the validation accuracy was flat and did not change. The
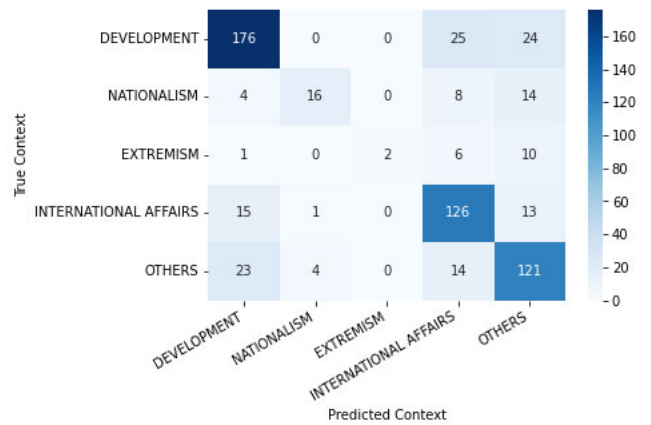

FIGURE 15. Confusion Matrix of EMPOLITICON-Context for Context Classification.

TABLE 12. Precision,Recall and F1-Score of Context Classification Using EMPOLITICON-Context.

	Precision	Recall	F1-Score
Development	0.80	0.78	0.79
Nationalism	0.76	0.38	0.51
Extremism	1.00	0.11	0.19
International Affairs	0.70	0.81	0.75
Others	0.66	0.75	0.70

TABLE 13. Architecture of Bidirectional LSTM for Context Classification.

Layer	Output Shape
Input_layer (InputLayer)	(None, 1)
Elmo_Embedding (Lambda)	(None, None, 1024)
BiLSTM (Bidirectional)	(None, 256)
dense (Dense)	(None, 128)
dense_1 (Dense)	(None, 64)
dense_2 (Dense)	(None, 5)
Total params: 1,222,149	
Trainable params: 1,222,149	
Non-trainable params: 0	

TABLE 14. Training Parameters of Bidirectional LSTM for Context Classification.

Name of the Parameter	Value
optimizer	adam
loss	sparse_categorical_crossentropy
metrics	accuracy

architecture of the bidirectional LSTM for context classification is given in Table 13.

Parameters for training the bidirectional LSTM for context classification are given in Table 14 and the training vs validation accuracy is shown in Figure 16.

The architecture of the LSTM model for context classification is given in Table 15.

Parameters for training the LSTM model for context classification are given in Table 16 and the training vs validation accuracy is shown in Figure 17.

The comparison between the models performance based on accuracy are shown in Table 17.
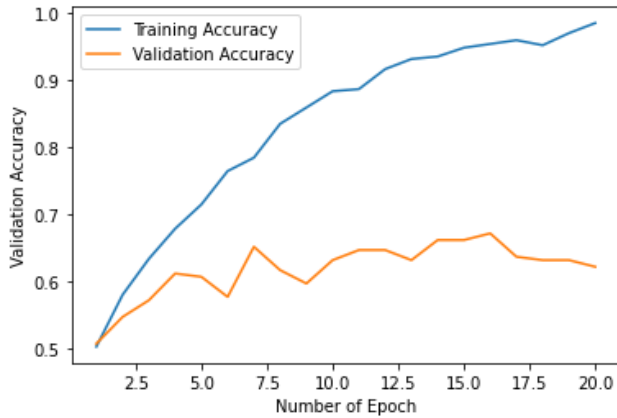
FIGURE 16. Training vs Validation Accuracy of Bidirectional LSTM for Context Classification.

TABLE 15. Architecture of LSTM Model for Context Classification.

Layer	Output Shape
Input_layer (InputLayer)	(None, 1)
Elmo_Embedding (Lambda)	(None, None, 1024)
LSTM0 (LSTM)	(None, None, 1024)
LSTM1 (LSTM)	(None, None, 512)
LSTM2 (LSTM)	(None, 128)
dense (Dense)	(None, 64)
dense_1 (Dense)	(None, 32)
dense_2 (Dense)	(None, 16)
dense_3 (Dense)	(None, 5)
Total params: 11,879,621	
Trainable params: 11,879,621	
Non-trainable params: 0	

Figure 18-24 represents the AUC-ROC curve of the different models for context classification. Looking at the AUC-ROC curve of the models for context classification, we can see that the BILSTM Model performs the best in classifying the development with our proposed EMPOLITICON-Context at the second. Similarly, the Cat Boost Classifier performs better in classifying the nationalism and extremism context with EMPOLITICON-Context at the second. On the other hand, the EMPOLITICON-Context performs better than the other models in identifying the international affairs and the others context. Moreover, the AUC value of EMPOLITICON-Context for all the classes is higher than 0.85. When we draw an overall conclusion on the AUC-ROC curve, we can see that the EMPOLITICON-Context performs better than the other models.

V. DISCUSSIONS
In this section, we discuss the results of our proposed EMPOLITICON-Emotion and EMPOLITICON-Context in classifying the emotion and context of political speeches respectively.

A. EMOTION CLASSIFICATION USING EMPOLITICON-EMOTION
EMPOLITICON-Emotion is a soft voting classifier that uses XGB classifier, Cat Boost classifier and Linear Discriminant

TABLE 16. Training Parameters of LSTM for Context Classification.

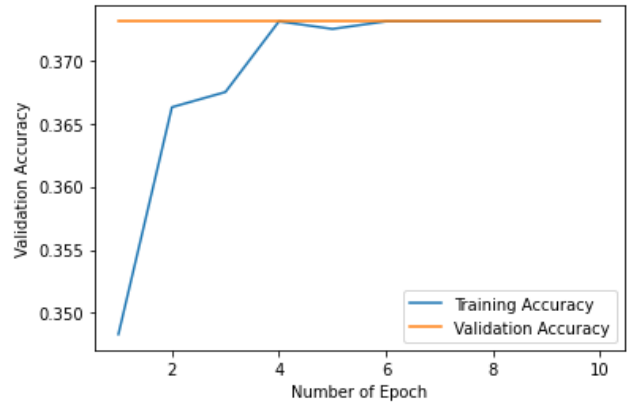| Name of the Parameter	Value
optimizer	adam
loss	sparse_categorical_crossentropy
metrics	accuracy

FIGURE 17. Training vs Validation Accuracy of LSTM for Context Classification.

Analysis as the baseline models. XGBoost and Cat Boost are boosting ensemble models that aim to maximize the model's performance by training weak learners. XGboost comes with a built in cross validation method that helps to prevent overfitting when the dataset is small. On the other hand, Cat Boost prevents overfitting using a gradient boosting scheme. Linear Discriminant Analysis finds a linear combination of features that best separates the classes in a dataset. These as baseline models make the EMPOLITICON-Emotion perform better than the other models in classifying the emotion of political speeches. From the confusion matrix of Figure 5, we can see that the proposed EMPOLITICON-Emotion model is quite good at identifying the emotions of the speeches. However, the confusion matrix shows that it also faced some struggles in identifying the correct emotion for all the speeches. The confusion matrix shows that the EMPOLITICON-Emotion model performed best in terms of classifying joy. After joy, the model performs better in classifying the neutral class. From the confusion matrix, we see that the model confuses between optimism and neutral. The fact that different subtexts of a speech might carry different emotions makes it difficult to draw an overall emotion label for the entire speech. Moreover, looking at the predictions of EMPOLITICON-Emotion, we found that the model predicted 33.50% speeches as neutral, 26.20% speeches as joy, 25.37% speeches as optimism and 14.93% speeches as upset.

B. CONTEXT CLASSIFICATION USING EMPOLITICON-CONTEXT
EMPOLITICON-Context is a soft voting classifier that uses Nu Support Vector classification, Cat Boost classifier and Linear Discriminant Analysis as the baseline models. As mentioned earlier, Cat Boost aims to maximize the

TABLE 17. A Comparison of Models based on Accuracy for Context Classification.

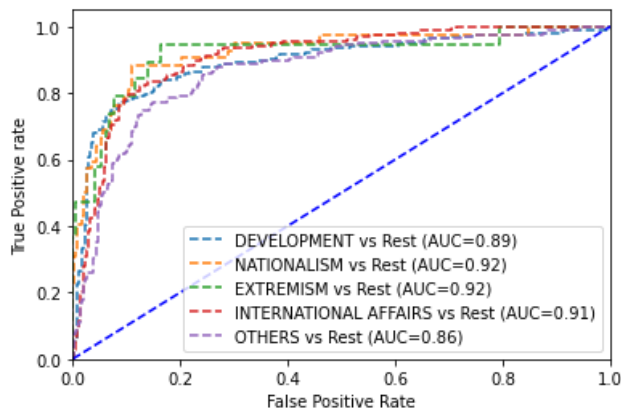| Model	Accuracy
Gradient Boosting Classifier	65.67%
XGB Classifier	69.15%
Cat Boost Classifier	69.82%
Linear Discriminant Analysis	64.51%
LSTM	37.31%
Bidirectional LSTM	63.18%
EMPOLITICON-Context	73.13%

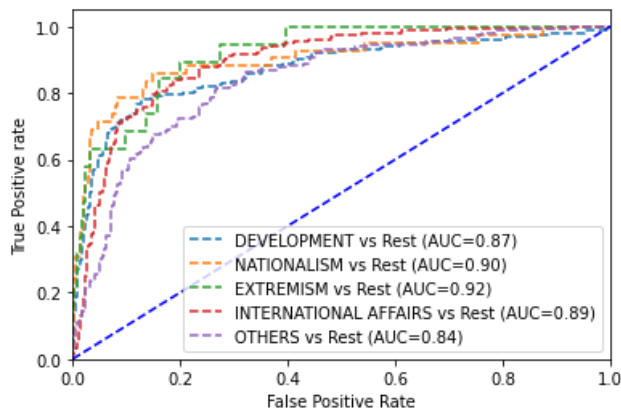**FIGURE 18.** AUC-ROC Curve of EMPOLITICON-Context for Context Classification.

FIGURE 19. AUC-ROC Curve of XGB Classifier for Context Classification.

FIGURE 20. AUC-ROC Curve of Cat Boost Classifier for Context Classification.

FIGURE 21. AUC-ROC Curve of Gradient Boosting Classifier for Context Classification.

FIGURE 22. AUC-ROC Curve of Linear Discriminant Analysis for Context Classification.

model's performance by training weak learners and preventing overfitting using a gradient boosting scheme. On the other hand, Linear Discriminant Analysis finds a linear combination of features that best separates the classes in a dataset. Moreover, Nu Support Vector Classification can control the number of support vectors with the help of the nu parameter. These as baseline models make the EMPOLITICON-Context perform better than the other models in classifying the context of political speeches. Looking at the confusion matrix of EMPOLITICON-Context in Figure 15, we see that the score
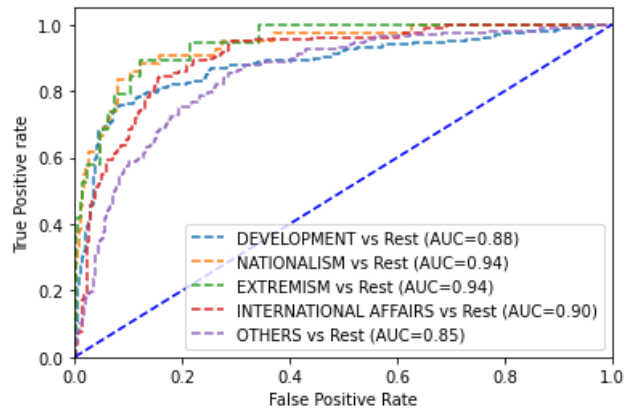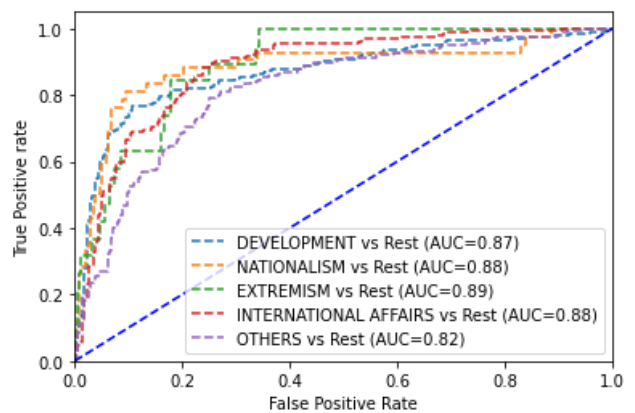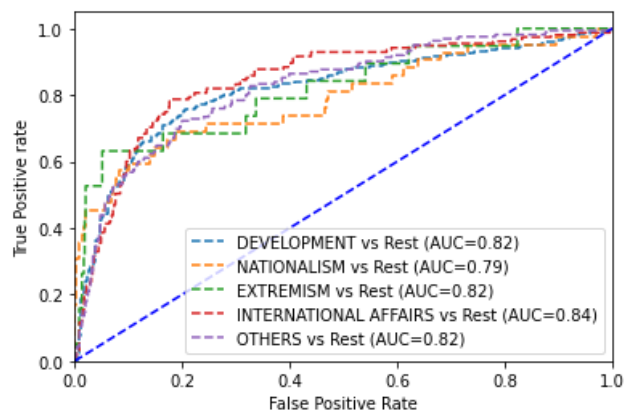
of development is the highest. It can also be said that the development speeches are commonly confused with speeches that convey the international affairs context. The main reason for such cases occurring is due to the fact that different international events related to economy, climate, sports, education etc. occur together. So, development gets confused with the
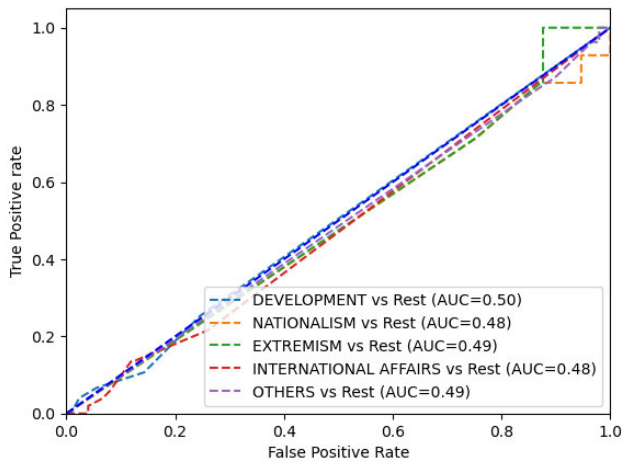
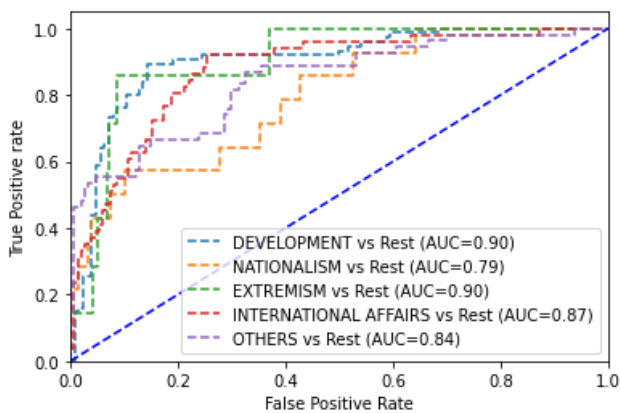**FIGURE 23.** AUC-ROC Curve of LSTM for Context Classification.

FIGURE 24. AUC-ROC Curve of Bidirectional LSTM for Context Classification.

speeches of international affairs more when compared to the other speeches. Next, looking at the nationalism context, we see that the model performed quite poorly in identifying nationalism. The model performed the worst in identifying extremism. The model confuses the extremism speeches with international affairs and nationalism speeches. A reason for this is that the politicians talk about other countries or the nation's unity while talking about extremism. Moreover, terrorism, hate speeches and inciteful comments are very often hidden cleverly in the speeches as the politicians do not want to cause a diplomatic issue. As a result, they tend to hide their true feelings by hiding them within the speeches addressing international terrorism and in other issues. However, we can see that the model was quite good at identifying the international affairs and other context in speeches. The performance of the proposed model for the classification of nationalism and extremism still has some room for improvement, which can be achieved by enriching the dataset more with samples from these classes. Moreover, from the predictions of EMPOLITICON-Context, we found that the model predicted 36.32% speeches as development, 30.18% speeches as others, 29.685% speeches as international affairs, 3.483% speeches as nationalism and 0.332% speeches as extremism.

VI. CONCLUSION

The primary focus of our work was to classify the political speeches in terms of the context and emotion that the speech consists of. To conduct the research, we at first built our own dataset consisting of 2010 speeches from the country leaders of China, Russia, United Kingdom and United States which are the permanent members of the United Nations Security Council. There were five classes in context and four classes in emotion. We have used longformer for generating embedding. Due to the imbalance of classes, it was required to oversample the data. Our models EMPOLITICON-Context and EMPOLITICON-Emotion performed very good in terms of identifying the context and emotion of political speeches respectively. We see that our models outperformed the traditional classifiers and also Bidirectional LSTM and LSTM models. EMPOLITICON-Context has achieved 73.13% accuracy in terms of context classification and EMPOLITICON-Emotion has achieved 53.07% accuracy in classifying the emotion of the political speeches. One of the key challenges of this research was creating the dataset. Moreover, the speeches that we have worked with are long textual data which created a challenge for the classification task. Different parts of a speech may contain different sub-context and emotions for which the classification of emotion and context was challenging. We see that the EMPOLITICON-Context struggled a bit in order to classify the extremism class in context classification on which more care needs to be taken. We also see that in emotion classification, the EMPOLITICON-Emotion sometimes gets confused between optimism, neutral and joy. It is because different parts of the speech contain different emotions and thus the model gets confused sometimes. Identifying the emotions and context of different sections of a speech can be a scope for further research.

REFERENCES

[1] R. Baly, H. Hajj, N. Habash, K. B. Shaban, and W. El-Hajj, "A sentiment treebank and morphologically enriched recursive deep models for effective sentiment analysis in Arabic," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 16, no. 4, pp. 1–21, Dec. 2017.

[2] A. Alnawas and N. Arici, "Sentiment analysis of Iraqi Arabic dialect on Facebook based on distributed representations of documents," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 18, no. 3, pp. 1–17, Sep. 2019.

[3] J. He, L. Wang, L. Liu, J. Feng, and H. Wu, "Long document classification from local word glimpses via recurrent attention learning," *IEEE Access*, vol. 7, pp. 40707–40718, 2019.

[4] J. Awwalu, A. A. Bakar, and M. R. Yaakub, "Hybrid N-gram model using Naïve Bayes for classification of political sentiments on Twitter," *Neural Comput. Appl.*, vol. 31, no. 12, pp. 9207–9220, Dec. 2019.

[5] E. H. R. Rho and M. Mazmanian, "Hashtag burnout? A control experiment investigating how political hashtags shape reactions to news content," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, pp. 1–25, Nov. 2019.

[6] T. Huang, Q. She, and J. Zhang, "BoostingBERT: Integrating multi-class boosting into BERT for NLP tasks," 2020, *arXiv:2009.05959*.

[7] J. A. Caetano, H. S. Lima, M. F. Santos, and H. T. Marques-Neto, "Using sentiment analysis to define Twitter political users' classes and their homophily during the 2016 American presidential election," *J. Internet Services Appl.*, vol. 9, no. 1, pp. 1–15, Dec. 2018.

[8] R. Castro and C. Vaca, "National leaders' Twitter speech to infer political leaning and election results in 2015 Venezuelan parliamentary elections," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2017, pp. 866–871.

[9] M. G. Folgado and V. Sanz, "Exploring the political pulse of a country using data science tools," *J. Comput. Social Sci.*, vol. 5, no. 1, pp. 987–1000, May 2022.

[10] I. Onyenwe, S. Nwagbo, N. Mbeledogu, and E. Onyedinma, "The impact of political party/candidate on the election results from a sentiment analysis perspective using #AnambraDecides2017 tweets," *Social Netw. Anal. Mining*, vol. 10, no. 1, pp. 1–17, Dec. 2020.

[11] R. Sandoval-Almazan and D. Valle-Cruz, "Sentiment analysis of Facebook users reacting to political campaign posts," *Digit. Government, Res. Pract.*, vol. 1, no. 2, pp. 1–13, Apr. 2020.

[12] S. P. Algur and R. H. Patil, "Sentiment analysis by identifying the speaker's polarity in Twitter data," in *Proc. Int. Conf. Elect., Electron., Commun., Comput., Optim. Techn. (ICEECCOT)*, Dec. 2017, pp. 1–5.

[13] M. Kaya, G. Fidan, and I. H. Toroslu, "Sentiment analysis of Turkish political news," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, Dec. 2012, pp. 174–180.

[14] T. Xiong and V. Cherkassky, "A combined SVM and LDA approach for classification," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Dec. 2005, pp. 1455–1459.

[15] S. Park, B. Bae, and Y. Cheong, "Emotion recognition from text stories using an emotion embedding model," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Feb. 2020, pp. 579–583.

[16] Y. Lee, C. Park, and H. Choi, "Word-level emotion embedding based on semi-supervised learning for emotional classification in dialogue," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Feb. 2019, pp. 1–4.

[17] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020, *arXiv:2004.05150*.

[18] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, vol. 2. New York, NY, USA: Springer, 2009, pp. 1–758.

[19] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018, *arXiv:1802.05365*.

[20] P. Bambroo and A. Awasthi, "LegalDB: Long DistilBERT for legal document classification," in *Proc. Int. Conf. Adv. Electr., Comput., Commun. Sustain. Technol. (ICAECT)*, Feb. 2021, pp. 1–4.

[21] K. Fiok, W. Karwowski, E. Gutierrez-Franco, M. R. Davahli, M. Wilamowski, T. Ahram, A. Al-Juaid, and J. Zurada, "Text guide: Improving the quality of long text classification by a text selection method based on feature importance," *IEEE Access*, vol. 9, pp. 105439–105450, 2021.

[22] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[23] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural Comput.*, vol. 12, no. 5, pp. 1207–1245, May 2000.

[26] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *Amer. Scientist*, vol. 89, no. 4, pp. 344–350, 2001.

ASIF ATIQ received the B.Sc. degree in computer science and engineering from BRAC University, Dhaka, Bangladesh. He is currently pursuing the Master of Applied Science in computer engineering with the Memorial University of Newfoundland, St. John's, NL, Canada. His research interests include natural language processing, blockchain, data science, machine learning, embedded systems, cyber security, and image processing.

ABRAR SHAHRIAR ABEED received the B.Sc. degree in computer science and engineering from BRAC University, Dhaka, Bangladesh. He has been with the Kona Software Laboratory, Dhaka, Bangladesh, since June 2022, as a Graduate Engineer. His research interests include image processing, blockchain technology, natural language processing, and quantum computing.

ARMANUL MOMIN received the B.Sc. degree in computer science and engineering from BRAC University, Dhaka, Bangladesh. He has been a Mobile Application Developer with Talent Pro, Dhaka, since February 2023. His research interests include natural language processing, machine learning, image processing, and cyber security.

AZHER AHMED EFAT received the B.Sc. degree in computer science and engineering from BRAC University, Dhaka, Bangladesh. He has been a Graduate Engineer with the Kona Software Laboratory, Dhaka, since June 2022. His research interests include natural language processing, blockchain, and image processing.

MD. GOLAM RABIUL ALAM (Member, IEEE) received the B.S. degree in computer science and engineering, the M.S. degree in information technology, and the Ph.D. degree in computer engineering from Kyung Hee University, South Korea, in 2017. He was a Postdoctoral Researcher with the Computer Science and Engineering Department, Kyung Hee University, from March 2017 to February 2018. He is currently a Full Professor with the Department of Computer Science and Engineering, BRAC University, Bangladesh. His research interests include healthcare informatics, mobile cloud and edge computing, ambient intelligence, and persuasive technology. He is a member of the IEEE IES, CES, CS, SPS, CIS, KIISE, and IEEE ComSoc. He received several best paper awards in prestigious conferences.

● ● ●