**Text Classification**

**Problem Statement:** The rise of social media has completely changed how individuals get information. Platforms like Facebook and Twitter make it possible to obtain information more quickly, widely, and without as many restrictions, but they also serve as a fertile ground for the spread of false information. A significant number of tweets in twitter about covid-19 during pandemic misleading the real news. According to the research, false information about the coronavirus is spreading more quickly than the truth. Additionally, these false tweets pose a serious threat to people's lives. To solve this issue regarding fake tweets we can use text classification which is a NLP technique to detect whether the tweets are real or fake by using a real world dataset associated with recent COVID-19 pandemic.

**Existing Models:** CTF, the first corona virus Twitter dataset, and cross-SEAN, a model for distinguishing between genuine and false tweets, are other contributions made by the group (cross stitch-based semisupervised end-to-end neural attention model). The total number of tweets in the CTF dataset is 45.26k, of which 18.55k have been determined to be real and 26.71k to be false. Their model exceeds all others by reaching the 0.95 F1 score when they compared the behaviour of their model with that of seven currently used approaches. For economic, political, and social reasons, fake news is now very common. The offline community has also been greatly impacted by news of the coronavirus. It is more crucial in these circumstances to tell the difference between the real thing and fake COVID-19 news to alleviate concerns about this harmful virus. They made use of a web-based dataset that dealt with COVID-19 false news binary classification. They pre-processed the dataset and used TF and IDF to extract the features. Then, they use decision trees and random forests to train their model, and different parameters to evaluate it. The proposed model achieved 92.07% accuracy with DT and 94.49% accuracy with RF classifier.

**Models:**

**Logistic Regression:** One of the most used classification methods is LR. It is a statistical model that uses a vector of variables to determine the weights assigned to each variable and, using this information, predicts the category of declared fake tweets on COVID-19 as a word vector. Only when the dependent variable is binary can LR be applied. In LR, the independent variable cannot be normally distributed or have an equal variance across groups, and there is no linear relationship between the dependent and independent variables.

**Naive Bayes:** NB is a probabilistic classifier that uses the Bayes theorem as its foundation. Its simplicity, accuracy, and dependability are the key factors contributing to its appeal. It has been used in many practical contexts, but natural language processing issues are where it has seen the most use. The fundamental premise of NB is that each characteristic contributes equally and independently to the result, which is why NB is referred to be "Naive." Using previously collected data that could be associated to a characteristic, it determines the likelihood of that attribute.

**Boosting:** By combining several weak classifiers, the ensemble modelling technique known as "boosting" aims to create a powerful classifier. It is accomplished by using weak models in series to develop a model. First, a model is created using the training set of data. The second model is then created in an effort to fix the previous model's flaws. Models are added in this manner until either the full training data set is properly predicted or the maximum numbers of models have been added.

**Bert**: BERT is just transformer architecture with an encoder stack. An encoder-decoder network using self-attention on the encoder side and attention on the decoder side is known as transformer architecture. The Encoder stack in BERTLARGE contains 24 layers compared to BERTBASE's 12 layers. These go beyond the Transformer design as it was originally defined in the paper (6 encoder layers). In addition, LARGE and BASE BERT architectures have more attention heads (12 and 16 respectively) and larger feed forward networks (768 and 1024 hidden units) than the Transformer architecture proposed in the original paper.

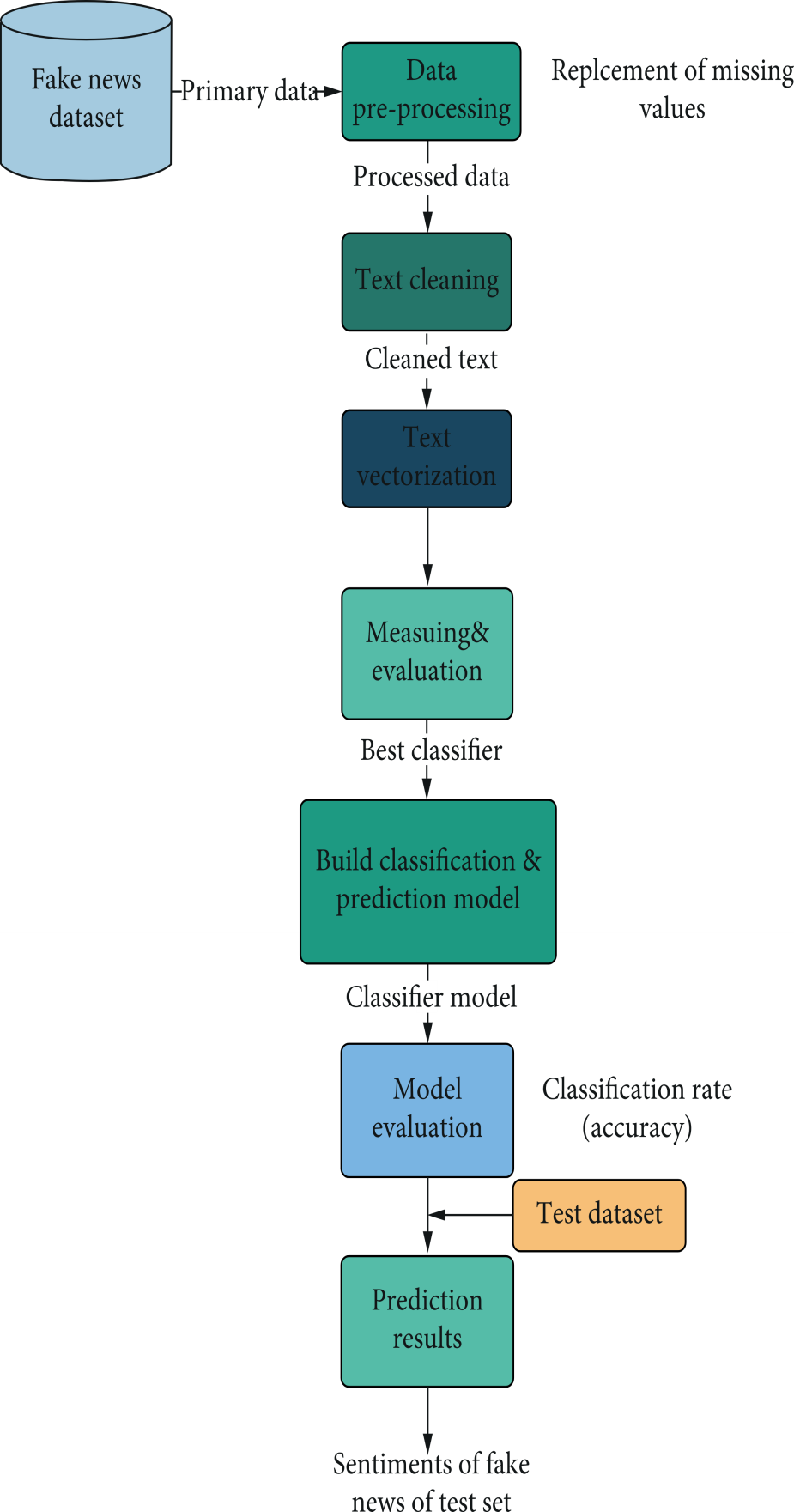
**Support vector machine:** We used SVM, an ML classifier, in our research. SVM has produced noteworthy results for numerous real-world applications and is applicable to both linear and nonlinear problems. SVM uses a line or hyper plane to divide the data into classes. Due to its functions known as kernels, which take input space in low dimensions and turn it into high dimensional space, it performs best for non-linear issues. SVM can quickly execute extremely complicated data transformations and classify data into appropriate groups.

**Multilayer Perceptron:** Feed-forward artificial neural networks include MLP as a subset (ANN). All human brains exhibit the phenomena that are reflected by ANN. The primary source of inspiration for ANN is the way the brain processes input, understands it, and produces answers. An artificial neural network (ANN) can learn from input data and relate it to the desired output variable. The basic building block of artificial neural networks is the perceptron. Each perceptron receives some weighted input and uses an activation function to produce output. They have a wide range of current applications, including protein secondary structure, character identification, data compression, pattern recognition, and computer vision.

**BI-LSTM:** A recurrent neural network used largely for natural language processing is called Bidirectional LSTM (Bi-LSTM). It may use data from both sides and, unlike regular LSTM, the input flows in both directions. In both directions of the sequence, it is a potent tool for modelling the sequential relationships between words and phrases. In conclusion, Bi-LSTM reverses the direction of information flow by adding one extra LSTM layer. It simply means that in the additional LSTM layer, the input sequence flows backward. The outputs from the two LSTM layers are then combined in a variety of ways, including average, sum, multiplication, and concatenation.

|  |  |  |  |
| --- | --- | --- | --- |
| **Models** | **Word Embedding Methods** | **Accuracy Score** | **F1 Score** |
| Logistic Regression | Count Vectorizer | 0.92 | 0.92 |
| Word Level TF-IDF | 0.92 | 0.92 |
| NGram TF-IDF | 0.84 | 0.84 |
| Char TF-IDF | 0.91 | 0.91 |
| Naïve Bayes | Count Vectorizer | 0.92 | 0.92 |
| TF-IDF | 0.90 | 0.90 |
| Boosting | Count Vectorizer | 0.92 | 0.92 |
| Word Level TF-IDF | 0.91 | 0.91 |
| NGram TF-IDF | 0.84 | 0.84 |
| Char TF-IDF | 0.95 | 0.95 |
| SVM | Count Vectorizer | 0.52 | 0.69 |
| Word Level TF-IDF | 0.52 | 0.69 |
| NGram TF-IDF | 0.52 | 0.69 |
| Char TF-IDF | 0.52 | 0.69 |
| MLP | Count Vectorizer | 0.91 | 0.91 |
| Word Level TF-IDF | 0.91 | 0.91 |
| NGram TF-IDF | 0.88 | 0.88 |
| Char TF-IDF | 0.90 | 0.90 |
| Bi-LSTM | - | 0.92 | 0.92 |
| Bert | - | 0.96 | 0.96 |

**Model Architecture:**



Classifying whether Real or Fake

**Dataset**: Dataset used in this project is tweets about covid-19 which consists of 6421 rows and 3 columns in which first column consists of serial numbers and second column consists of tweets and third column consists of real or fake.

**Data preprocessing:** During the preprocessing step, we apply several cleaning and filtering techniques on this fake news on COVID-19, such as removing links, identifiers, deleting words that contain several less than 3 characters, and filtering empty words.

**Data Vectorization**: Because most automatic learning algorithms do not take text directly but digital vectors, the transformation of text to digital vectors is performed based on the count vectorizer technique with the TF-IDF method for calculating the score of each word.

##### **Classification Model Building:** We choose the most efficient classification algorithm based on the results, and then we built our classification model.

##### **Evaluation and Testing:** Training and evaluation of classification model by performance measures test the model on a set of test data that represents a set of unclassified tweets on COVID-19, to predict whether they are real or fake COVID-19 among this set.

**Model results and discussions:** The findings of all machine learning-based models that were tested can be seen in the aforementioned tables. All of the classifiers performed admirably; although, certain classifiers outperformed others in terms of f1-score and accuracy, when compared to other machine learning classifiers. We have tested various models on a large dataset comprising COVID-19 Tweets. We have divided our dataset into two halves, one for training and one for testing. Out of all Bert model and Boosting model performed well with 0.96 and 0.95 accuracy score.

**Conclusion**: According to the results obtained in the project, to detect whether real or fake tweet using text classification technique, we used seven different classifiers: NAÏVE BAYES, BOOSTING, SVM, BERT, BI-LSTM, MLP, and LR. From the comparison of the different measures, we find that BERT model perform better than other learning methods even though machine learning algorithms give a good accuracy score with 0.96.We audited various techniques and conducted experiments on the dataset from a reliable repository to find or adapt the best classifier. Furthermore, the system has been analysed in the aspects of accuracy and F1-score for all the algorithms.