Introduction:

Internet and social media have made the access to the news information much easier and comfortable. When people need to follow their event of interest, they just use their phones online. Unfortunately, some challenges have become apparent in disseminating information. More clearly, the prevalence of information on the Internet and social media have made intricate on users to distinguish between the fake and real news. As a result, mass media which has a great influence on the society may manipulate information in different ways. It means that people can easily produce mixed news of true and false information or false information completely. This has given motivation that there are several websites generating false information exclusively. Deliberately, they publish propaganda, disinformation, and hoaxes for sake of personal reasons, one of which is affecting public opinions on certain matters: political mostly. They amplify the effect of such information by publishing it on social media. A classic example of fake news is the mainstream news during U.S. 2016 president election (). Thus, there is a need for a solution to distinguish between fake and real news. Since artificial intelligence algorithms have started to work much better than before on classification problems, such as image recognition and sentiment analysis detection, and datasets of fake news have become ubiquitous, scientists believe that this problem can be addressed by means of machine learning and artificial intelligence. Thus, the researchers of the paper have been catalyzed to understand how machine learning can detect fake news. This paper discusses different types of machine learning algorithms for detecting fake news from text. In this sense, the research project will use multiple NLP models, such as BERT. Regarding datasets, LIAR, Fakeddit, Real News, and Snopes datasets are used. The goal of the research is to see which machine learning algorithm explained in the class will be the most accurate one for detecting the fake news.

Literature Review:

<https://www.sciencedirect.com/science/article/pii/S0957417419301988>

**First Paper: “Behind the cues: A benchmarking study for fake news detection”**

They have proposed a model for fake news detection using content-based features and machine learning algorithms. More specifically, they made use of linguistic features (e.g.words belonging in certain categories, part of Speech tags and others) in a combination with ML algorithms (the most popular ML classifiers).

**Feature Set Benchmarking pipeline:**

They created three tables for linguistic features from the experience of other papers in the literature review. The first paper separated the features to four categories, namely grammatical complexity, vocabulary complexity, quantity, and Specificity and Expressiveness [a]. The second paper focused more on multivariate linguistic profile of deception by presenting the total features in three categories, namely psychological processes, standard linguistic dimensions, and relativity [b]. The third paper increased the number of features for a hope of improving the accuracy of fake news detection. They categorized the feature to nine features: quantity, complexity, uncertainty, non-immediacy, expressivity, Specificity, and affect [c].

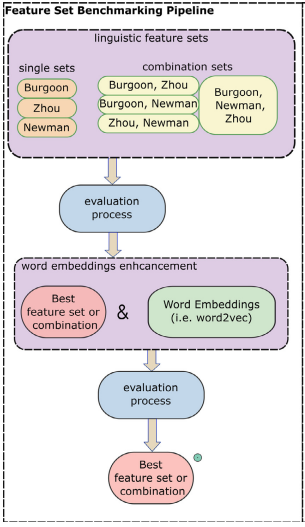
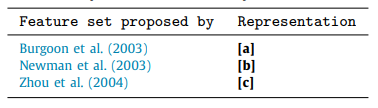


Figure 1.

Table 1



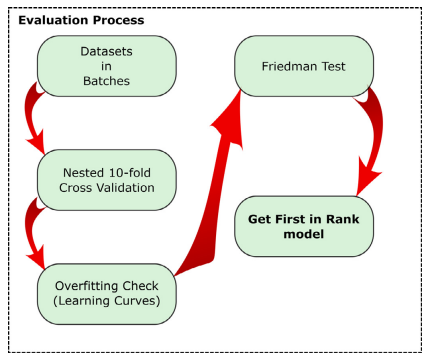
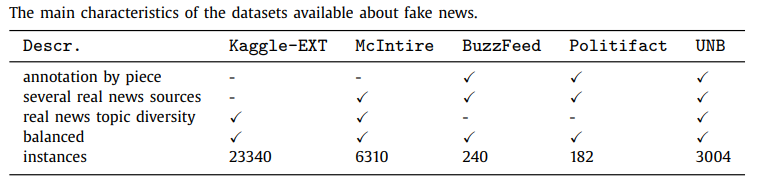


Figure 2

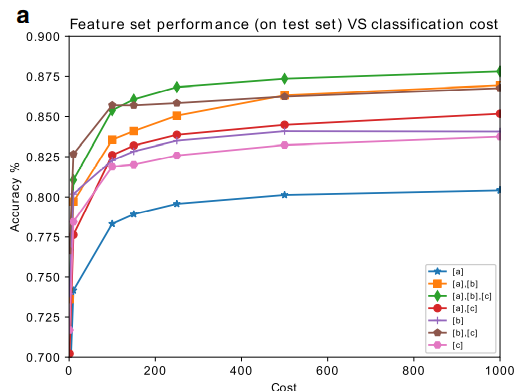
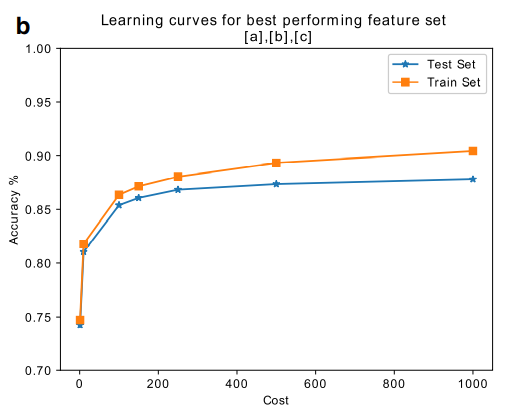
From figure 1, the feature set benchmarking pipeline is comprised of different processes. In the first process, the Georgies Gravanies and his researchers succeeded in taking an advantage of the previous papers by experimenting the whole combination of three tables, together for sake of boosting the accuracy and the usage of linguistic features as presented in the table 1. The second process is “Evaluation Process”. To ensure the equivalent evaluation of all feature sets and algorithms, they followed the steps as shown in figure 2. The figure 2 illustrates systematic steps to for equivalent evaluations for both features and algorithms. First, they split the dataset up to 1000 random batches. Second, in each dataset batch, they used 10-fold cross-validations. To avoid the overfitting, they referred to learning curves plots and got the optimum parameters values. Then, to compare the significance of multiple methods over several datasets, they used the Friedman Test. The third process is for word embedding. They utitlized the Word2vec and GloVe for learning word embedding from raw text. When experimenting both, they did not notice any difference in terms of accuracy. Finally, they tried to use the same aforementioned evaluation process to compare the best performance after enhancement between the combination of linguistic feature, word embeddings, and both of them together. They did not step at this point, but they created an unbiased fake news dataset with taking into consideration four important standards. First, Fake news articles should be reviewed by experts. Second, Real news should be published from credible organizations. Third, Fake news should originate from several sources. Fourth, it is a must to obtain several articles from a varied number of categories. From the table below, it is apparent that they created UNB (unbiased dataset) for covering all mentioned description compared to the other datasets.

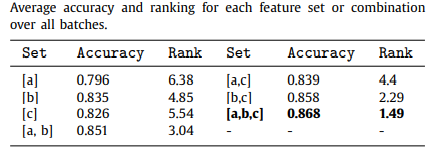


**Machine Learning algorithms:**

**The Results:**

Experimenting all combinations of linguistic Features:

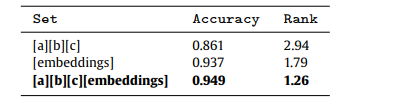
 

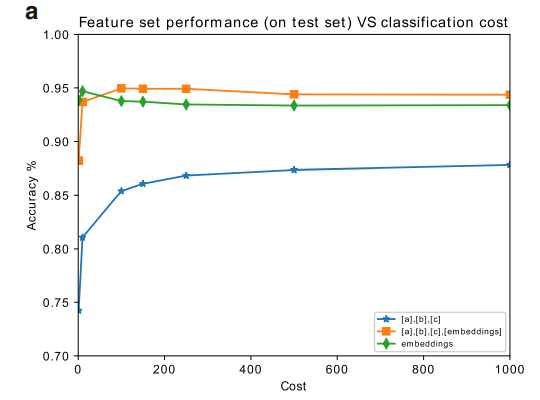
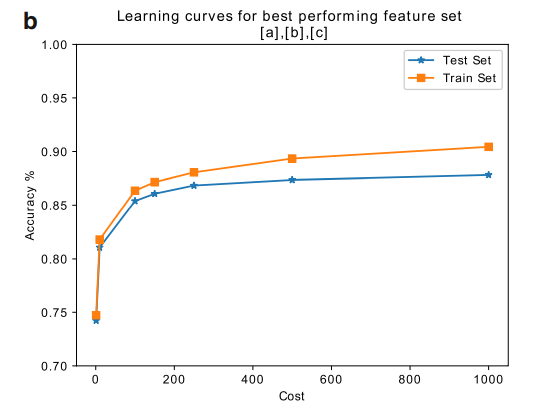


Using the SVM classifier with a linear kernel using 70% of the data and tests the rest 30%. From the figure above, the reader can notice that the combination of a& b& c features achieves the best accuracy compared to the other combinations as presented in the table above. To use the optimum cost for the best feature set and avoid overfitting, the researchers of the paper plotted the learning curve as shown in figure x. b. Also, to balance between bias and variance, they chose to resume using Friedman test with a cost value = 250. With regard to the mean rank by using the Friedman test.

Experimenting [a][b][c] combination with word embeddings:

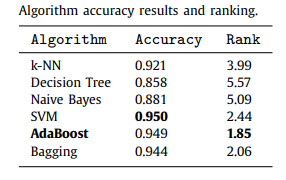
The next step is the use of Word embedding, specifically Word2vec. They started with taking into account a pre-trained model (Google News corpus) in order to extract 300-word embeddings vector for each article. Then, they checked the performance with word embeddings alone and word embeddings with the three combinations of linguistic features and noticed that the [a][b][c] linguistic feature combination with embedding gives the best accuracy as presented in the table and figure below.

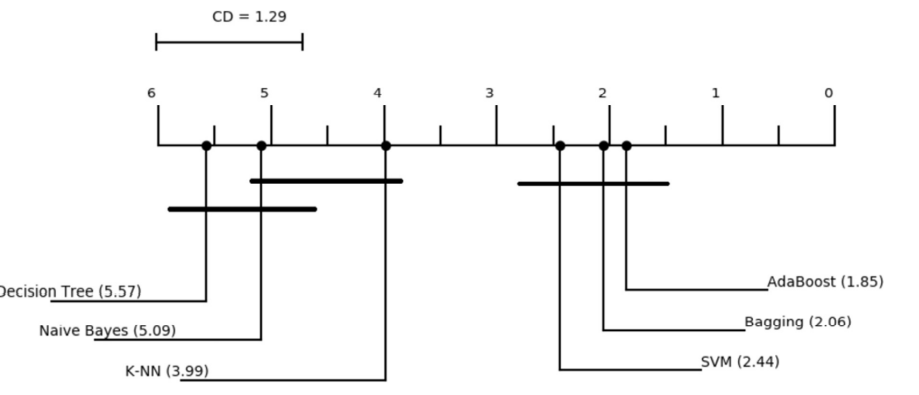


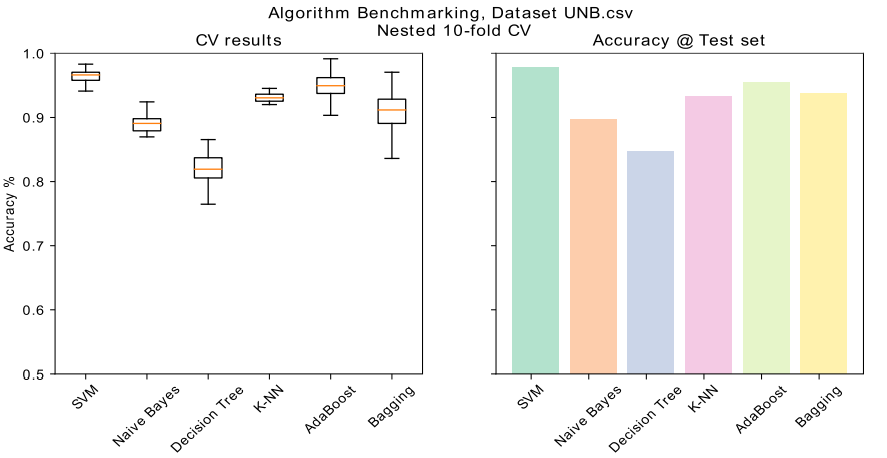
Experimenting the six Machine Learning algorithms:

The final step is that they experimented K-NN, Decision Tree, Naïve Bayes, SVM, AdaBoost, and Bagging for deciding which model is the best for detecting fake news with the dataset adjusted to [a][b][c] lingustic feature combination and word embeddings enhancement. SVM is the best in terms of accuracy.





Finally, the paper tried to prove with experimenting it on SVM, bagging, and Adaboost algorithms that Politilifact, Kaggle-EXT, and BuzzFeed do not have all unbaised features, and have dire influence on the results as seen in the figure above. Finally, with the UNB dataset, the results of test sets using the six machine learning algorithms are shown on the figure below.



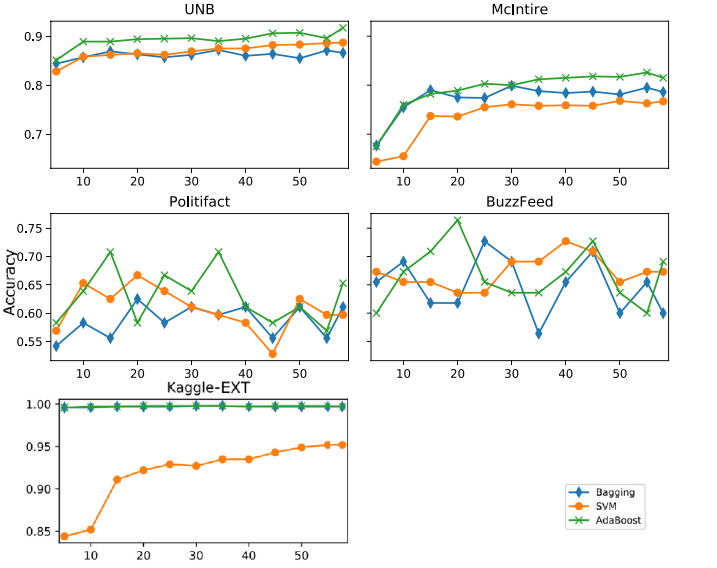


Fig. 9. Accuracy per algorithm and dataset based on feature selection by mutual information. The horizontal axis corresponds to the number selected features.

From

Datasets: