Machine Learning Endsem Project Report
FAKE AND REAL NEWS DETECTION

Anshak Goel
2020283

Vibhor Agarwal
2020349

Sahil Goyal
2020326

Deeptorshi Mondal
2020294

Motivation (Abstract)

In the recent years of information transfer we have seen how major shortcomings in the field of technology have affected the lives of the people. The times of social media has catalyzed the process of propagating a lot of fake news from anti-social elements all across the world. We as a group want to solve this problem by applying the concepts of Machine Learning learnt in the class and get a result which enables us to solve the problem. We thought about this project by first interacting with many of our friends and peers which made us aware of the issue. One can also see how in the past as well as in the present there has been chaos which has also led to loss of human lives due to transfer of incorrect news in places. So all in all our group wants to contribute in ensuring peace and sanity by identifying which is fake and real news through the use of Machine Learning Concepts.

1. INTRODUCTION TO THE PROBLEM STATEMENT

Though, technology has been the reason for the recent positive developments in the human history it also has had its fair share of disadvantages too. One can see that there was a time when we had to search books for gathering information or maybe read newspapers for reading news but now people have both information and news in their pockets in the form of mobile phones. With regards to news which comes from various sectors in the form of social media, digital news etc. people tend to rely on certain things which are not true. This results in the propagation of information which is wrong. This is happening extensively nowadays due to bias with which journalists are reporting incidents due to their involvement of a particular political organization. Just recently we saw how there were riots in India due to circulation of news where a person belonging from a certain community was accused killing someone from the other communities. We have also observed how political parties instigate the public by using their IT cell networks to hide truth as well as polarize the voters. So all in all our problem statement is to analyze news that we get from social media and label them as fake or real. In this project we will get the data in form of paragraphs and would see which person has spoken the words from which political party. Depending on the sentiment analysis which would be handled by applying Natural Language Processing where we would be able to analyze the text and convert it into numeric data which would help us to apply our algorithms properly. Several machine learning models have already been applied to solve this problem. But of the people have tried to test the models on particular datasets thereby inducing dataset biases. So if a particular algorithm works on a specific type of dataset it may work poorly on the other. In our case we preprocess the data and train six different models which are logistic regression, Naïve Bayes, Decision Tree, Random Forrest, Neural Networks, Support Vector Machine and ADABoost Classifier. Here Naïve Bayes serves us the threshold level of accuracy and we accordingly work on the other models. The neural Networks was involved in the application of Deep Learning. It helps in the better learning of the data which enables us to give better accuracies then the models used earlier. While using Neural Networks we have used to algorithms which are used for better data visualization. The algorithms are respectively PCA and t-SNE which helps in reducing dimensionalities thereby helping in noise filtering and feature extraction. Now getting insights become easier for us hence enabling us to get proper results. The best part of the algorithms is that preserves the global data even after preserving the data. Feature extractions have been done which has enabled us to form 4 types where we consider the speaker and the party and then further divide it on the basis of the vectorization techniques like BOW(Bag of Words) and TF-IDF which involves converting the text into numeric vectors and the features that we have.

2. LITERATURE REVIEW

Before going into making the code for the following data it becomes very important to search about how research has been done in the field that we want to work upon. We

have analysed quite a few papers that had done work upon fake news detection. Many types of model were trained which had many issues and had obtained many results which provided a lot of help in our project. The researchers have applied a lot of algorithms ranging from linear regression to deep learning algorithms. All the papers have first argued about how fake news has been troubling the world since a long time which has resulted in a lot of chaos including death in many cases. They have talked about the importance of classification of such news and how it becomes important to remove such propaganda to prevent treating misinformation as news. The research papers themselves have analysed several papers before proceeding on with their project to get an idea what goes wrong and how to add novelty to their project. Their papers discussed how the conversion of text to numeric values have been done where different methods of vectorization techniques have been used ranging from TF-IDF to Bag of Words(BOW). The process of data cleaning has been discussed where how data has been made into a proper dataset has been discussed and the use of NLP has been shown. Different type of algorithms has been discussed by their research papers like SVM, Random Forrest to name a few. The use of deep learning algorithms like CNN have been shown and the final accuracies have been shown where importance to data classification has been given. Now coming to the research papers, we can observe most of them have picked the dataset from LIAR dataset. Some other datasets are also included for example combined corpus by Junaed Younus Khan , Md. Tawkat Islam Khondaker , Anindya Iqbal and Sadia Afroz. There has been a proper classification that has been done for the type of data they are getting for example visual based and user based. This has been discussed in detail by Syed Ishfaq Manzoor, Dr Jimmy Singla and Dr Nikita in their research paper. For data cleaning different methods have been employed to remove all the unnecessary IP and the URL addresses. Whitespaces have been removed using stemming. TFI-IDF has been used extensively for the vectorization techniques by most of the papers. The above two works have been done by Junaed Younus Khan. BOW has been used in the research paper by Dr Singla. Another important point about data pointed out in the research papers was the issue of bias in data aligning with the models. Next all the 3 research papers have done the feature extraction where empath tool has been used for classifying the type of news as violent, misleading etc. Another important method used here is Lexical and Sentiment Feature extraction has been done where word count, word length has been used as lexical while positive and negative has been marked as lexical. This works also has been done by the research paper made by Junaed Younus Khan. Next traditional models have been used such as SVM, Linear Regression, Decision Trees, Naïve Bayes and K-NN model by professors at Dhaka University. XG Boost and Random Forrest were the new algorithms which were implemented by professors at LPU. The paper made by Harsh Khatter argued about SVM being used to solve the problem and proposed a model combining of News Aggregator, Authenticator and Suggestion recommendation. Further deep learning algorithms have been implemented for the better learning of the data so that better accuracies are obtained. The paper by Dr Khatter implemented simple neural Networks for the same while the paper by Anindya Iqbal discussed about the CNN model and used several new deep learning algorithms like Hierarchical Attention Networks(HAN) and Convolutional HAN. Three types of LSTM were also used which includes LSTM,C-LSTM and Bi-LSTM. LSTM is basically Linguistic Inquiry and Word Count (LIWC) dictionary which includes a word classification and count tool. The results were divided into two parts by professors at Dhaka University were one analysed before the neural networks while the other talked about after that. The best accuracy was reported by Naïve Bayes with 94 percent after using n-gram (bigram TF-IDF) features. For the paper by Harsh Khatter it reported Naïve Bayes to be the best with a accuracy of 93.5 percent and the paper by professors at LPU argued about XG Boost being the best. In conclusion all papers argued that perfect accuracy cannot be obtained and scope of future work was there.

3. DATASET WITH PRE-PROCESSING TECHNIQUES

Here we have taken the data from the Liar dataset for fact-checking and fake news detection in our paper. This data has evidence sentences which are straight extracted from report written by journalists in Politifact.

3.1. DATA DESCRIPTION

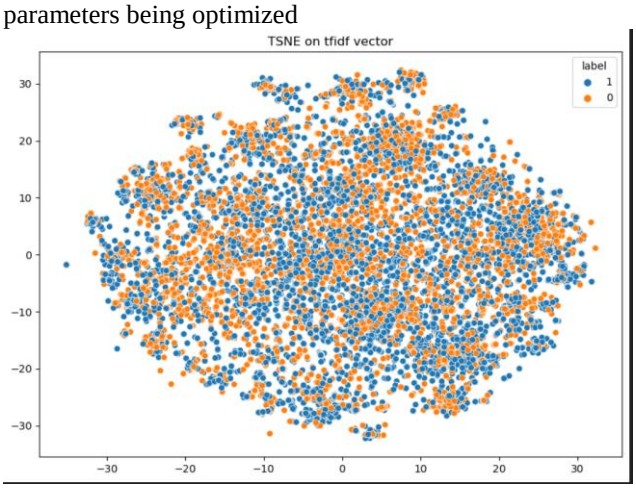
The dataset had 16 columns namely index, the ID of the statement([ID].json), label, statement, subject, speaker, speaker's job title,, the state info, party affiliation, the total credit history account, including current statement(comprises of 5 columns together which are barely true counts, false counts, half true counts, mostly true counts, pants on fire counts), context(venue /location of speech statement) and extracted justification. There are a total of 12788 rows. There are 10239 rows for testing,1266 rows for testing and 1283 rows for validation

3.2. DATA CLEANING

We have made several changes to the dataset so that our model is easy to train and also gives accurate results. We have merged the data such that it becomes one type. Next we have dropped some columns as well which are index, id of the statement, state info, context, justification, barely true counts, false counts, half true counts, mostly true counts, pants on fire counts. We then converted our data into binary classification where true, half-true, mostly-true were treated as real (1) while false, pants-false, barely true were treated as fake (0). Next we merged the statement and subject column into one for better analysis. Wherever missing values were there we dropped them, next we converted them into lowercase. So now our total rows are 12786 while total columns are 4. A heat map has also been made to show the correlation between the different feature about the data given to us. This allows us to remove a certain amount of data which has high correlation. A scatter was also made to help us categorise the 6 classification into a two class classification.

3.3. DATA PREPROCESSING [USE OF NLP]

As our data is in the form of text, we need to convert it in the form of numerical data and vectorization. For doing so we will take the help of natural language processing. So first we need to refine the data for actually converting the text to numbers. We first remove all the punctuation marks, links and extra white spaces except the commas by normal methods. Next, we do our first NLP where we tokenize the data. For tokenization we use the library as it is where it's work is to split paragraphs and sentences into smaller units giving it an actual meaning. We have lemmatized the text afterwards which basically means to switch any kind of word to its root node, basically grouping words having the same meaning. This is done by cutting down the suffices. We have removed the stop words as well example "and"," the" which do not really add any meaning to the sentence. After that we have joined the text column into string for vectorization while applying the NLP algorithms which would actually covert the text in to numerical value. We have visualized the data as well by word-cloud denoting the frequencies. Our next job is to use the NLP algorithms BOW(Bag of Words) and TF-IDF(term frequency inverse document frequency). These algorithms basically help in vectorization of the data thereby converting text to numeric data. In BOW each key in the dictionary is set to word and the value is set to the number of times the word appears, while for TF-IDF it quantifies the importance of relevance of the string representation (lemma, phrases, words) in a document among a collection of documents. After that we take the help of label encoder which converts label into numeric form thereby converting it into machine readable form. We concatenate the data available into four parts where no speaker and no party is there and we are using either tfidf and bow separately. In the next we take both the speaker and the part he belongs and apply both the NLP algorithms. Next Grid search would take care of the hyper parameters being optimized



4. METHODOLOGY AND MODEL DETAILS

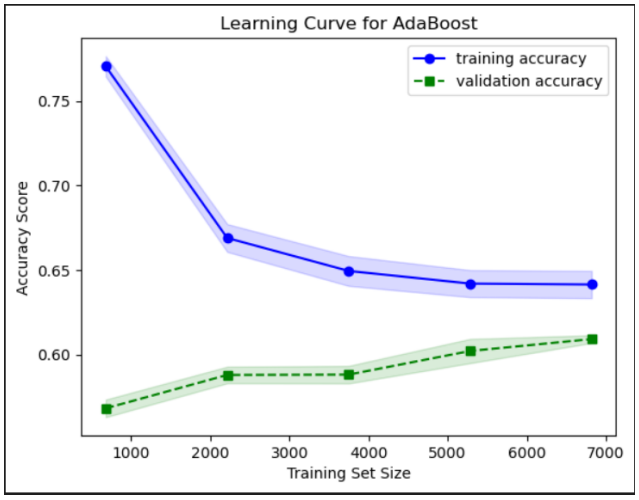
4.1. METHODOLOGY

The methodology used in determining whether the data is a real or not is a combination of data cleaning, data processing. NLP and the different algorithms that we are applying to get the best accuracy. In data cleaning as written above we saw that we dropped certain columns and rows, tinkered with the classification as well. We merged the statements as well. We did data pre-processing which include tokenizing, removing stop words and lemmatization. Next we applied TF-IFD and BOW to the data for converting it into numeric data. Next we apply PCA and t-SNE are unsupervised linear dimensionality reduction and data visualization technique for very high dimensional data. They helped in reducing dimensions which allowed us to gaining insights from it. The main methodology now applied is that we have divided the whole data into four parts where in 1 part we remove both the speaker and the party he belongs form and then once apply TF-IFD and then again apply BOW for the vectorization to numeric data. We will apply the same above mentioned algorithms to the case when both the speaker and his/her party is also there. So we have 4 cases where we would apply our algorithms and check for ourselves the accuracy as well. After applying the algorithms, we have a table with data being put in a tabular form such that a table is formed with columns containing the numeric data and the five features as label. Next we have used the label encoders to make sure that we can assign a number to the feature that we are using. We have used Grid search as well which helps us to estimate the

hyper-parameters to the optimal best. The grid search brings out the best parameters for the algorithms we are using. For example, for logistic regression it is the learning rate while for a decision tree it is the depth which would improve our accuracy as well. We have also applied deep learning algorithms using neural networks so that the accuracies improve. t-SNE and PCA have also been applied such that the dimensions are reduced. At last learning curves have been plotted to analyse the accuracies.

4.2. MODEL DETAILS

When the above methodology is done where we divided our data set consisting of testing, validation and training into the above four cases after data pre-processing we will now apply our machine learning algorithms. For all the four cases we would apply Logistic Regression, Naïve Bayes, Decision Tree and Random Forest. Our job is to observe the frequencies and analyze the accuracies accordingly. So for Logistic regression it basically is used to solve binary classification problems. It is nothing but a statistical model which models the probability of an event by applying logit function and the event is a combination of more than one independent variables. Basically it is used to estimate the parameters of a logistic model. Next we have used Naïve Bayes which is basically based on the concept of probability which makes strong independent assumptions between features. It is nothing but calculating the probability of our aim given there is some event that has happened. After this we have used decision tree to analyze the data which mainly deals with the fact that whether it is classifying properly or not. It is nothing but a flow chart where internal node denotes the test data, each branch denotes outcome while the leaf node denotes the class label. The data travels through the leaf nodes till the whole iteration is not complete. Next we have used Random Forrest which is the ensemble learning for classification and regression that works by constructing a multitude of trees at the same time. For classification the tree chosen the maximum number of times a class has been selected by most trees while for regression the average of the output given by the trees is given. ADABoost Classifier has also been used. AdaBoost also called Adaptive Boosting is a technique in Machine Learning used as an Ensemble Method. The most common algorithm used with AdaBoost is decision trees with one level that means with Decision trees with only 1 split. These trees are also called Decision Stumps.We have next used support vector machine to further find out if we could improve our accuracies for our model. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. At last we have tried to incorporate deep learning for better performance of our model where we have used neural networks through the use of MLP classifier. The MLP classifiers that have been used also have gone through PCA and t-SNE algorithms which has enabled us to reduce the amount of dimensions we got while applying TF-IDF and BOW, so that better learning of data is done and getting insights become easier by removing the highly co-related data. While applying all the algorithms we have made sure that the grid search algorithm is also applied through all the models such that the best parameters are obtained for better accuracies. At last the learning curves have been plotted for each and every algorithm with respect to different algorithms. While all algorithms have been plotted w.r.t their dataset and accuracies only the MLP algorithm was plotted w.r.t its loss. The learning curves indicate the difference between the training and validation data thereby informing us about the variance and bias of our data.



5. RESULTS AND ANALYSIS

LR-Logistic Regression, NB-Naïve Bayes, DT-Decision Tree, RF-Random Forrest

5.1. RESULTS

Accuracy Scores for every model				
Model	Tfidf without speaker and party	Bow without speaker and party	Tfidf with speaker and party	Bow with speaker and party
Naïve Bayes	58.09	54.65	60.04	54.41
Logistic Regression	59.57	58.405	60.98	59.34

Decision Tree	56.91	57.23	58.95	59.89
Random Forest	59.57	60.59	60.906	62.93
Adaboost Classifier	57.38	58.17	60.43	60.75
SVM	59.34	58.405	58.32	58.09
MLP Classifier	58.95	59.65	57.38	59.42
MLP with PCA	57.70	59.42	57.93	58.56
MLP with TSNE	55.27	56.91	54.73	59.34

5.2. ANALYSIS

One can observe that the accuracy is highest in Random Forrest where we have used bow as the method of NLP for vectorization of the textual data and then applied the algorithm of Random Algorithm. Also both the speaker as well as its political party has been considered. The Accuracy is coming out to be 62.93 percent. Accuracy determines how correct our model is. In fact, the precision, recall and the F1-score is highest for Logistic Regression applying Bow taking into consideration the speaker as well as the party. So one can infer that after providing more data in the form of the speaker and the party he belongs to we get better results on our model while using bow also gives us an upper hand rather using TF-IFD. Also we have seen for each case it is not necessary that only one model dominates. So you never know which model dominates in which situation. But fake news is a very complex model and even when ML experts try to incorporate their own models they do not obtain very high accuracies which calls for the scope of improvement. Another interesting result was when algorithms were used for t-SNE and PCA were used in reducing the dimensionalities it reduced the accuracy score for the neural Networks part. By using the learning curves we obtained the main results which tells us that in Random Forest data is overfitting on the training set. But in the other models as the data is increased the overfitting issue is removed. The Gini criteria dominates the entropy criteria in the Random Forrest part which gave us the best accuracy.

6. CONCLUSION

6.1. LEARNING FROM THE PROJECT

Our main learning from the project was it is not possible to obtain a high accuracy in the field of fake news. Though one can clearly argue that the models are much better performing than the humans making random guesses but still it cannot fully tell whether the news is fake or not. We could also conclude how data defining becomes important in order to successfully train the data set and obtain a desirable result. We see how logistic regression dominated even though algorithms like random forrest were expected to perform better. Even deep learning algorithms like Neural Networks obtained a less accuracy than the simple logistic regression model. Random Forrest was the one which gave us the maximum accuracy of 62.93 percent where the speaker and vector and BOW vectorization was done. More the sentiment more the accuracy like we observed as soon as we provided the data for the person and the political party he belongs to, immediately the accuracy rose. But in some models this did not occur like the MLP Classifier and SVM. In Decision tree the bias and variance is high compared to other models. More data should have been given. We used the Grid search algorithm which also has the ability to optimize the hyper parameters like the learning rate or the depth of the decision algorithm. We saw the NLP algorithms in action and how tokenization, lemmatization and stop words were important to refine the data. One could also observe that it is not necessary that each model works for a different topic of fake news because generally we make the algorithms according to our needs. Even after applying deep learning algorithms and different models with standard NLP models we could not make the perfect model signifying that scope of improvement is still there.

6.2 WORK DIVISION THROUGHOUT THE PROJECT

We maintained proper work distribution for achieving the best possible results in the project in the most efficient way. Anshak Goel took care of the NLP part of the project which helps us to convert the fake news English sentences to data which could actually be given to a ML model. He also ensured that the dataset is properly tested and suffices the scope of the project. His work was also involved in doing the data processing as well. Sahil Goyal took care of cleaning and optimizing the data for better performance in the ML models. He also oversaw the preprocessing part and took the decision regarding what attributes to choose and handled the visualization part. He was also actively involved in implementing models and data preprocessing. Vibhor Agarwal is responsible for choosing which model to use depending on the accuracy and effectiveness of the model. He also took care of the data analysis. Deeptorshi Mondal handled the documentation part of the project and also helped in the NLP part. He tracked the progress of the project and ensured that it is properly documented in the final project report which he also wrote. He also helped in the data analysis part of the project. Lastly we learnt a lot about the models and Natural Processing Algorithm and then implement them.

6.3 TIMELINE OF THE PROJECT

We ensured that we proceed with the project in a well-planned manner and stuck with the timeline that we submitted in the initial proposal. In the first week starting from 19th September, we decided upon the dataset and did some data cleaning. We also decided which attributes to use. Next we discussed and implemented the NLP part of the project which is crucial for good results when applying ML models. In the subsequent weeks we implemented the ML models like Logistic Regression, Naïve Bayes, Decision Tree, and Random Forest. We spent the next weeks learning about new models like SVM so that we could implement the code which would enable us to improve our accuracies. Deep Learning algorithm which involved using neural networks(MLP) to improve the accuracies. The accuracy was observed by using TSNE and PCA. We saw the difference without using TSNE and PCA and with using them which enabled us to reduce the columns, so that we could improve the learning rate and gain insights about it. The documentation of the project was done in parallel with the work done so we don't miss out on any necessary details. The complete analysis was done about the models and also identified the future work and sorted out the limitations.

6.4 WORK TO BE DONE IN THE FUTURE

One potential direction for a university undergraduate machine learning project involving fake news detection could be to explore the use of natural language processing (NLP) techniques to automatically identify fake news articles. This could involve training a machine learning model on a dataset of known fake and real news articles, and then using the model to make predictions on new, unseen articles. The model could be evaluated using various metrics, such as accuracy and precision, to determine its effectiveness at detecting fake news. Another potential direction for the project could be to focus on the development of an interactive system that allows users to input a news article and receive feedback on its credibility. This could involve the use of natural language understanding (NLU) techniques to analyze the content of the article and provide users with a credibility score or other indicators of its veracity. The system could also incorporate other features, such as the ability to flag potentially fake news articles for further review by human experts. In the near future we could also take care of the visual and audio news so that the scope can be increased. The dataset bias should also be handled as models can be made inclined to a certain dataset. So next time we will also use different datasets in the near future

7. REFERENCES

1. https://github.com/manideep2510/siamese-BERT-fake-news-detection-LIAR
2. https://paperswithcode.com/paper/liar-liar-pants-on-fire-a-new-benchmark
3. https://arxiv.org/pdf/1705.00648.pdf%E2%80%8B https://www.overleaf.com/latex/templates/cvpr-2022-author-kit/qbmjsdxryffn
4. https://www.researchgate.net/publication/336436870_Fake_News_Detection_Using_Machine_Learning_approaches_A_systematic_Review
5. https://www.researchgate.net/profile/Harsh-Khatter/publication/339022255_A_smart_System_for_Fake_News_Detection_Using_Machine_Learning/links/5f9112b0299bf1b53e3a2aa1/A-smart-System-for-Fake-News-Detection-Using-Machine-Learning.pdf
6. https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/
7. https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html
8. https://www.geeksforgeeks.org/nlp-how-tokenizing-text-sentence-words-works/#:~:text=Tokenization%20is%20the%20process%20of,a%20token%20in%20a%20paragraph.&text=How%20sent_tokenize%20works%20%3F,of%20PunktSentenceTokenizer%20from%20the%20nltk.
9. https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc
10. https://www.javatpoint.com/machine-learning-ran
11. dom-forest-algorithm

8. GITHUB LINK

https://github.com/Anshak-Goel/Machine-Learning-Project