CS725: Foundations of Machine Learning (Project Report)
Fake News Detection

Niteen Pawar(22M0800)
Tejas Rathod(22M0801)
Nikhil Solanki(22M0814)
Aniket Jadhav(22M0817)

1 Abstract

False information created to mislead readers is referred to as fake news. It is now quite simple to produce fake news, which can alarm the public and mislead them. As a result, the study of fake news identification is becoming more popular. This study seeks to identify the most effective algorithms for spotting fake news as a solution. In this study, text preparation is done separately using term frequency-inverse document frequency (TFIDF).
Using evaluation measures like accuracy and F1 score, four machine learning algorithms passive aggressive classifier (PAC), random forest (RF), logistic regression (LR), and support vector machine (SVM) are evaluated. The most effective algorithms for the dataset under consideration are Random Forest and SVM.

2 Keyword

Fake news Supervised learning,Online learning,Passive-aggressive classifier,TF-IDF,

3 Introduction

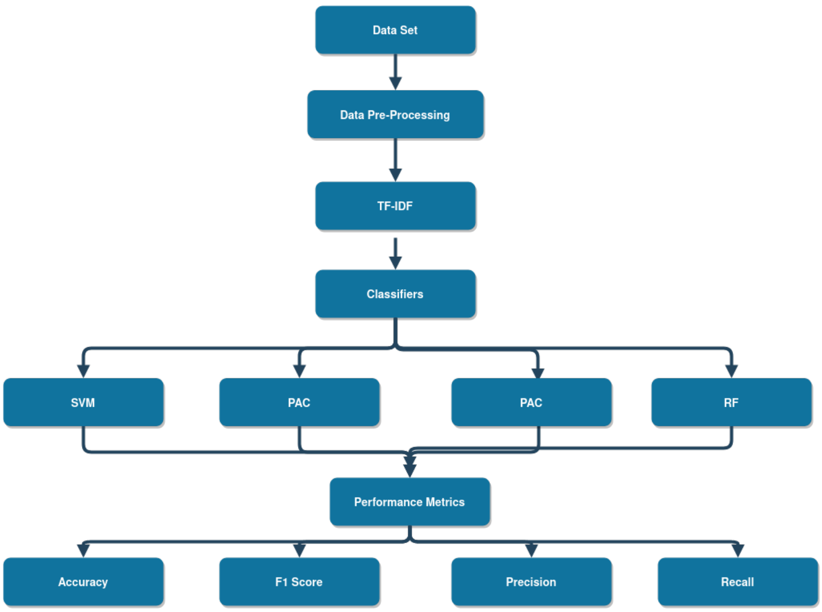
The effects of fake news on contemporary culture are profound. Finding fake news is a crucial step. This study suggests using machine learning methods to identify fake news. It is unknown how much of the material shared on social media is accurate. When an incident occurs, a large number of people use social networks to chat in the cloud . If this information were accurate, the concept of going too far is really straightforward because spreading false information can cause confusion and using lies to influence another person's behaviour may have long-term effects.[1] Additionally, anyone has the option of sharing any content they like. People are exposed to fake news, hoaxes, rumors, conspiracy theories, and misleading news through unreliable sources of information.

Unpredictable news includes some types, including tragic natural disasters or climatic changes. In addition to misleading news being aired when unexpected events occur, the events themselves cause bewilderment. The majority of people trust the information that comes from reliable acquaintances or family members. Fake news originates from inaccurate information, misunderstandings, or incredible information from a reliable source. When they acquire the news information, these find it difficult to decide whether to believe it or not.

Many scientists think that artificial intelligence and machine learning can be used to address the problem of fake news . There is a reason for this; lately, artificial intelligence algorithms have begun to perform significantly better on many classification issues (image recognition, voice detection, and so on) due to cheaper technology and more readily accessible larger datasets.[2]

We suggest strategies to identify bogus news in this research. We suggest a variety of approaches, focusing on SVM, Passive Aggressive Classifier, Logistic Classifier, and Random Forest.

4 Methodology



4.1 Dataset

For our project we have used a dataset which has 20800 samples, three input features title, author and text and one output feature which is the label.

- ID is a unique identifier to uniquely identify each instance.
- Title is the headlines of the news.
- Text is the content of the news.
- Label is the class assigned to the instance as real or fake.

By deleting the null values, data cleaning is completed first. The 'text' input element is the most pertinent to our work among the three input options because it is only through the content of the news that we can determine if it is true or false. It falls into the binary category. The goal feature has two classes, real and fake, with equal priorities. News is labelled as true or false depending on whether it is actual or false. The dataset is divided into 20% for testing and 80% for training.

4.2 Text Preprocessing Techniques

TF IDF is used for data preprocessing TF-IDF stands for **TF (term frequency)** and **IDF (inverse document frequency)**.This is a technique to vectorise words in a set of documents. We generally compute a score for each word to signify its importance in the document.

tf = (Frequency of a word in the document)/(Count of words in the document)
idf = log(Count of documents/Count of documents possessing that word) + 1

Finally, by taking a multiplicative value of TF and IDF, we get the TF-IDF score.

tf-idf = tf * log(idf)

As the name suggests, TF-IDF determines values for each word in a document by dividing the frequency of a word in a given document by the proportion of documents in which the word appears. High TF-IDF words signal a close connection to the document they occur in, indicating that if the word appeared in a query, the user could be interested in the document. We offer proof that this straightforward approach successfully classifies pertinent phrases that can improve query retrieval.[3]

4.3 Classifiers

All of the models that will be used in our project are covered in this section.

4.3.1 Passive Aggressive Classifier

The Passive Aggressive Classifier(PAC) is an online learning algorithm which is used in case of social media sites where data keeps on flowing 24*7 and we don't need to store the data to make correct classification on later stages.
When the right categorization is made, the classifier is passive; otherwise, the rule becomes quite aggressive. It searches for the new weight that matches the L value and is

closest to the prior weight.The PAC is very similar to *Perceptron* and *SVM* except that PAC is faster than SVM and has a different weight update rule as compared to the two.

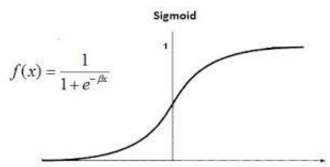
$$\bar{W}_{t+1} = \bar{w}_t + \frac{\max\left(0, 1 - \mathrm{y}t\left(\bar{w}^T \cdot \bar{x}_t\right)\right)}{\left\|x_t\right\|^2 + \frac{1}{2C}} y_t \bar{x}_t[4]$$

Working

intialize W=(0,0,0...0)
receive a new text x=(x_1,x_2,x_3...x_n)
apply tf-idf,normalize ||X||=1
predict positive if x^Tw > 0
observe true class: y=±1
want to have:
x^Tw ≥ +1 if positive (y=+1)
x^Tw ≤ -1 if negative (y=-1)
same as: y(x^Tw)≥ 1
loss L=max(0,1-y(x^Tw))
update w_{new}=w + yLx

4.3.2 Logistic Regression

Logistic regression is a classification algorithm used when we want to classify labels in the form of (0/1) eg- True/False in the case of Fake news detection. Logistic regression is less likely to overfit on the data, but we have ridge and lasso regularizers to solve the overfitting problem if it does also generates great accuracy's for linearly seperated data It uses sigmoid function which outputs probability and input can be -∞ to +∞



[5]

Loss function in case of Logistic regression is as follows

$$Loss = \sum_{i=1}^{N} (y_i Log(p_k) + (1 - y_i) Log(1 - p_k))$$

We used imports like TF-IDF mentioned above for the code also linear_model We checked the accuracy for different features in the dataset as to which feature combination gave higher accuracy. For single individual feature like (author,title,text) accuracy as high as 91% , But for two features namely (title of news + author) accuracy of 95% was achieved . But When used against all information i.e all three features in the dataset ,

accuracy was 93%
Changing the regularizer from L2 to L1 does not significantly improve the accuracy on the dataset. F1 score by this algorithm was around 97% which was good as compared to all algorithms used in this project [6].

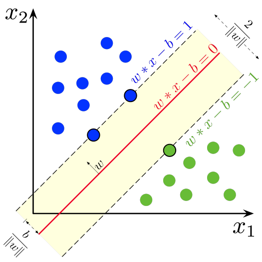
4.3.3 SVM

Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. However, it is mostly used in classification problems. Finding a hyperplane in an **N-dimensional** space (N is the number of features) that categorises the data points clearly is the goal of the support vector machine algorithm.[**?**]he decision boundary created by SVMs is called the maximum margin classifier or the maximum margin hyper plane.

$argmin(w^*, b^*) \left\| w \right\| /2 + c \sum_{i=1}^{n} \zeta i$

$_1$ **How an SVM works**
Making a straight line between two classes is how a straightforward linear SVM classifier functions. In other words, the data points on one side of the line will all be assigned to one category, while the data points on the other side of the line will be assigned to a different category. This implies that the number of possible lines is unlimited.

Because it selects the optimal line to categorise your data points, the linear SVM algorithm is superior to several other algorithms like k-nearest neighbours. It selects the line that divides the data and is as far from the nearest data points as it may be.



[7]

A mathematical model containing a number of parameters that must be learned from the data is referred to as a machine learning model. Hyperparameters, on the other hand, are those parameters that cannot be learned directly. Prior to starting training, humans frequently choose them based on some intuition or trial and error. By enhancing the model's functionality, such as increasing the model's complexity or learning rate, these parameters demonstrate their significance. Models may have a large number of hyper-parameters, and determining the ideal set of parameters can be approached as a search issue.
We did Hyperparamenter tuning by change values of C,tolerance and changing max feature in TF-IDF.

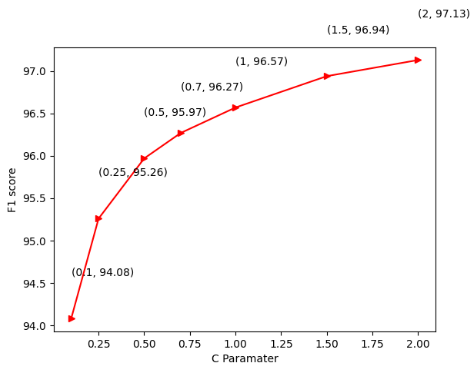
- tol is the tolerance for the stopping criteria.Basically it tells Svm to stop when difference in error in nth and n+1 iterations reaches tolerance value it stops.

- C value is used regularisation,The C parameter indicates to the SVM optimizer how much misclassification of each training example is to be avoided. For large values of C, the optimization will pick a smaller-margin hyperplane if it performs better at accurately classifying every training point. The optimizer will search for a larger-margin separating hyperplane even if it misclassifies more points if C is set to a relatively small value, on the other hand. Even if your training data can be linearly separated, you should frequently encounter misclassified cases for extremely low values of C.

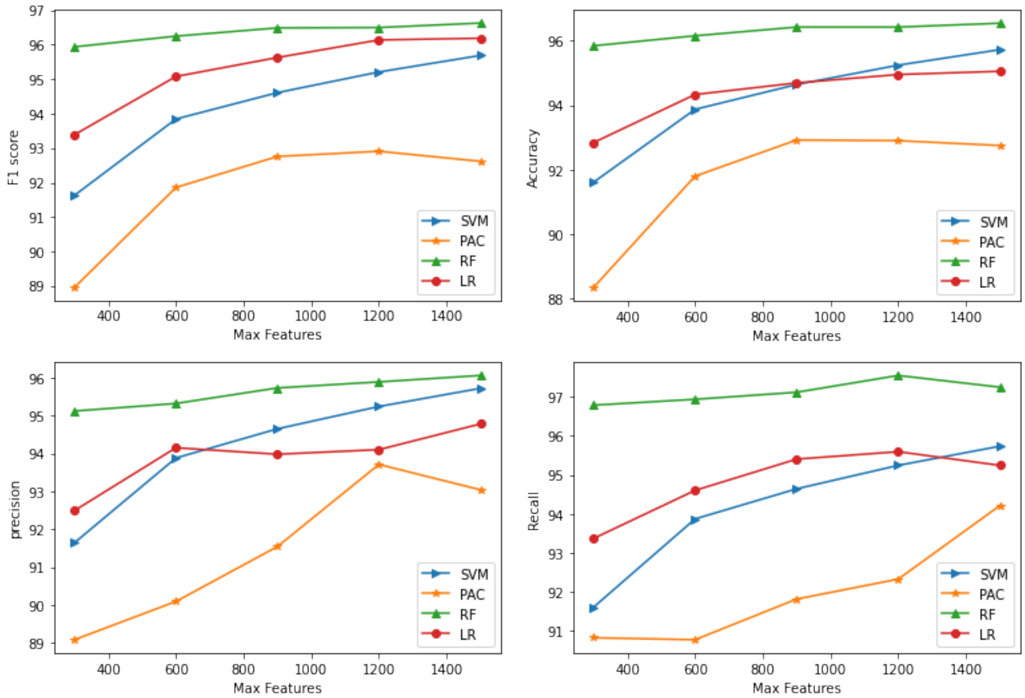
4.3.4 Random Forest

Like its name suggests, a random forest is made up of numerous independent decision trees that work together as an ensemble. Every tree in the random forest spits out a class forecast, and the classification that receives the most votes becomes the prediction made by our model.
The fundamental concept behind random forest is a simple but powerful one - the wisdom of crowds. The best results come from a large number of very uncorrelated models (trees) working together as a committee. The key is the low correlation between models. Uncorrelated models have the ability to generate ensemble forecasts that are more precise than any single prediction. The fact that the trees shield one another from each other's mistakes accounts for this lovely effect. Many trees will be right while some may be wrong, allowing the group of trees to move in the proper direction.[8]
Decisions Related to Bagging (Bootstrap Aggregation) Because of how sensitive trees are to the data they are trained on, even minor adjustments to the training set can produce dramatically different tree architectures. By enabling each individual tree to randomly sample from the dataset with replacement and produce various trees as a consequence, random forest takes advantage of this. This procedure is referred to as bagging.

5 Result



In SVM changing Hyperparameter C and calculating corresponding F1 score



Model	F1 Score	Accuracy	Precision	Recall
PAC	92.62	92.75	93.04	92.21
SVM	95.69	95.72	95.72	95.73
LR	97.16	93.26	93.08	93.54
RF	96.63	96.56	96.06	97.54

The performance of several algorithms on our pre-processed dataset is shown in the following table, and we can compare the algorithms using various criteria.

6 Conclusion

Spreading information that isn't accurate to its source is known as fake news. In some instances, they might be deceptive and have negative effects. Knowing how to spot fake news is crucial because of this. This study compares various machine learning algorithms, including as PAC, RF, LR, and SVM, in an effort to find the best algorithms to identify bogus news.

Our dataset's preparation steps produce better outcomes. The accuracy of the classification was significantly improved as a result of these actions.We have used a TF-IDF vectorizer for pre-processing which takes into consideration the importance of words.As we know that F1 Score is a good metric to classify different algorithm and using F1 score we can say that Support Vector Machine(SVM) and Random Forest are giving us the best score.

From the results we can see that most of the algorithms give an accuracy of more than 90% but out of all of them PAC gives the least scores. This is happening because it is not getting trained enough which was concluded when we increased the training data to 90% from intital 80% and started getting an accuracy of more than 95%. Whereas, Logistic gives highest F1-score among all the algorithms and also has good accuracy of 93%. So we can conclude that the the best algorithms for this dataset are SVM ,Logistic Regression and Random Forest.

References

[1] K. Nagashri and J. Sangeetha, "Fake news detection using passive-aggressive classifier and other machine learning algorithms," in *Advances in Computing and Network Communications* (S. M. Thampi, E. Gelenbe, M. Atiquzzaman, V. Chaudhary, and K.-C. Li, eds.), (Singapore), pp. 221–233, Springer Singapore, 2021.

[2] M. Granik and V. Mesyura, "Fake news detection using naive bayes classifier," in *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, pp. 900–903, 2017.

[3] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242, pp. 29–48, New Jersey, USA, 2003.

[4] R. R. Mandical, N. Mamatha, N. Shivakumar, R. Monica, and A. N. Krishna, "Identification of fake news using machine learning," in *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pp. 1–6, 2020.

[5] GeeksforGeeks.

[6] N. N. P. A. Muhammad Syahmi Mokhtar, Yusmadi Yah Jusoh, "Fakebuster: Fake news detection system using logistic regression technique in machine learning."

[7] Wikipedia contributors, "Support vector machine — Wikipedia, the free encyclopedia," 2022. [Online; accessed 25-November-2022].

[8] T. Yiu, "Understanding random forest," Sep 2021.

Link to Github repository

https://github.com/TejasN99/Fake-News-Detection