International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

FAKE NEWS DETECTION USING MACHINE LEARNING

[1]S.Fowzia Sultana, [2]Y.Indravathi, [3]Shaik Asma, [4]Shaik arziya, [5]Modala Harika
[1] Assistant Professor [2,3,4,5]B.Tech Scholar,
[1,2,3,4,5]Department Of Electronics And Communications Engineering
[1,2,3,4,5]G. Pullaiah College of Engineering and Technology, Nandikotkur Rd, near Venkayapalle, Pasupula Village, Kurnool, Andhra Pradesh 518002, India.

Abstract:

Recent political events have led to an increase in the popularity and spread of fake news.As demonstrated by the widespread effects of the large onset of fake news, humans areinconsistent if not outright poor detectors of fake news. With this, efforts have been madeto automate the process of fake news detection. The easy access and exponential growthof the information available on social media networks has made it to distinguish betweenfalse and true information. the goal of this project was to create a tool for the languagepatterns that characterize fake and real news through the use of machine learning andnaturallanguageprocessingtechniques.Thespreado ffakenewshasfar-reachingconsequences like the creation of biased opinions to swaying election outcomes for thebenefitofcertaincandidates. Moreover,spammers useappealingnewsheadlinestogenerate revenue using advertisements via click- baits. In this project, we aim to performbinary classification of various news articles available online with the help of conceptspertaining to Artificial Intelligence, Natural Language Processing and Machine Learning.We aim to provide the user with the ability to classify the news as fake or real and alsochecktheauthenticityofthewebsitepublishingthene ws.

Key words Online fake news, Machine learning, fake news, Text Classification, social media.

1. Introduction

Fake news detection has always been a problem because of its long-term repercussions and consequences. Its root can be traced back to the 17th century in propaganda, which became misinformation in the cold war [1]. In modern days, this problem has become grave due to the emergence of social media platforms. Specifically, in the past few years, social media channels, such as Facebook, Twitter, and Instagram, have emerged as platforms for quick dissemination and retrieval of information. Figure 1 shows a snapshot of some fake news in recent years. According to various studies [2], almost 50% of the population of developed nations depend on social media for news. The importance of social media cannot be denied, and it has emerged as an effective medium at the time of crises in regard to the role it plays in breaking news, for example [3]. However, one drawback of the convenience provided by social media is the quick dissemination of fake news.In contrast to conventional mediums such as print media or television, the content of social media can be modified by users, thereby enriching the content with their opinions or biases. This can alter the meaning or context of the news altogether [5]. According to various studies, social media is a fertile ground for quick sharing of information without fact checking [1]. Fake news can be defined as the creation or modification of news content by social media user to deliberately or non-deliberately change its apparent meaning or context, contaminating it with their opinion or biases, where the intent may be to jeopardize or harm a person, organization, or society, monetarily or morally. Examples of fake news are sarcasm, memes, fake advertisements, fake political statements, and rumors [3]. A fakester is a term used for a person responsible for spreading fake news. News can have various degrees based on its credibility, i.e., true, half-true, and false [5]. Fake news can be transmitted in the form of images, video, and text. The life cycle of fake news has been described in [6] as the creation, publication, and propagation of the news.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

2. Literature Review

In the world of rapidly increasing technology, information sharing has become aneasy task. There is no doubt that internet has made our lives easier and access to lots ofinformation. This is an evolution in human history, but at the same time it unfocusses thelinebetweentruemediaandmaliciouslyforgedmedia. Todayanyonecanpublishcontent credible or not that can be consumed by the world wide web. Sadly, fake newsaccumulates a great deal of attention over the internet, especially on social media. Peopleget deceived and don't think twice before circulating such mis-informative pieces to theworld.Thiskindofnewsvanishesbutnotwithoutdoingt heharmitintendedtocause.Thesocial media sites like Facebook, Twitter, WhatsApp play a major role in supplying thisfalse news. Many scientists believe that counterfeited news issue may be addressed bymeansofmachinelearningandartificialintelligence.

Existing System:

Shloka gilda presented concept approximately how NLP is relevant to stumble onfakeinformation.Theyhaveusedtimeperiodfrequency -inverserecordfrequency(TFIDF)ofbi-gramsandprobabilistic contextfreegrammar(PCFG)detection. Theyhaveexamined their dataset over more than one class algorithms to find out the great model.They locate that TFIDF of bi-grams fed right into a stochastic gradient descent modelidentifiesnon-credibleresourceswithanaccuracyof77%.Mykhailogr anik proposed simple technique for fake news detection the usage ofnaïve Bayes classifier. They used buzzfeed news for getting to know and trying out thenaïve Bayes classifier. The dataset is taken from facebook news publish and completedaccuracyupto 74%ontestset.Codybuntain advanced a method for automating fake news detection on twitter.They applied this method to twitter content sourced from buzzfeed's fake news Dataset.Furthermore, leveraging non-professional, crowdsourced peopleinstead ofJournalistspresents a beneficial and much less costly way to classify proper and fake Memories ontwitterrapidly.Marco L. Della offered a paper which allows us to recognize how social networksandgadgetstudying(ML)strategiesmaybeused forfauxnewsdetection.Theyhaveusednovel ML fake news detection method and carried out this approach inside a FacebookMessenger chatbot and established it with a actual-world application, acquiring a fakeinformationdetectionaccuracyof81%.ShivamB.Par ikhaimstopresent aninsightofcharacterizationofnewsstoryinthemoderndi asporacombinedwiththedifferentialcontenttypesofnews storyanditsimpacton readers. Subsequently, we dive into existing fake news detection approaches that areheavily based on text- based analysis, and also describe popular fake news datasets. Weconclude the paper by identifying 4 key open research challenges that can guide futureresearch. It is a theoretical Approach which gives Illustrations of fake news detection byanalysingthepsychologicalfactors.Himank Gupta et. al. gave aframework based on different machine learningapproachthatdealswithvariousproblemsincludi ngaccuracyshortage,timelag(BotMaker) and high processing time to handle thousands of tweets in 1 sec. Firstly, theyhave collected 400,000 tweets from HSpam14 dataset. Then they further characterize the150,000 spam tweets and 250,000 non- spam tweets. They also derived some lightweightfeatures along with the Top 30 words that are providing highest information gain fromBag-of-Wordsmodel.4.Theywereabletoachieveanaccuracyof9 1.65%andsurpassedtheexistingsolutionbyapproximatel y18%.

3. Proposed Method

Hardware Description:

Processor : Intel i5
Ram : 8GB Hard Disk
Space : 100GB

Software Description:

Operating System: Windows 10
Back-end Design: Python3
Tool: Wampserver

Wampserver:

WampServer is a Windows web development environment. It allows you to create web applications with Apache2, PHP and a MySQL database. Alongside, PhpMyAdmin allows you to manage easily your databases. WAMP acts like a virtual server on your computer. It features without any consequences allows you to test all WordPresssince it's localized on your machine and is not connected to the web.

Installation:

Download Pycharm

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Go to https://www.jetbrains.com/pycharm/download/ and Click the "DOWNLOAD" link under the Community Section.

The modules that should be installed are listed below:

1.NumPy

2.Django

3. Pandas

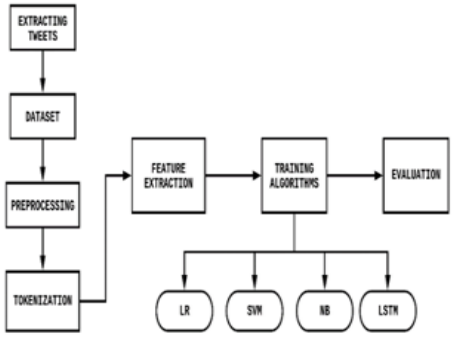

Fig 1: Block Diagram

As shown in the block diagram it is Fake news detection module, where the proposed WELFake model for fake news detection in two steps.

1) collection of various linguistic features from state-ofthe-art methods and identification of a subset that performs well on the larger WELFake data set, and

2) ensemble learning on WE features using various ML methods.

Working:

Data Collection: Firstly, Dataset can be collected from various sources of any organization. The right dataset helps for the prediction and it can be manipulated as per our requirement. Our data is in the form of images it may be based on night, fog, and rainy. The data can be collected from the organization based on the areas where the weather can be seen unusual. By collecting these it makes accurate in prediction.

Pre-processing:

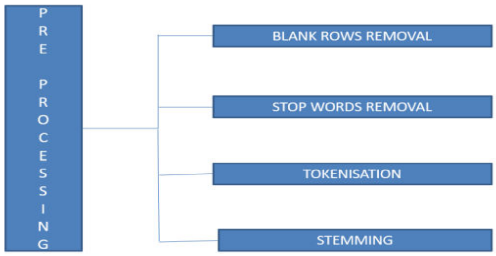

Fig 1: Data Pre-processing

Training: Training the Data after the data has been prepared and transformed, the next step was to build the classification model. This technique was selected because the construction of support vector classifiers does not require any domain knowledge. By using the attribute, we have considered in the dataset we train the model. The training sets are used to tune and fit the models.

Deploying: In this model the classification rules are generated from the support vector. The trained data can be used for the Testing the data and it helps to give output.

Tokenization:It is the process of turning a meaningful piece of data, such as an account number, into a random string of characters called a token that has no meaningful valueif breached. Tokens serve as reference to the original data, but cannot be used to guess those values.

Feature Extraction: It involves reducing the number of resources required to describe a large set of data. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy.

Data flow Diagram:

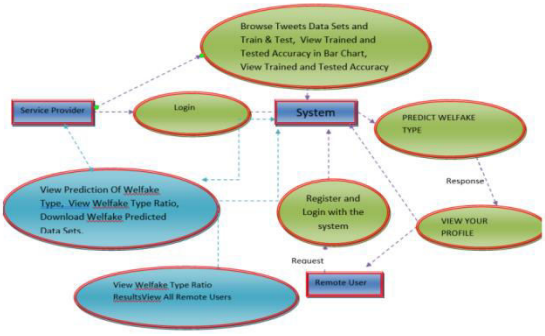

Fig 3: Data flow Diagram

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Architecture Model:

The purpose of aarchitecture model is to describe the process that takes place in this fake news detection. It consists of a webserver, welfake dataset, and service provider. The service provider provides login, browse tweets data sets and train & test, view trained and tested accuracy results, view prediction of welfake type, view welfake type ratio.

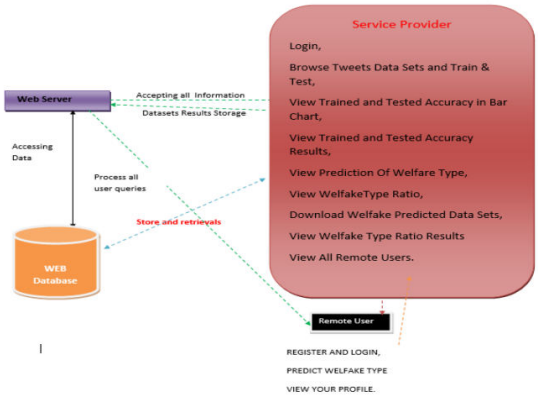

Fig 4: Architecture model

FlowChart :

Flow chart diagram itself clarifies the purpose of the diagram and other details. It describes different states of a process in a system. The flowchart of the fake news detection gives the instructions to predict the type of news.

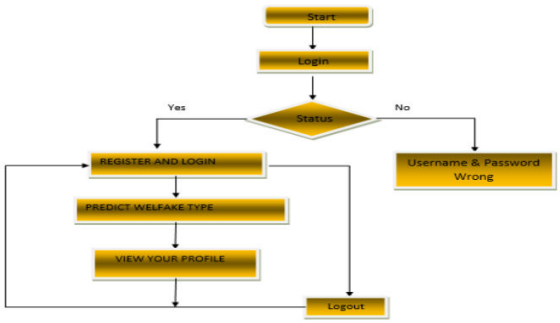

Fig 5: Flow Chart

Training the dataset:

Many datasets for fake news detection on Kaggle or many other sites. We downloaded these datasets from Kaggle. There are two datasets one for fake news and one fortrue news. In true news, there is 21417 news, and in fake news, there is 23481 news. Both datasets have a label column in which 1 for fake news and 0 for true news. We are combined both datasets using pandas built-in function[5]. A word cloud was made for the headline and body text of fake and real news in the selected dataset,Word cloud is a visualization technique of word frequency. The more regularly terms show up in the content being assessed, the bigger the word in the image created.

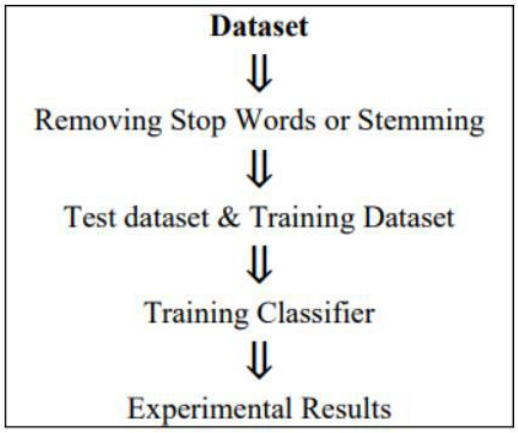

Fig 6: Training the dataset

ML Algorithms:

In ML, analyses are often grouped into broad categories for ease of reference.To categorize things, we looked at how machines learn and use feedback[6]. Supervised learning, which uses human-labelled data and information to train algorithms, and unsupervised learning, which provides no clear examples of a technique and requests thatmeaning be discovered in its input data, are the most commonly used methods in ML.

Random Forest:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. It generates accurate and diverse decisions that are dynamic algorithms for this classifier [6]. In a random forest, the individual decision trees are an ensemble, and they operate on average to increase the accuracy of the prediction of the model. This model also focuses on the reduction in over-fitting. The sub-

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

samples are drawn with replacement, keeping their size the same as the original input sample size.

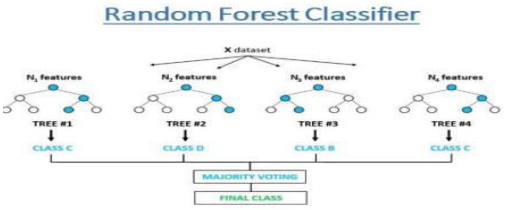

Fig 7: Random Forest Classifier

K-Nearest Neighbor:

KNN is a simple algorithm for categorizing new cases based on how similar existing ones are. The KNN was used to identify trends and patterns in statistical data. Aninstance (x) is predicted by KNN by scanning the whole training program for most similarexamples and summarizing result variables for those k cases.

Example:

➢ Training dataset consists of k-closest examples in feature space

➢ Feature space means, space with categorization variables (non-metric variables)

➢ Learning based on instances, and thus also works lazily because instance close to the input vector for test or prediction may take time to occur in the training dataset.

Logistic Regression:

Logistic regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name logistic regression is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. The name multinomial logistic regression is usually reserved for the case when the dependent variable has three or more unique values, such as Married, Single, Divorced, or Widowed. Although the type of data used for the dependent variable is different from that of multiple regression, the practical use of the procedure is similar.

Support Vector Machine:

To categorizing data, a method known as a support vector machine uses supervised learning to sort it. Because it is previously been trained, it uses a set of data, that is, already been divided into two groups when creating the model. It is the job of an SVM algorithm to figure out where a new piece of data fits in the scheme of things. As a result, the SVM may be considered a nonlinear linear classifier[7]. SVM is a discriminant technique because it solves the convex optimization problem analytically. In classification tasks a discriminant machine learning technique aims at finding, based on an independent and identically distributed (iid) training dataset, a discriminant function that can correctly predict labels for newly acquired instances.Less powerful than generative approaches, which are mostly used whenpredictioninvolves outlier detection, discriminant approaches require fewer computational resources and less training data.

Naïve Bayes

A probability-based classification method predicts class membership based on the likelihood of all possible characteristics being present in a sample at a given time. If the decision of the target class is influenced by a combination of several characteristics known as evidence, this method is used. It is possible for NB to examine characteristics that will have little influence if taken individually, but when taken together may have a substantial effect on the likelihood of an instance belonging to a certain class. Assuming all features are equal, the value of one feature does influence the value of another feature. As a result, the functions are distinct[8].

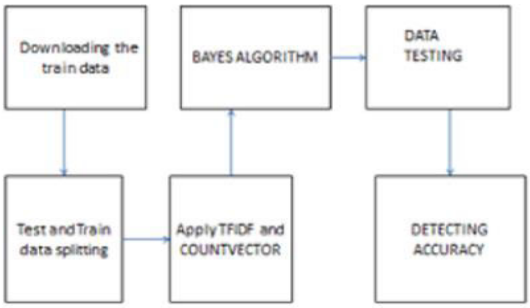

Fig 8: Naïve Bayes

Decision Tree:

The decision tree is an example of a supervised learning model. We can use it for classification and regression since it is a powerful nonparametric approach. The source set is divided into subsets in a

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

decision tree depending on the results of an attribute value test. Every subset is divided recursively, and the procedure is repeated. When all of the nodes in a subset have the same variable, the recursion is over. The decision tree's nodes and leaves are the outcome. In a decision node, the leaf node indicates either categorizationor decision. It is capable of dealing with both category and statistical data[9].

LSTM:

One of the most famous of them is the Long Short Term Memory Network(LSTM). In concept, an LSTM recurrent unit tries to "remember" all the past knowledge that the network is seen so far and to "forget" irrelevant data. This is done by introducing differentactivation function layers called "gates" for different purposes[10].

These LSTM cells, which are a kind of recurrent neuron that has been shown to provide extremely intriguing results in sequence modelling issues because they can "remember" information from the past, constitute the basis of this design. They can avoidthe vanishing gradient issue by using LSTM units, which are made up of several gates thatare responsible for keeping track of a hidden cell state.

4. Experimental Results

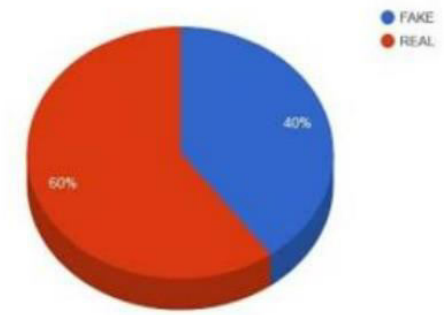

Fig 9: Pie chart

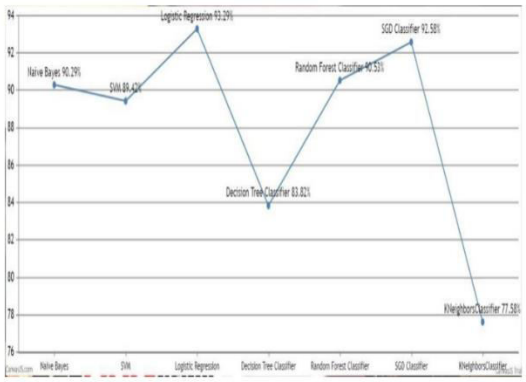

Fig 10: Line chart

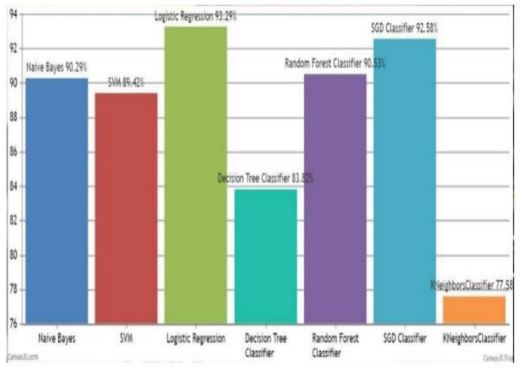

Fig 11: Bar graph

View Welfake Type Trained and Tested Results

Model Type	Accuracy
Decision Tree Classifier	81.53117600631413
Naive Bayes	89.97632202052091
SVM	88.55564325177585
Logistic Regression	92.10734017363852
Decision Tree Classifier	84.5303867403315
Random Forest Classifier	89.02920284135753
SGD Classifier	92.65982636148382
KNeighborsClassifier	80.18942383583267
Random Forest Classifier	88.08208366219415
SGD Classifier	91.31807419100237
KNeighborsClassifier	78.68981846882399

Fig 12: Trained and Tested results

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

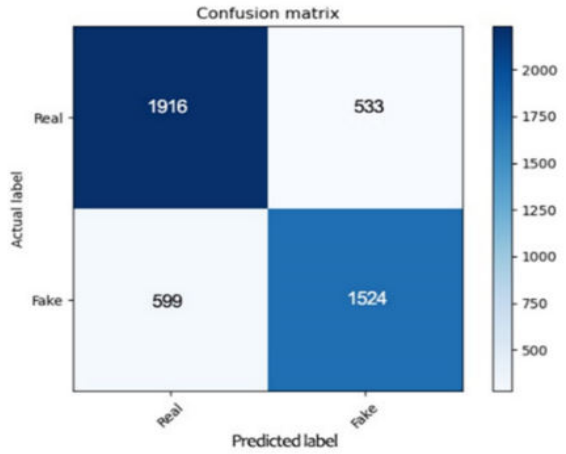

Fig 13: Confusion matrix for decision tree algorithm

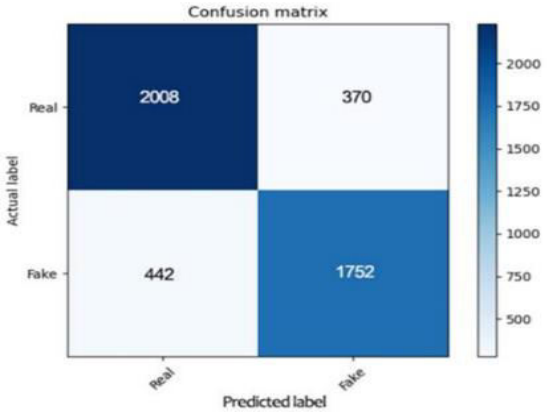

Fig 14: Confusion matrix for Random forest classifier

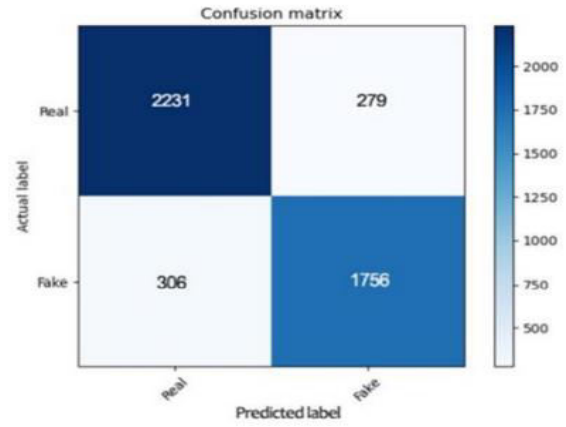

Fig 15: Confusion matrix for logistic regression

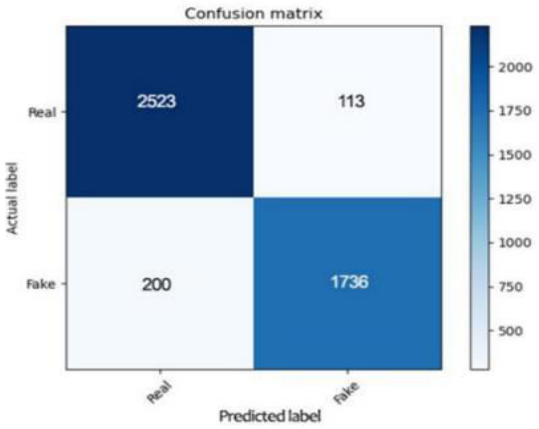

Fig 16: Confusion matrix for SVM algorithm

5. Conclusion

A new model called WELFAKE for text fake news detection. For this purpose, a larger data set called WELFAKE with over 72 000 news articles combining four open-source data sets (i.e., Kaggle, McIntire, Reuters, and Buzz Feed) to reduce their individual limitation and bias is used. Afterward, over 80 linguistic features from state-of the- art works is analyzed and selected 20 significant ones to minimize the computational complexity and increase the standard classifiers accuracy. Two WEB based methods (i.e., TF-IDF, CV) over these linguistic features using six ML models (i.e., KNN,SVM, NB, DT, Bagging, and Ada Boost) are applied and found out that CV produces better overall accuracy than TF-IDF with an SVM model. Therefore, CV over LFS is used and classified.20 features based on four categories: writing pattern, readability index, psycho linguistics, and quantity. As the number of predictors that participate in the voting classifier needs to be odd, three LFS are prepared by distributing the twenty selected features in a balanced manner across these categories. Afterward CV is embedded with these LFS and applied all six ML models. The most accurate ML model is considered and took its predicted results from each WE-enabled LFS data set for voting classification.

6. References

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access 6 (2018), 52138–52160.

[2] Naveed Akhtar and Ajmal Mian. 2018. reat of adversarial aacks on deep learning in computer vision: A survey. IEEE Access 6 (2018), 14410–14430.

[3] Abdullah Almaatouq, ErezShmueli, Mariam Nouh, Ahmad Alabdulkareem, Vivek K Singh, Mansour Alsaleh, Abdulrahman Alarifi, Anas Alfaris, et al. 2016. If it looks like a spammer and behaves like a spammer, it must be a spammer: analysis and detection of microblogging spam accounts. International Journal of Information Security 15, 5 (2016), 475–491.

[4] Jonathan Bakdash, Char Sample, Monica Rankin, Murat Kantarcioglu, Jennifer Holmes, Sue Kase, Erin Zaroukian, and Boleslaw Szymanski. 2018. Œe Future of Deception: MachineGenerated and Manipulated Images, Video, and Audio?. In 2018 International Workshop on Social Sensing (SocialSens). IEEE, 2–2.

[5] MelikaBehjati, Seyed-Mohsen Moosavi-Dezfooli, MahdiehSoleymaniBaghshah, and Pascal Frossard. 2019. Universal Adversarial AŠ acks on Text Classifiers. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 7345–7349.

[6] Sreyasee Das BhaŠacharjee, AshitTalukder, and BalaVenkatramBalantrapu. 2017. Active learning based news veracity detection with feature weighting and deep-shallow fusion. In 2017 IEEE International Conference on Big Data (Big Data). IEEE, 556–565.

[7] Juan Cao, Junbo Guo, Xirong Li, ZhiweiJin, Han Guo, and Jintao Li. 2018. Automatic rumor detection on microblogs: A survey. arXiv preprint arXiv:1807.03505 (2018).

[8] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twiŠ er. In Proceedings of the 20th international conference on World wide web. ACM, 675– 684.

[9] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2013. Predicting information credibility in time-sensitive social media. Internet Research 23, 5 (2013), 560–588.

[10] Rose Catherine, Kathryn Mazaitis, Maxine Eskenazi, and William Cohen. 2017. Explainable entity-based recommendations with knowledge graphs. arXiv preprint arXiv:1707.05254 (2017).

[11] Tong Chen, Xue Li, Hongzhi Yin, and Jun Zhang. 2018. Call aŠention to rumors: Deep attention based recurrent neural networks for early rumor detection. In Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 40–52.

[12] Wenji Chen, Yang Liu, and Yong Guan. 2013. Cardinality change-based early detection of large-scale cyber-aŠ acks. In 2013 Proceedings IEEE INFOCOM. IEEE, 1788–1796.