ISSN: 0975-3583, 0976-2833 VOL 11, ISSUE 02, 2020

An Integrated Feature Engineering Approach to Identify Falsity in COVID-19 Information

Haritha Akkineni [1]* Pratuisha Koripilli[2] VenkataSuneetha Takellapati [3] Deepthi Gurram [4]

[1] Department of Information Technology, PVP Siddhartha Institute of Technology, Vijayawada, India
[2]Department of CSE, KL Deemed to be University, Vijayawada, India
[3]Department of CSE, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India
[4] Department of Computer Science, St.Ann's College for Women, Hyderabad, India
* akkinenih@gmail.com

Abstract: There is huge flow of data in social media which obviously causes a rise in the amount of information. This is more evident particularly in the information regarding COVID-19 pandemic. Despite of the information being periodically updated on regular basis, lot of disturbances still exists. Countering misinformation is a mammoth task. The misinformation during this time should be monitored properly as it creates havoc in the society if it's left unattended. So identifying fake information is the main objective of this investigation. The analysis is based on a dataset of 1100 posts related to COVID 19 collected from social media. We propose an Integrated Feature Engineering (IFE) approach which uses a combination of content based and count vectorization methods. This study focuses on applying the proposed approach and gives a comparison with in individual methods and finally assesses the performance of different machine learning classifiers for classifying fake news.

Keywords: Vector Representation, Classifiers, Content Based Features, COVID 19 Fake

1. Introduction

Social Media is becoming the major source of information today and many are relying on social media. It can disseminate the news more quickly than any other medium. Trust in social media is becoming a primary concern. There is major scope for misinformation propagation in today's digital world. Fake information gets traveled in really a jet speed than what is treated as the original. The after effects will really be threatening and scary. Fake news in social media will surely have an adverse effect in the general public and it will create wrong notion in society.

"We're not just fighting an epidemic; we're fighting an infodemic ". as stated by Director-General of WHO Tedros Adhanom Ghebreyesus proclaimed at the Munich Security Conference*[1]*. It has been identified that the misinformation available in the social media might be the main cause for the spread of COVID 19 pandemic at a really disturbing rate *[2]*. There are some sought of scenarios where online gossipy tidbits blaming 5G arrangements for causing COVID-19 prompted cell phone poles being assaulted in the UK. Vinegar is more successful than hand sanitizer against COVID-19.

This misinformation not only contributes to the spread of fake rumors, it shoots up fear among public and could lead to societal disaccord and lead to direct damage. There are many studies showing that some of the people have lost their lives absolutely due the fear of Corona rather than affected by the disease.

Normally, users who believe in such fake news could continue to disrupt general wellbeing. The activities of individual residents guided by the nature of the data they have nearby are essential to the achievement of the worldwide reaction to this wellbeing emergency.

So in order to address this health crisis there is a need for us to continuously monitor the outspread of misinformation related to COVID-19. If not handled it has potentially dangerous consequences. A potential use case in this scenario is to devise classifiers and techniques to stop this flow. A data set containing information related to COVID 19 is collected. The data is preprocessed. The combination of vectorizers and content based methods is used in feature extraction. The content based features concentrate on text style, syntax and components related to grammar. The count vectorizer focuses on the frequency of the words. This proposed method, Integrated Feature Engineering is given as input to different classification models to find the best parameters that would give the highest accuracy score. So that it is regarded as the one most capable of classifying the fake from original information. Further parts are arranged in the following manner. *Section II* described the related work in this field. *Section III* discusses the methodology and the architectural framework. *Section IV* discuss the mathematical modeling of the complete procedure. *Section V* represents results and discussion of the work done.

ISSN: 0975-3583, 0976-2833 VOL 11, ISSUE 02, 2020

2. Related Work

The problem of misinformation propagation during this situation could really have adverse effects on the society. Many researchers have come forward to address this problem. One of such kind is the data science community. “Fake News Challenge” is conducted by Kaggle. Facebook is using AI to filter fake from original posts. All the social media giants are having an eye on the spread of misinformation. There is some related work in this area.

Zhang and Ghorbani *[3]* investigated in online news and introduced an extensive outline of the ongoing inventions connected to the fake news. In addition, they portrayed the effect of online fake news, introduced futuristic detection recognition strategies, and talked about the ordinarily utilized datasets utilized to construct models to classify news.

Collins and Erascu *[4]* gave a review on different algorithms to recognize various kinds of fake news. They concluded that the results are satisfactory with combined effort of humans and machines rather than when contrasted with frameworks that rely just upon either one. There is no explicit framework discussed to identify fake news.

Al Asaad et al. *[5]* planned a new model to check validity of the news using machine learning classifiers. The compare the effectiveness of the models with used different algorithm like Multinomial. The work done here confines only to the developing the model.

Elhadad et al. *[6]* came up with a model that works on social network platforms to detect fake from real news. To build the feature vector they focused on hybrid features from Meta data of news articles. The efficiency was tested on comparing with nine different machine learning algorithms with three different datasets. In this paper a feature building utilizes metadata which can be a major source of noise. As there are too many metadata standards the data retrieval process is a tedious task here.

Wang *[7]* presented another dataset for fake news. Dataset comprises 12.8K physically marked short declarations in different settings from PolitiFact.com. Prominently, this dataset was viewed as the primary huge dataset identified with fake news recognition. It is a significant degree bigger than past public fake news datasets. Focus is on building dataset and not on any model building or assessing classifier performance.

Fake Health *[8]* is gathered from medical care data survey site Health News Review, which audits whether news is acceptable from 10 standards. Every day distribution of new substance on this site was stopped toward the finish of 2018. The writers slithered clients' answers, re-tweets and profiles by

utilizing article URL and article title by utilizing Twitter API.

Yang et al. *[9]* investigated the usage of the volume of tweets connecting to low-believability data was contrasted with the volume of connections to the New York Times and Centre for Disease Control and Prevention. This paper concentrated only on the volume of tweets, nothing has been discussed with respect to modeling and performance evaluation.

3. Methodology

To begin exploring the content of news articles related to COVID-19, Over 1,100 news articles and posts on social media related to COVID-19 pandemic from a variety of novel sources were collected. Proposed model process is presented in the form of an architectural framework. Thereafter, The characteristics of the dataset is discussed, data preprocessing, feature extraction and classification model selection were focused on. The steps of the proposed model are shown in *Figure. 1*.

3.1 The Architectural Framework of the Integrated Feature Engineering (IFE)

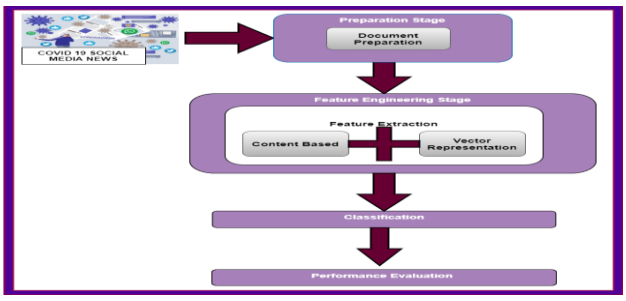

Figure. 1 Architectural Framework of Integrated Feature Engineering (IFE)

ISSN: 0975-3583, 0976-2833 VOL 11, ISSUE 02, 2020

3.2. The Data

The first step is to collect the data, and labeling of the data as fake and original is done. The data set used for analysis is collection of over 1,100 news articles and posts on social media related to COVID-19 pandemic from a variety of new sources. The data is subsequently tagged as real and fake.

3.3. Preprocessing

The noise removal process is being done during this stage. There will be plenty of irrelevant words or text in the textual data which is collected from social media like the source names, stop words to sanitize the text content.

3.4. Feature Extraction

To arrange news articles the crude content information should be transformed into something more valuable. This is called feature extraction. It can take numerous structures: n-gram counts, punctuation usage, word counts, Term frequency, inverse document frequency and numerous others. The extracted features would then be able to be utilized to classify the article. By determining the best features for classifying process can be more accurate. To study the fake and the original news used proposed Integrated Feature Engineering.

The feature engineering involves a combination of two different methods

3.4.1. Content based features

These are extracted from the COVID 19 news data set the main focus here is the stylistic features. They are based on text style, syntax and components related to grammar. This process is carried out using POS tagger which uses NLTK to keep a monitor on each tag count within an article. Like the count on proper nouns, number of verbs, nouns, comparatives, determinants and superlatives is being taken into count. Along with this, Word Count dictionaries are used to keep track of the frequencies of belief, surprise, negation, existential, conditional, modal, and interjection words.

For this purpose textstat - a Python library is used to calculate statistics from text to resolve complexity, grade level and readability of any article.

Type token ratio gives the ration of number of unique words to total amount of words in a given segment of language.

3.4.2. Features based on vectorization

Count Vectorizer is the second technique in the integrated approach. The whole of the text is transformed to a vector form based on term/token counts. The document is divided into tokens

If Document= *[" One cent, two cents, old cent, new cent: about money"]*

This text is changed to a sparse matrix as represented in tables below. This is a count vectorizer sparse matrix representation of the text data in the document. The *table 1.* shows how does it represent according to theory. *Table 2.* represent the actual way it works

Table 1. Vectorization in theory

	all	cent	cents	money	new	old	one	two
Doc	1	3	1	1	1	1	1	1

Table 2. Vectorization in practice

	0	1	2	3	4	5	6	7
Doc	1	3	1	1	1	1	1	1

As there are 9 unique words, there are 8 columns. The column in the matrix designates the distinctive word in the vocabulary, whereas the document in the dataset is shown in form of a row.

ISSN: 0975-3583, 0976-2833 VOL11, ISSUE 02, 2020

The document taken as example has got only one sentence, so there is only one row. The values given in the cells are the frequency of the words. The value could be zero if the word is absent in the document.

3.5. Classification

The extracted features will be to train classifiers. The classifiers used here are

3.5.1. Naive Bayes

It is a simple probabilistic classifier to make predictions on the basic of probability of an object. It considers each feature and treats it as unrelated to any other feature. Then the probability of the feature belonging to a particular class is identified. It does that for each feature and then aggregates each individual probability to calculate the final classification. It works on Bayes theorem *[10]*

3.5.2. Logistic Regression

A log-linear classifier is viewed as general linear classifier, since its result relies upon the whole of its sources of info and boundaries [12].

Once more, the point is to learn, from the preparation set, the mapping function which maps hidden information, which is a vector of features, to a particular class.

3.5.3. A Decision Tree

It is analogous to the human decision making process and make it easy to understand. It is a graphical representation for getting all the possible solutions to a problem based on given conditions. It is build based on tree structure. It starts at the root compares internal nodes and jumps to the leaf where it gets to a particular class. This leaf will give the class label [13].

3.5.4. Neural Network

Training a neural network includes masterminding all the weightage by replaying two significant steps, forward and in backward propagation. Collection of weights in forward propagation is given to the input to have an estimate on the output. As in the case of backward propagation the gauge is kept on the margin of error of the estimated output and then work towards modification of the weight to reduce the error. The forward and backward propagations are replayed until all weights get adjusted to predict the accurate output.

A function is used in the hidden layers of the neural network that sums the inputs with the weights and maps the accurate output. Some of the functions are linear, sigmoid, hyperbolic tangent etc. *[11]*

3.6. Evaluation

To assess the performance of the classifier a common simple metric used is accuracy. This is not appropriate for unbalanced datasets. Apart from this metric others like *confusion matrices* are also explored.

4. The Mathematical model

The COVID-19 social media news textual data is a collection of N documents pertaining to COVID news $D1,\ldots,D_N$ and associated labels $y1,\ldots,y_N$. Here we have two class labels Fake and Original.

The main problem of focus is we are interested in classifying or extracting the fake from the original news. For the purpose of classification the class label $yi \in C$ is one of $C=|Original, Fake|$. This is exactly Binary classification. D represent the set of documents with corresponding labels with N training examples.

$$D = \{D_i, y_i\}N_{i=1} = \{(D_1, y_1), \ldots . (D_i, y_i), \ldots (D_N, y_N)\} \ldots\ldots\ldots (1)$$

The document D_i in Eq. (1) comprises of sequence of tokens which is represented by

$$S_i=(t_{i,1},\ldots,\square_{,j},\ldots,t_{i,ni}) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

ISSN: 0975-3583, 0976-2833 VOL 11, ISSUE 02, 2020

with variable length n $_i$ in Eq. (2) contains Tokens which can be punctuation, words or other meaningful units.

$$A = \bigcup_{i=1}^{N} \bigcup_{j=1}^{n_i} \{t_{i,j}\}$$

In Eq. (3) has tokens which are elements of alphabet and *t i,j* belongs to *A*

POS Tagging

The Stochastic approach is being used for POS Tagging. A special variant is the word frequency approach which works on the probability that a word occurs with a particular tag. The words in document D in (4) are designated with w. Assign each word w, its most likely POS tag. If w has tags tag1, tag 2, tag k, then

$$P(tag_i|w) = c(w, tag_i)/\, c(w|tag_1) + \cdots + c(w, tag_k))$$

c(w, tag $_i$)= number of times w/t $_i$ appears in the corpus

4.1. Count Vectorization

Along with POS tagging to extract the features form the COVID 19 news articles count vectorization is being used.

The documents *d $_1$ to d $_N$* has *N* unique tokens extracted. The N tokens are extracted out from the document and size of CountVectorMatrix CVM is given by DXN. Each row represents the frequency of each token which comprises all *A*=|*A*| tokens of training set *d*. A is a result of combination of POS and Count Vectorization.

On considering document D, for which the class label is unknown *y*∈*C*. The probability of the document which has not been classified belongs to c∈C is denoted as *P(y=c|d,d,M),* which is conditional on the unclassified document *D*, data set D, and machine-learning model M.

The task of the Identifying fake from original COVID-19 post problem is to estimate the label y^∈C from an unlabeled document D as

y^=argmax c∈CP(y=c|D,D,M) ……. (6)

where *M* takes the values *M $_1$ to M $_4$*.

{M1= Naive Bayes, M2= Logistic Regression, M3=Decision Tree, M4 =Neural Network} as in *Eq. (6)*

The probability *P* is effected by the removal of input- output pair from the training set. Any changes to the machine-learning model *M*, which includes tokenizing, feature extraction and the classification algorithm that is chosen will also contribute to the probability of the document being fake. The accuracy of all the models in the set *M* are calculated and the one with highest accuracy can be regarded as the algorithm which performs better than the remaining ones in the set.

5. Results and Discussion

5.1. Exploratory Data Analysis (EDA)

The data set used for analysis consists of 1164 rows in which each of those has *4 attributes*. *Figure. 2* represents a balanced data set because both the fake and true data are equally represented. *Number of fake data: 575 Records containing true data: 584*

	title	text	source	label
0	Due to the recent outbreak for the Coronavirus...	You just need to add water, and the drugs and ...	coronavirusmedicalkit.com	Fake
1	NaN	Hydroxychloroquine has been shown to have a 10...	RudyGiuliani	Fake
2	NaN	Fact: Hydroxychloroquine has been shown to hav...	CharlieKirk	Fake
3	NaN	The Corona virus is a man made virus created i...	JoanneWrightForCongress	Fake
4	NaN	Doesn't @BillGates finance research at the Wuh...	JoanneWrightForCongress	Fake

Figure. 2 Data Set View

ISSN: 0975-3583, 0976-2833 VOL 11 ISSUE 02, 2020

5.2. Feature Engineering

5.2.1. Content based features

These are formulated using Capital Letters in Title for Integrated Feature Engineering. *Figure. 3* show the percentage of capital letters in each article body is computed rather than counting the number, because the length of the articles is very different.

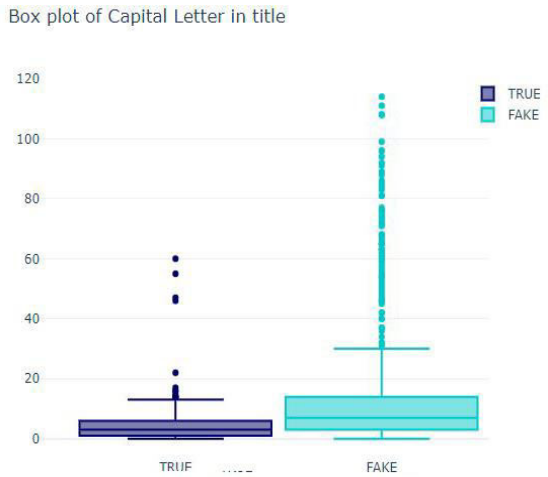

Figure 3 Capital letters in title for Integrated Feature Engineering

On average, title in fake news has more words in capital letters. This shows that fake news is mostly focusing on the readers who are likely to be inclined by titles.

5.2.1.1. Stop Words in Title

Figure. 4 show the percentage of stop words for Integrated Feature Engineering in every article body is computed.

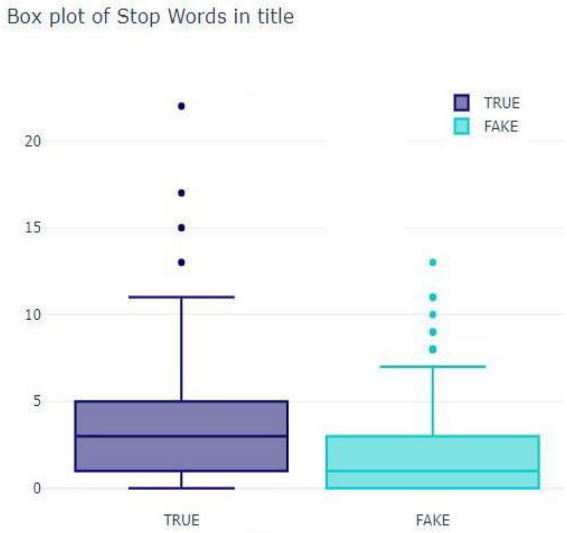

Figure. 4 Stop Words in title for Integrated Feature Engineering

Stop words are rather lower in Fake news titles compared to real news.

ISSN: 0975-3583, 0976-2833 VOL 11, ISSUE 02, 2020

5.2.1.2. Proper Noun in Title

To have an estimate on number of proper nouns in each title for Integrated Feature Engineering. Proper nouns are more predominant in Fake news. Evidently the utilization of nouns in titles are exceptionally huge in separating fake from genuine.

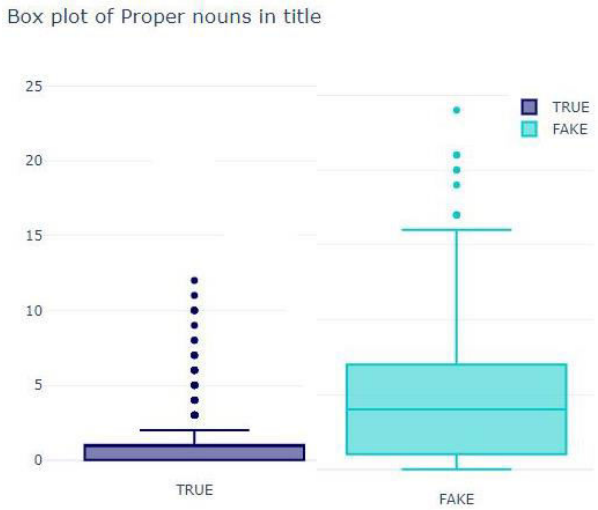

Figure. 5 Stop Words in title for Integrated Feature Engineering

On the whole, these results state that the authors of fake news are attempting to attract attention by using all capitalized words in titles, and squeeze as much substance into the titles as possible by skipping stop-words and increase proper nouns as in Figure. 5.

5.2.2. The Stylistic Features

Figure. 6 depicts the count on number of nouns for Integrated Feature Engineering

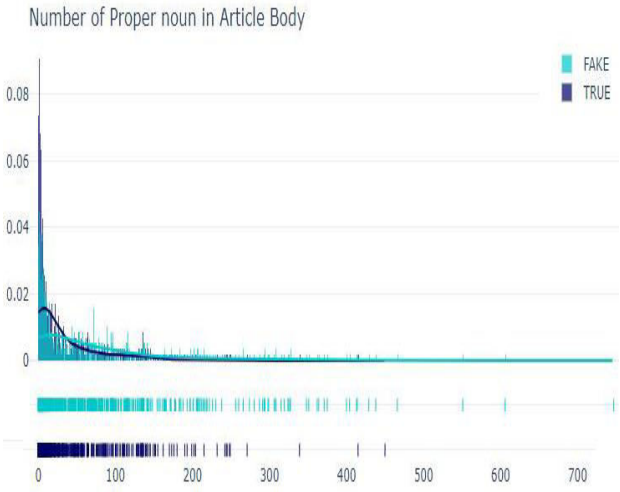

Figure. 6 Stop Words in title for Integrated Feature Engineering

Figure. 7 shows the count on the number of verbs for Integrated Feature Engineering

ISSN: 0975-3583, 0976-2833 VOL 11, ISSUE 02, 2020

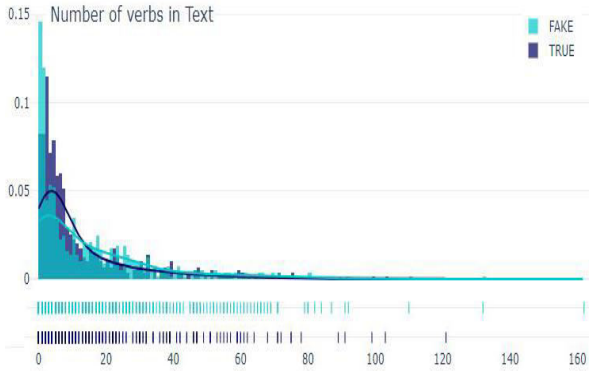

Figure. 7 Count on the verbs for Integrated Feature Engineering

Figure. 8 gives the frequencies of negation for Integrated Feature Engineering

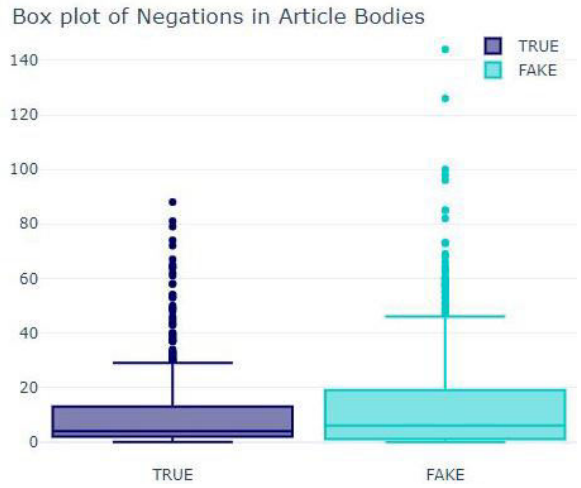

Figure. 8 Number of Negations for Integrated Feature Engineering

Fig. 9 shows the stop words in the article body for Integrated Feature Engineering

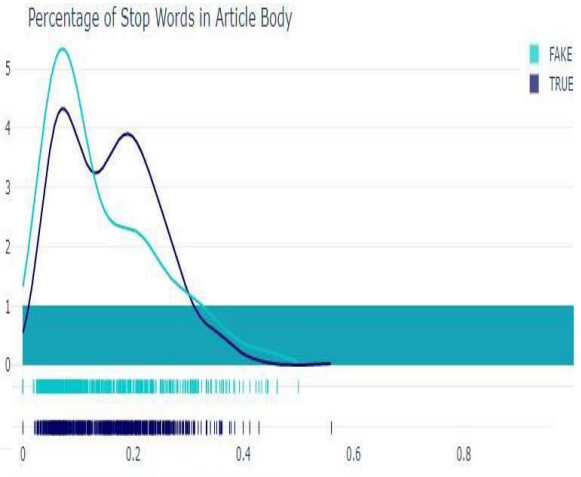

Figure. 9 Stop Words for Integrated Feature Engineering

ISSN: 0975-3583, 0976-2833 VOL 11, ISSUE 02, 2020

Figure. 10 shows the type token ratio for Integrated Feature Engineering

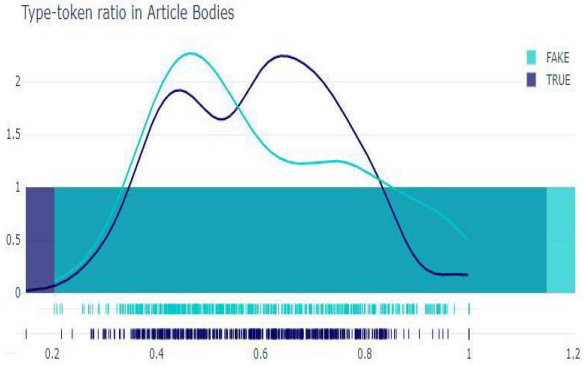

Figure. 10 Type Token Ratio for Integrated Feature Engineering

5.2.3. Count Vectorization

The dataset is converted into array of word occurrences. The shape of the array above is (1159, 21117) which represents the number of samples and the feature vector size of each sample respectively.

5.3. Classifiers

The confusion Matrix for Decision Tree Classifier using for Integrated Feature Engineering approach is given in Figure. 11

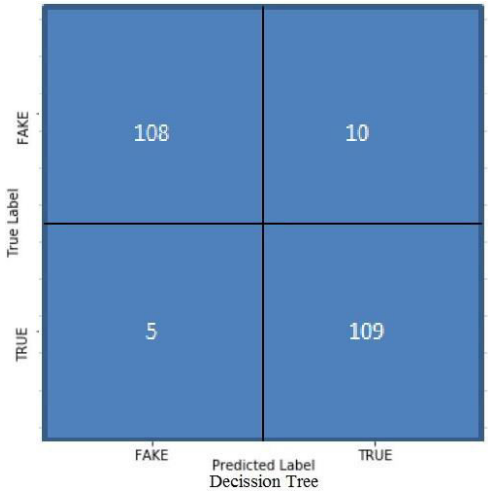

Figure. 11 Confusion Matrix

TP=108; TN=109; FP=10;FN=5

Classifier Accuracy: 217/233=0.935

Table 2. Confusion matrix of all the classifiers

Machine Learning Classifier	True Positive	False Positive	False Negative	True Negative
Decision Tree	108	10	5	109
Logistic Regression	113	4	6	109
Naive Bayes	108	10	5	109
Neural Network	110	6	5	90
IFE	130	1	5	90

ISSN: 0975-3583, 0976-2833 VOL 11 ISSUE 02, 2020

5.3.1. Performance evaluation: Accuracy of all the classifiers along with IFE

Accuracy of classifiers using only Count Vectorization approach is depicted in table 3.

Table 3. Classifiers with count vectorization

Classifier	Accuracy on Train Data	Accuracy on Test Data
Naive Bayes classifier	0.963322545846817	0.93534482758620
Logistic regression	0.974522155522522	0.951546265956556
Decision Tree	0.955484515455554	0.9353448275862069
Neural Network	0.9654845254615	0.953546586206
IFE	0.995484515455554	0.985344827586206

Accuracy of classifier using only Content Based Features is depicted in Table 4.

Table 4. Classifiers with content based features

Classifier	Accuracy on Train Data	Accuracy on Test Data
Naive Bayes classifier	0.9523225458468	0.92348275862069
Logistic regression	0.9445221555225	0.931546265956556
Decision Tree	0.9444845154555	0.922344827586206
Neural Network	0.96879687787787	0.957876758787
IFE	0.9814845154555	0.972344827586206

The count vectorization and content based features are applied to all the classifiers in comparison with IFE is shown in Figure. 11. IFE has 0.99% on train data and 0.98% on test data with count vectorization and 0.98% on test and 0.97% on test data with content based features.

ISSN: 0975-3583, 0976-2833 VOL 11, ISSUE 02, 2020

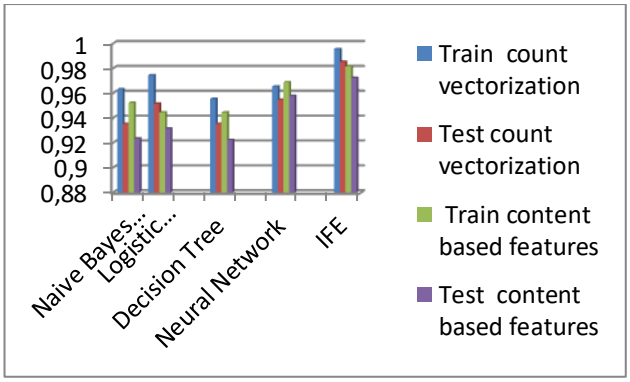

Figure.11 Comparison chart of Count Vectorization& Content based feature with the Classifiers.

accuracy of the classifiers using Integrated Feature Engineering approach is shown in Figure. 12

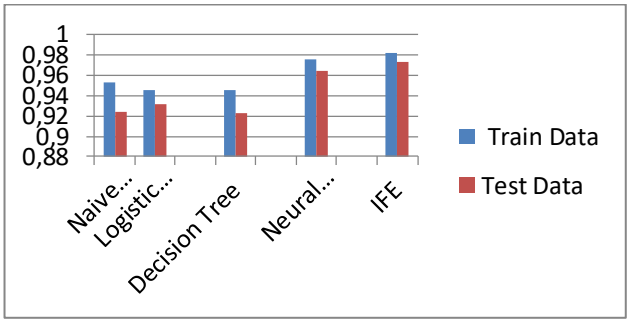

Figure.12 Accuracy of the Classifiers compared with IFE

The accuracy of the classifiers using the Integrated Feature Engineering outperforms the individual classifiers. The proposed framework used a combination of content based features and count vectorization to find the best parameters that gave the highest accuracy score of 0.98 on train data and 0.97 on the test data.

6. Conclusion

In this paper, the methodology which included an integrated feature engineering approach was applied in detecting misleading information related to the COVID-19 outbreak. Off-the-shelf NLP models, however, do not perform well on this data, indicating a need for further research and development on this topic. The novelty of this framework is in the feature engineering stage where an integration of content based and count vectorization is being used to extract the features. The classification algorithms are used and the performance is being assessed. It is evident through the results that the data collected is valid and gave superior perception of the performance of different classification algorithms on them. Considering the accuracy, The proposed framework used a combination of content based features and count vectorization to find the best parameters that gave the highest accuracy score of 0.98 on train data and 0.97 on the test data as compared with the remaining classifiers.

References

[1] Zarocostas, J., 2020. World Report How to fight an infodemic. The Lancet 395, 676. doi:10:1016/S0140-6736(20/30461-X.

[2] M. D. Ibrishimova and K. F. Li, ``Amachine learning approach to fake news detection using knowledge verification and natural language processing," in Proc. INCoS, Oita, Japan, 2020, pp. 223_234.

[3] X. Zhang and A. A. Ghorbani, ``An overview of online fake news: Characterization, detection, and discussion," Inf. Process. Manage., vol. 57, no. 2, Mar. 2020, Art. no. 102025, doi: 10.1016/j.ipm.2019.03.004.

[4] B. Collins, D. T. Hoang, N. T. Nguyen, and D. Hwang, ``Fake news types and detection models on social media: A state-of-the-art survey," in Proc. ACIIDS, Phuket, Thailand, 2020, pp. 562_573.

[5] B. Al Asaad and M. Erascu, ``A tool for fake news detection," in Proc. 20th Int. Symp. Symbolic Numeric Algorithms Scienti_c Comput. (SYNASC), Sep. 2018, pp. 379_386, doi: 10.1109/SYNASC.2018.00064.

[6] M. K. Elhadad, K. F. Li, and F. Gebali, ``A novel approach for selecting hybrid features from online news textual metadata for fake news detection," in Proc. 3PGCIC, Antwerp, Belgium, 2019, pp. 914_925.

[7] W. Y. Wang, ```liar, liar pants on Fire': A new benchmark dataset for fake news detection," in Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (Short Papers), vol. 2, 2017, pp. 1_5, doi: 10.18653/v1/P17-2067.

ISSN: 0975-3583, 0976-2833 VOL 11, ISSUE 02, 2020

[8] Enyan Dai, Yiwei Sun, and Suhang Wang. Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. arXiv preprint arXiv:2002.00837, 2020.
[9] Yang, K.C., Torres-Lugo, C., Menczer, F., 2020. Prevalence of Low-Credibility Information on Twitter During the COVID-19 Outbreak. ArXiv preprint URL: http://arxiv:org/abs/ 2004:14484http://dx:doi:org/10:36190/2020:16, doi:10:36190/2020:16, arXiv:2004.14484.
[10] Shlok Gilda,"Evaluating machine learning for fake news detection", 2017 IEEE 15th Student conference on Research and Development, NSPEC. algorithms for fake news detection", 2017 IEEE 15th Student conference on Research and Development, INSPEC Accession Number: 17613664.
[11] Chaitanya Naik, Vallari Kothari, Zankhana Rana, "Document Classification using Neural Networks Based on Words", In: International Journal of Advanced Research in Computer Science,2015.
[12] Sebastian Raschka. Why is logistic regression considered a linear model? 2013. https://sebastianraschka.com/faq/docs/logistic regression linear.html. pages 21,
[13]Kumar, K. & Suresh, Y. & Srinivasu, S.V.N (2018). Image mining and viewpoint patterns: Review and challenges. Journal of Advanced Research in Dynamical and Control Systems.376-381.

Authors Contributions

Dr Haritha Akkineni received her Ph.D in Computer Science and Engineering. She is working as Associate Professor in PVP Siddhartha Institute of Technology Her research interests are Data Science, Image Mining, Artificial Intelligence, Data Analytics, Deep Learning and Machine Learning. She has published about 35 papers in reputed Journals like SCOPUS UGC etc. She has published 2 patents.

Ms Pratuisha Koripilli is a Research Scholar at KL Deemed to Be University. She is working in the area of Data Sciences and Machine Learning.

Ms VenkataSuneetha Takellapati is working as Assistant Professor in the department of Computer Science and Engineering at Gokaraju Rangaraju Institute of Technology, Hyderabad. Her research interests include developing algorithms and models for building systems and applications in areas like Computer Networks, Data Mining with security applications in which she has various publications, in various journals and conferences.

Deepthi Gurram is currently working as Assistant Professor in Department of Computer Science, St.Ann's College for Women, Hyderabad and pursuing her Ph.D in area of medical machine learning. She has various publications in reputed Journals.