7

by 7 7

Submission date: 31-May-2022 05:45PM (UTC+0530)
Submission ID: 1847777466
File name: DMA_Co19322_FInal_Report_no_code.docx (602K)
Word count: 2768
Character count: 14966

10

Chandigarh College of Engineering and Technology
(Degree Wing)

Compiler Design (CS – 605C)
Final Project File

Submitted By
Dipesh Singla
Roll: - CO19322
C.S.E 6th Semester

Submitted To
Dr. Varun Gupta
Professor CSE, HOD (A.S)

Table of Contents

1. Problem Introduction

News is being circulated in an ever-increasing social context, whether on social media platforms or the internet and that too in bulk. With so much news arriving from everywhere, it becomes difficult to check the source and validity of the news, that is, to distinguish between false and accurate news. This project provides an overview of news detection, which uses existing Machine Learning Models and Deep Learning Algorithms to determine whether the news is correct, i.e. authentic, and trustworthy.

Why fake news is a problem?

Fake news is defined as misinformation, disinformation, or malicious information conveyed by word of mouth and conventional media, as well as, more recently, digital modes of communication such as manipulated videos, memes, unconfirmed adverts, and social media promoted rumours. Fake news on social media has become a severe concern, with the potential for it to lead to mob violence, suicides, and other negative outcomes as a result of disinformation spreading on social media.

Keywords

True and Fake News, Data Mining, Twitter, Google News Vector, word2vecgensim model

2. Project Methodology

Clement Bisaillon at [1] has donated the dataset used. It comprises two CSV files, one for bogus news called Fake.csv and the other for factual news called True.csv.

First, we displayed each of the data columns (such as Title, Subject, and Date) and examined how they are connected and what sort of problem we can deduct from that situation. And concluded that it is a binary classification problem in which a news item can be either phony or truthful. Using these representations, we vectorized the textual data in several ways and predicted it using the Logistic Regression Model, Decision Tree Classifier, and Artificial Neural Network. The Decision Tree Classifier surpasses other models in identifying false and real with an accuracy of 99.63 percent in the final model tests. As of now, a comprehensive solution for the use case is also attached.

3. Flow of Work

The flow of work is as shown:

3.1 ETL (Exploration, Transformation, and Loading)

a) Analysis of Dates
b) Analysis of Subjects
c) Analysis of Title
- Based on the length of the title
- Based on the word count of the title
- Visualizing the most used words (100) (for this we used df_data_1 and df_data_2)

d) Analysis of Text
- Based on the length of text
- Based on word count of column text
- Visualizing the most used words (100) (for this we used df_data_1 and df_data_2)

3.2 Feature Engineering

a) Data Cleaning
b) Pre-Processing
c) Feature Creation
d) Visualizations

3.3 Model Definition and Training (Machine Learning Models)

a) Using Logistic Regression (from scikit-learn) with 96D vector features provided by spaCy
b) Using Decision Tree Classifier (from sci-kit-learn) with 96D vector features provided byspaCy
c) Iteration/Approach 1: Using the above models mentioned with spaCy vectors as inputfeatures for predictions.
d) Iteration/Approach 2: Using the above models mentioned by adding CountVectorizer andTFIDF Transformer from sklearn.feature_extraction

3.4 Model Definition and Training (Deep Learning Algorithms)

a) Using Artificial Neural Network from a Sequential Model
b) Iteration 2 uses more features with accuracy and F1-score as the primary metrics of evaluation.

3.5 Model Evaluations

a) Based on test accuracy
b) Based on F1 scores

4. Architectural Decisions/ Tools Used

4.1 Brief description of the dataset

This dataset consists of about 40000 articles consisting of fake as well as real news. We aim to train our model so that it can correctly predict whether a given piece of news is real or fake. The fake and real news data are given in two separate datasets with each dataset consisting of around 20000 articles.

4.2 Data Source

a) **Technology Choice**:

The data source makes data available in CSV format as Fake.csv and True.csv [1].

b) **Justification:**

It is simple to access, gather and explore such datasets, and they also includeresearch on a variety of people, which provides us with the foundational approach for working on machine learning or deep learning challenges.

4.3 Enterprise Data

a) **Technology Choice:** Not required initially. We can use Lift for the enterprise data if required.

b) **Justification:** The lift enables us to efficiently shift on-premises data to cloud databases. In our situation, we used a preset data set and verified that it worked in any similar real-world scenario.

4.4 Data Repository

a) **Technology Choice**: Cloud Storage (IBM)

b) **Justification**: Cloud object storage allows you to store almost unlimited amounts of data. It's commonly utilized for scalable and long-term data archiving and backup, online and mobile apps, and analytics storage.

4.5 Discovery and Exploration

a) **Technology Choice:** Python, Jupyter, sci-kit-learn, pandas, Matplotlib, Seaborn

b) **Justification**: Open source and supported in IBM Cloud are Jupyter, Python, sci-kit-learn, pandas, Matplotlib, and Seaborn. Some of these components have overlapping

characteristics, while others have complementing characteristics.

4.6 Actionable Insights

a) **Technology Choice:** Python, pandas, nltk, and sci-kit-learn for Machine Learning Models and Keras and Tensorflow for Deep Learning Algorithm

b) **Justification**: Python, pandas, nltk, and sci-kit-learn are great alternatives to R/R-Studio for data research. Python is also a cleaner and simpler programming language to learn than R. Pandas is Python's equivalent of R data frames, giving access to relational data. Nltk is a natural language processing toolkit that comes in handy when dealing with vast quantities of text. Finally, scikit-learn conveniently organizes all of the machine learning methods necessary. It is also available on IBM Cloud via IBM Watson Studio.

Keras and TensorFlow are the most popular technologies for deep learning algorithms for framing neural networks. One of the most commonly used deep learning frameworks is TensorFlow. It is a linear algebra library that supports automated differentiation at its core. TensorFlow's Python-based syntax is rather complicated. As a result, Keras adds an abstraction layer to TensorFlow. Watson Studio and Watson Machine Learning on IBM Cloud provide seamless support for both frameworks.

4.7 Applications / Data Products

a) **Technology Choice**: Export of an already run, static Jupyter Notebook

b) **Justification**: IBM Watson Studio's Jupyter Notebook is one of the simplest, fastest, and most efficient methods to provide a data product based on what we have worked on. It not only provides a fantastic platform for doing everything at once, but it also keeps everyone on the same page, whether it's data visualization, cleaning, pre-processing, training the model, or delivering the entire project.

4.8 Security, Information Governance and Systems Management

a) **Technology Choice**: Cloud Object Storage

b) **Justification:** When it comes to contemporary, cost-effective cloud storage, Object Storage is the gold standard.

5. Proposed approach/ Methodology and Model:

In this work, in Phase-1 first of all we've cleaned the data, then a list of indices where the publisher is not mentioned, then Separating Publication info, the from the actual, text and The listing text column with new text there were 630 rows in fake news with empty text, then we created world cloud for both true and fake news and both were pretty different. In pre-processing text we first combined the title and text for real and fake news. Then in Phase-2 we used vectorization in Word2Vec, Word2Vec is one of the most popular technique to learn word embeddings using shallow neural network. Then we created Word2Vec model with genism. Then we feeded US presidents, etc to form some vectors. Further we tokenized the text and Checking the first 10 words of first news # every word has been represented with a number, made Histogram for no of words in news shows that most news article are under 700 words, and created a weight matrix. Then embedding vectors from word2vec and usings it as weights of non-trainable keras embedding layer and defined Neural Network. Then finally predict probability of news being real so converting into classes.

Lastly we used pre-trained Word2Vec explored them and got the required accuracies and F-1 Scores.

6. **Block Diagrams**:

Block Diagram for fake news detection in Previous research works.

1. **Framework Block-Diagram of Fake News detection with NLP and ML.**

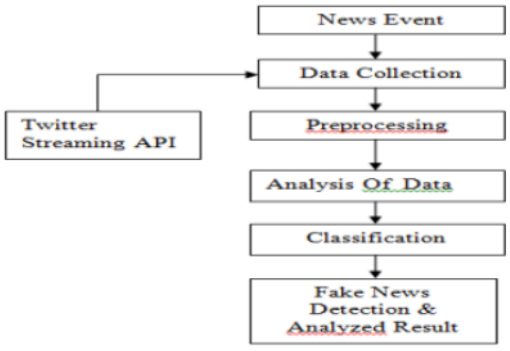

Figure 1: Framework Block-Diagram of Fake News detection with NLP and ML [6].

2. **Fake News detection using supervised Learning Algorithms**

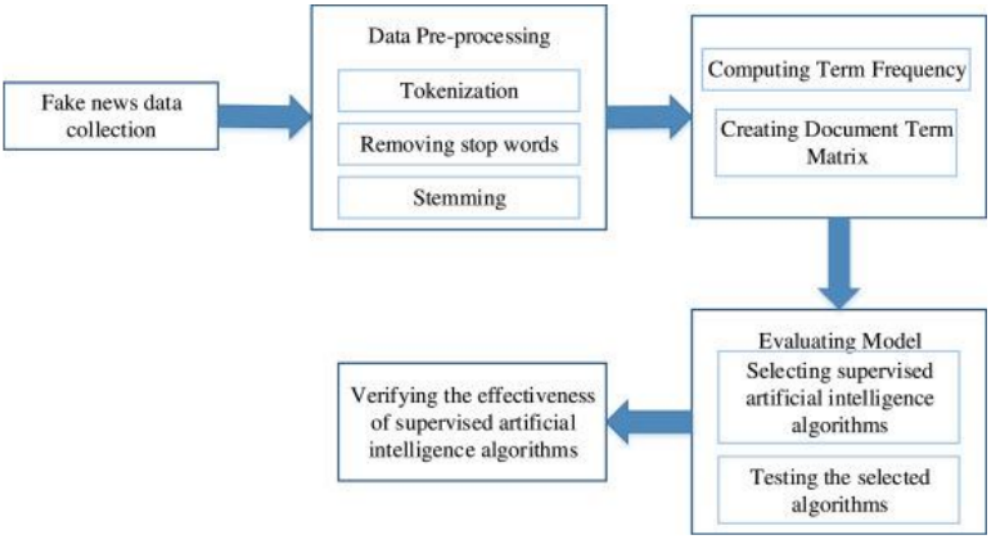

Figure 2: Fake News detection using supervised Learning Algorithms [7].

3. MBi-LSTM method in the Fake news detection method

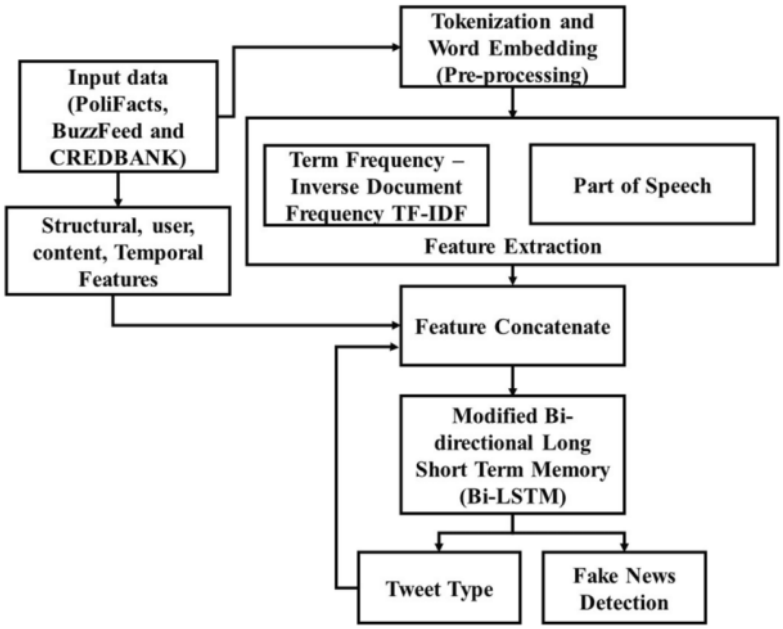

Figure 3: Block diagram of MBi-LSTM method in the Fake news detection method [8].

4. Fake news detection approaches:

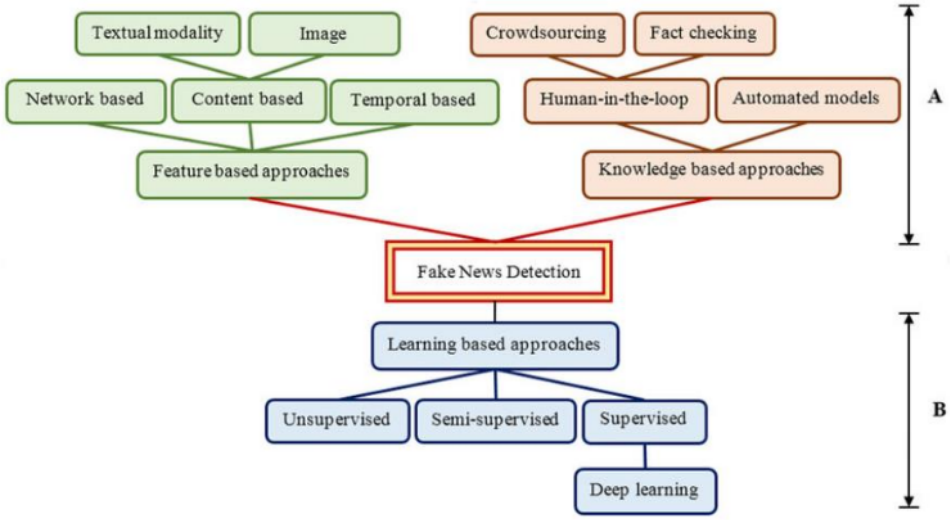

Figure 4: Fake news detection approaches [9]

5. Development of a smart system for fake news detection

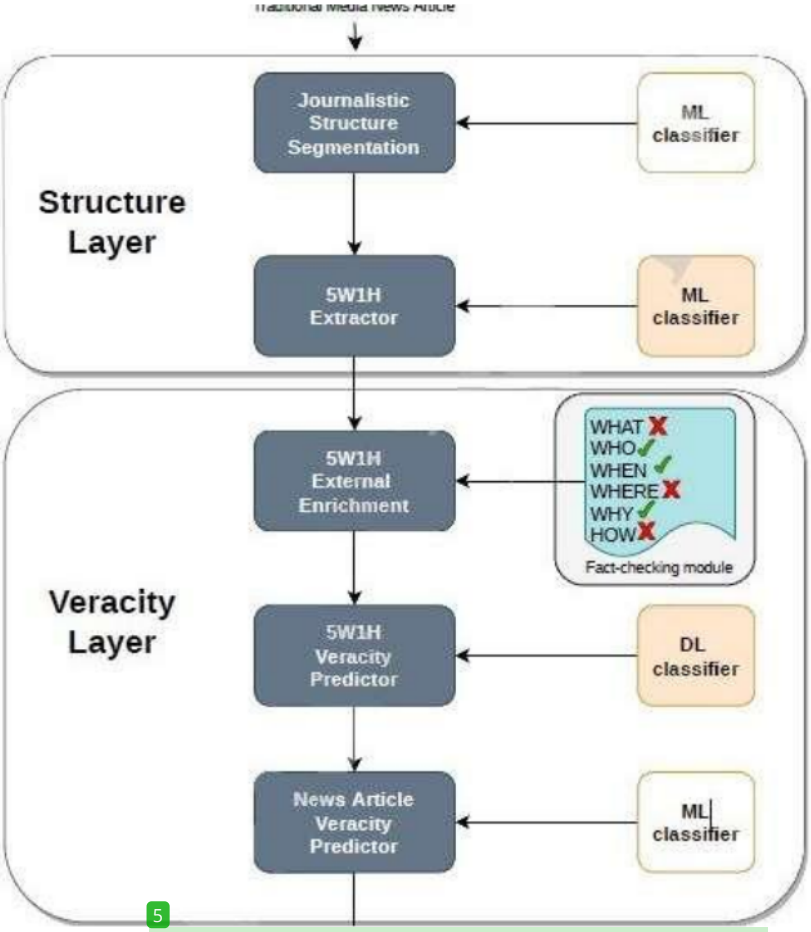

Figure 5: Development of a smart system for fake news detection [10]

6. **EchoFakeD: improving fake news detection in social media with an efficient deep neural network**

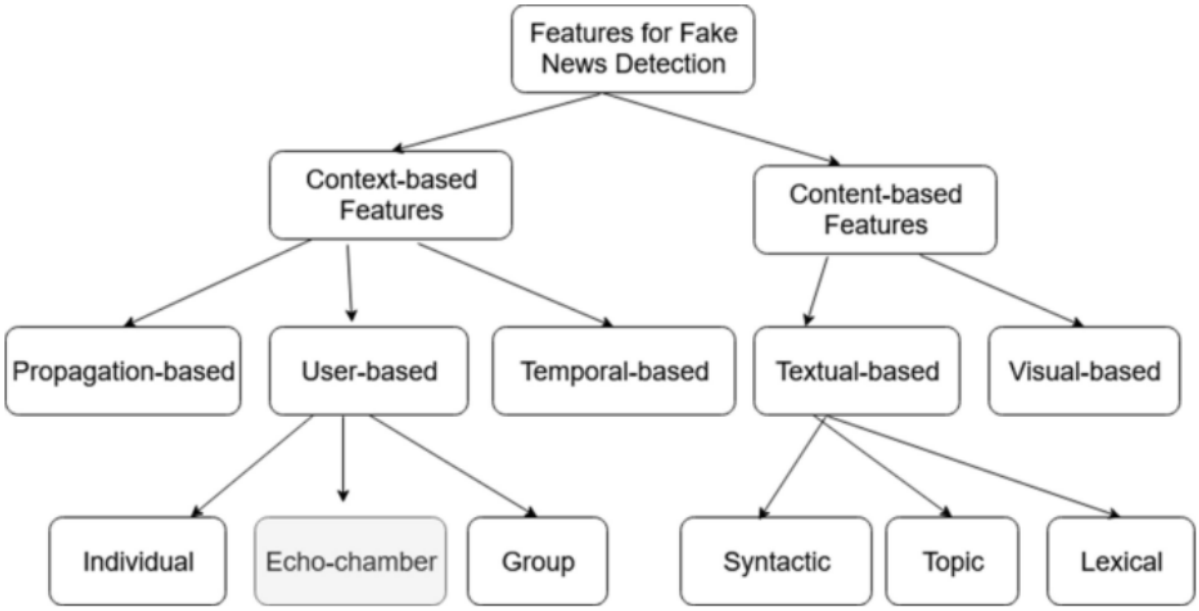

Figure 6: Features of Fake news detection [11]

7. Final Results

Along with a comparison from other top models Based on the chosen model's accuracy.

1. By selecting blogs as the solver and increasing the maximum number of iterations to 200, we can get an accuracy of 90.77 percent on the testing data for Logistic Regression with spaCy vectors as input features.
2. With spaCy vectors as input features, the decision tree classifier achieves an accuracy of 85.51 percent.
3. Logistic Regression with textual data as an input feature and Count Vectorizer andTFIDF Transformer in pipeline achieves 98.93% accuracy.
4. A Decision Tree Classifier using textual data as an input feature and Count Vectorizer and TFIDF Transformer in the pipeline achieves 99.63 percent accuracy.
5. Using 1000 input features, an artificial neural network constructed with Keras and TensorFlow backend with the Sequential Model and dense layers achieves an accuracy of 99.28 percent.
6. An Artificial Neural Network built using Keras and Tensorflow backends with the Sequential Model and dense layers achieves an accuracy of 99.21 percent on testing data with 10000 input features. Based on the F1 scores of the selected model, we can calculate the number of false positives and false negatives. Overall, we get an accuracy of 98.93%.
7. By selecting blogs as the solver and increasing the maximum number of iterations to 200, we can obtain a f1-score of 90.42 percent on the testing data for Logistic Regression with spaCy vectors as input features.
8. A decision tree classifier with spaCy vectors as input features achieves an f1-score of 84.49 percent.

9. Logistic Regression with textual data as an input feature and Count Vectorizer andTFIDF Transformer in the pipeline yields an f1 of 98.81%.
10. A Decision Tree Classifier using textual data as an input feature and Count Vectorizer and TFIDF Transformer in the pipeline produces an f1-score of 99.61 percent.
11. Keras with Tensorflow backend Artificial Neural Network with Sequential Model and Dense Layers supply us with an f1-score of 99.25 percent on the testing data with1000 input features.
12. An artificial neural network built using Keras and Tensorflow backends with the Sequential Model and dense layers achieve a f1-score of 99.17 percent on testing data with 10000 input features. Also, an F1 score comparable to that of accuracy demonstrates how accurate our model is in forecasting

Future Scope:

1. Gathering specific news-related tweets.
2. Using the same Decision Tree Classifier and pipeline, we will forecast whether ornot this is false news.
3. Another suggestion is to utilize the same for the credibility score assignment of acertain tweet, which is slightly different from this.

Table of Comparison

Table of comparison of fake news detection and clean is shown in Table 1.

Model/ Classifier	Epochs/ Iterations	Batch Size	Input Dimensions	Test Set Accuracy	Test Set F1-Score	Other Parameters
Logistic Regression	200		spaCy Vectorized Features	90.77	90.42	solver = lfbgs
Logistic Regression	200		Features from Count Vectorizer	98.93	98.88	solver = lfbgs
Decision Tree Classifier	-		spaCy Vectorized Features	85.51	84.49	criterion=entropy max_depth = 10 splitter=best, random_state=2020

Decision Tree Classifier	-		Features from Count Vectorizer	**99.63**	99.20	criterion=entropy max_depth =10 splitter=best random_state=2020
Neural Network 1	7	512	1000 Features from Count Vectorizer	99.19	99.16	optimizer=adam loss=binary_crossentropy
Neural Network 2	7	512	10000 Features from Count Vectorizer	99.24	99.12	optimizer=adam loss=binary_crossentropy

8. Conclusion

Using Pre-Trained Word2Vec Vectors in a pipeline with we've achieved maximum accuracy, recall, and F1 score of about 99%, but there is always room for improvement. Various things, such as tweaking the hyperparameters, can be done in the future to improve the accuracy and f1 score based on another dataset. The use of tensors mayenhance the overall results. Since the beginning of the project, the authors were enthralled while working, facing different challenges: may it be in terms of preprocessing the textual data, model training, or evaluation. Finally, the maximum accuracy and F1 score as per the models trained are achieved by using the Decision Tree Classifier in a pipeline with the Count Vectorizer and TFIDF Transformer, while with other models the team believes that there is always a scope of improvement accommodating various changes. Pretrained 300-D spacy model (en_core_web_lg()) can be used which can take the models having a lower accuracy to the next level, on the contrary, it will require a huge amount of system resources to be compromised. Nonetheless, more hyperparameter tuning can be done in other prospects depending on another dataset using, this project will help us in our future research inin which collection of certain news-based tweets can be performed using twtwinsnd then using the same Decision Tree Classifier with the said pipeline, it can predict whetherthat is a piece of fake news or not. Apart from this, another idea is to use the same for the credibility score assignment of a particular tweet. As a whole, the project will help the contributors build models or research.

9. References

1. Dataset Link: https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset
2. Ahmed H, Traore I, Saad S. “Detecting opinion spams and fake news using text classification”, Journal of Security and Privacy, Volume 1, Issue 1, Wiley, January/February 2018.
3. Ahmed H, Traore I, Saad S. “Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In: Traore I., Woungang I., Awad A. (eds) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science, vol 10618. Springer, Cham (pp. 127-138).
4. Fake News Vs Real News Dataset
5. The lightweight IBM Cloud Garage Method for data science
6. Dataset Link: https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset
7. Ahmed H, Traore I, Saad S. “Detecting opinion spams and fake news using text classification”, Journal of Security and Privacy, Volume 1, Issue 1, Wiley, January/February 2018.

8. Ahmed H, Traore I, Saad S. “Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In: Traore I., Woungang I., Awad A. (eds) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science, vol 10618. Springer, Cham (pp. 127-138).
9. Fake News Vs Real News Dataset
10. The lightweight IBM Cloud Garage Method for data science
11. Ahmed H, Traore I, Saad S. “Detecting opinion spams and fake news using text classification”, Journal of Security and Privacy, Volume 1, Issue 1, Wiley, January/February 2018.
12. Ahmed H, Traore I, Saad S. (2017) “Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In: Traore I., Woungang I., Awad A. (eds) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science, vol 10618. Springer, Cham (pp. 127-138).
13. Ahmed H, Traore I, Saad S. “Detecting opinion spams and fake news using text classification”, Journal of Security and Privacy, Volume 1, Issue 1, Wiley, January/February 2018.
14. Ahmed H, Traore I, Saad S. (2017) “Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In: Traore I., Woungang I., Awad A. (eds) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science, vol 10618. Springer, Cham (pp. 127-138).

ORIGINALITY REPORT

6%	2%	5%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to CSU, San Jose State University Student Paper	1%
2	Chetan Agrawal, Anjana Pandey, Sachin Goyal. "Fake news detection system based on modified bi-directional long short term memory", Multimedia Tools and Applications, 2022 Publication	1%
3	libx.bsu.edu Internet Source	1%
4	Mohamed K. Elhadad, Kin Fun Li, Fayez Gebali. "Chapter 86 A Novel Approach for Selecting Hybrid Features from Online News Textual Metadata for Fake News Detection", Springer Science and Business Media LLC, 2020 Publication	1%
5	Submitted to University of Wales Institute, Cardiff Student Paper	1%

6	"Combating Fake News with Computational Intelligence Techniques", Springer Science and Business Media LLC, 2022 Publication	1 %
7	www.safetylit.org Internet Source	1 %
8	Muhammad Shahid Anwar, Jing Wang, Sadique Ahmad, Wahab Khan, Asad Ullah, Mudassir Shah, Zesong Fei. "Impact of the Impairment in 360-degree Videos on Users VR Involvement and Machine Learning-Based QoE Predictions", IEEE Access, 2020 Publication	1 %
9	"Disinformation, Misinformation, and Fake News in Social Media", Springer Science and Business Media LLC, 2020 Publication	<1 %
10	"Data Engineering for Smart Systems", Springer Science and Business Media LLC, 2022 Publication	<1 %
11	doctorpenguin.com Internet Source	<1 %
12	dokumen.pub Internet Source	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		