Characterization, Classification and Detection of Fake News in Online Social Media Networks

Xavier Jose*, S.D Madhu Kumar† and Priya Chandran‡
Department of Computer Science and Engineering
National Institute of Technology Calicut
Calicut, India
Email: *xavier.jose55@gmail.com, †madhu@nitc.ac.in, ‡priya@nitc.ac.in

2021 IEEE Mysore Sub Section International Conference (MysuruCon) | 978-1-6654-3888-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/MYSURUCON52639.2021.9641517

***Abstract*—Due to its increasing popularity, low cost, and easy-to-access nature, Online Social Media (OSM) networks have evolved as a powerful platform for people to access, consume, and share news. However, this has led to the large-scale distribution of fake news, i.e., deliberate, false, or misleading information. Fake news is a pressing dilemma, as it has serious negative implications for individual users and for society as a whole. The news contents in the OSM networks are distributed rapidly, so the identification systems should predict news items as soon as possible to avoid spreading false news. Therefore, it is extremely crucial and technically challenging to detect fake news in social media networks. In this paper, we have discussed different characteristics and types of fake news and also propose an effective solution to detect fake news in OSM networks. The stance detection model and the fabricated content classifier are the main two components of the solution. The stance detection model achieved an accuracy of 90.37% with Logistic Regression, and the fabricated content classifier achieved an accuracy of 93.46% with Bi-directional LSTM.**

***Index Terms*—Online fake news, Social media analytics, Natural Language Processing, Machine learning, Fake news detection.**

I. Introduction

In all sectors, including journalism, Online Social Media (OSM) networks have emerged as an essential medium of communication. As social media platforms are becoming more popular, more users prefer to search for news in social media instead of conventional news sources. Although consuming news on social media offers many benefits, the absence of control and convenient access has led to the wide dissemination of fake news or misinformation in OSM networks. Fake news in online social media networks poses many challenges in multiple dimensions. The widespread distribution of fake news can have a significant adverse effect on users and society as a whole. Propagandists usually use fake news to transmit or propagate misinformation for economic or political benefits. Also, malicious user accounts like social bots can be used to publish and propagate fake news in online social media networks.

Fake news propagates quickly in online social media networks and reaches a wider audience in very little time. Considering the dynamic nature of social media networks, there is a severe demand for effective automatic real-time fake news detection mechanisms. Identifying relevant features that can distinguish fake news from real news makes fake news detection a challenging problem. Fake news identification on traditional news media depends predominantly on the features extracted from the news content only. But in the case of fake news in OSM networks, social context features can also be utilized along with news content features in recognizing fake news. The features extracted from the user's social engagements in news consumption and sharing on social media networks are termed as social context features [1].

The rest of this paper is organized as follows. Section II is a review of related works, and Section III presents the classification of fake news in OSM networks. The design of the proposed solution is described in Section IV, and Section V explains the implementation. Results and discussion are presented in Section VI, and Section VII makes concluding remarks and discusses future work.

II. Related Works

Over the past few years, fake news detection on social media has generated a lot of attention among researchers, and several attempts were made to address the problem. Most approaches proposed for fake news detection have been based on text classification using machine learning models. Various machine learning models like Logistic regression, Naive Bayes, Random forests, Decision Trees and Support Vector Machines have been built and trained to classify a particular news item as reliable or not [2]. Among the supervised machine learning algorithms, SVMs are one of the most widely used methods for classification in many research works concerning fake news detection. As per the experimental results of [3], SVMs have shown better accuracy than other supervised machine learning approaches for fake news prediction. Classification based on deep learning techniques are also found to be very useful in identifying fake news [4].

The research work of [5] shows that there are considerable differences between fake news and real news, particularly in the title of the news articles. The authors find that the length of fake news titles is more compared to real news titles and fake news titles used fewer nouns overall but more proper nouns and fewer stopwords. Also, the length of body text of fake news articles is less compared to real news articles and uses simple words, more occasional quotes, fewer punctuations, etc. They tried various machine learning algorithms, and the best performing model could achieve an accuracy of 91%. They

978-0-7381-4662-1/21/$31.00 ©2021 IEEE

conclude that real news articles persuade users through sound arguments while fake news convinces users through heuristics.

The authors in [6] try to come up with a solution that can identify sites containing misinformation based on *clickbaits*. They suggested that most fake news contains what is called *clickbaits*. *Clickbait* is an article, mostly sensational or having a catchy headline, targeted at clicking. The authors in [7] propose a prototype for identifying fake news items from twitter posts by using machine learning classifiers trained on datasets like the FNC-I [8]. The solution compares the main tweet object with its child objects if they agree on the topic or not. This might help to detect a fake tweet if one of the child objects discusses a different topic. On comparing different learning algorithms, SVM and Naive Bayes classifier gave better accuracy.

An approach based on n-gram analysis and machine learning techniques has been implemented in [9]. The paper presented a machine learning model for fake news using n-gram analysis with the help of different feature extraction techniques. Six different machine learning techniques were compared, employing two different feature extraction techniques. The best performing model was found to be the linear SVM classifier using unigram features.

The authors in [10] proposes a novel hybrid deep learning model that combines convolutional and recurrent neural networks for fake news classification. The classification accuracy of the model was significantly better than other non-hybrid baseline methods. A similarity-aware multi-modal method is proposed to predict fake news in [11]. In this method both textual and visual features of news content are extracted and investigated. The experimental results show that both multi-modal features and cross-modal similarity are important in fake news detection.

A novel automatic fake news detection model based on geometric deep learning is proposed in [12]. The model was trained and tested on news articles from Twitter. Experimental results show that social network structure and propagation are key features enabling highly accurate fake news detection. The results imply that propagation-based approaches can be used as an alternative to content-based approaches for fake news detection.

The following section presents the classification of fake news in online social media networks.

III. Classification of Fake News in OSM Networks

In general, "fake news refers to all kinds of false stories, rumors or news that are mainly published and distributed on the Internet, in order to mislead, befool or lure readers for financial, political or other gains" [13]. First Draft News project has identified seven types of fake news in its preliminary findings [14]. Following are the different types of fake news in OSM networks:

1) ***Satire/Parody****:* These are entertainment-oriented articles presented in conventional news media format mostly created without any intention to spread misinformation. But these articles can mislead readers when shared out of context.

2) ***Image/Video Manipulation****:* There can be mainly two types of image/video manipulation in news articles. One uses real images/videos that are not associated with the news content and tries to create a false association to mislead users purposefully. Other uses edited images/videos inside the news article to establish the claims made by the news.

3) ***Imposter content****:* Impersonation of legitimate/credible news sources, for example, by using an established news agency's branding.

4) ***Sponsored content****:* These are news article that claims to be unbiased media content when, in fact, it is public relations or advertising campaigns.

5) ***False connection****:* News articles in which the headline/title doesn't support/agree with the news content. This includes clickbait which contains a catchy/sensational headline but the content will be unrelated.

6) ***Misleading content****:* Selective reporting of real information by a news article to develop an issue or create a false narrative is called misleading content. For example, reporting only partially chosen real facts related to an incident to frame issues.

7) ***False context****:* These are news articles in which real facts are shared with incorrect background information.

8) ***Fabricated content****:* These are news stories that are completely made up but appear as legitimate news articles. It tries to mimic legitimate news articles but differs from legitimate journalism in writing styles/patterns, structural and other ways. These are intentionally produced for some benefits and not to report facts, so they frequently contain opinionated and biased language.

The design of the proposed solution is described in the following section.

IV. Design of Solution

From the types of fake news listed in Section III, our solution focuses on detecting false connection and fabricated content. The input to the system is the tweet containing the news article. The solution consists of two main phases. In the first phase, relevant words, phrases, and sentences are extracted from the news article using NLP techniques and query them in reputable/credible sources using Google News API. A credibility score is assigned based on whether the news content in the tweet agrees, disagrees, or unrelated to the news content from credible/reputable sources. In the second phase, the tweet is given to the machine learning models for classification. Figure 1 shows the design of the solution.

- *Input:* URL/Tweet Id of the tweet containing the news article
- *Output:* There will be four output parameters
  - List of news articles from credible sources related to the news content
  - Credibility score (Between 0 and 1)
  - False connection : Yes or No
  - Fabricated content : Yes or No

The steps involved are :

1) Fetching the tweet and social context information associated with the given tweet using Twitter Search API.
2) Using the links present in the tweet, extract the content from the pages using web scraping techniques, and append it to the news text.
3) The news text is converted to lowercase and tokenized into n-grams using the NLTK module. Stop-words are removed from the list of n-grams.
4) The remaining n-grams are reconstructed into the original strings and query them in reputable/credible sources using Google News API. This will return a list of news articles from credible sources related to our news content.
5) For each news content in the list, check whether the given news content in the tweet agrees, disagrees, or is unrelated to it using the stance detection model. Based on this a credibility score is assigned to the tweet.

   Let a be the number of news articles from credible sources that agree and d be the number of news articles from credible sources that disagree, then credibility score,

$$C = \begin{cases} 0 \text{ if } a + d = 0 \\ \frac{a}{a+d} \text{ otherwise} \end{cases}$$

6) Pass the headline/title and content of the news article to the trained stance detection model exposed as a REST web service. This will predict whether given headline-content pairs agree, disagree, discuss the same topic, or are not related at all.
   a) If the headline-content pairs disagree or are not related at all, then it is classified as false connection.
   b) Not false connection otherwise.
7) Classify the tweet containing the news article as fabricated content or not using the trained machine learning fabricated content classifier exposed as a REST web service. This is a binary classifier that outputs yes or no.

A. *Stance Detection Model*

Stance Detection is a natural language processing problem to detect the text's stance towards another target text. Stance detection is used in numerous applications such as textual entailment, opinion summarization, emotion detection, etc. Our stance detection model classifies body text stance towards a target into one of the four stance categories.

- *Input:* The news headline and the news body text (news content).
- *Output:* The stance of the news body relative to the claim asserted in the news headline:
  - *Agree:* The news body agrees with the news headline.
  - *Disagree:* The news body disagrees with the news headline.

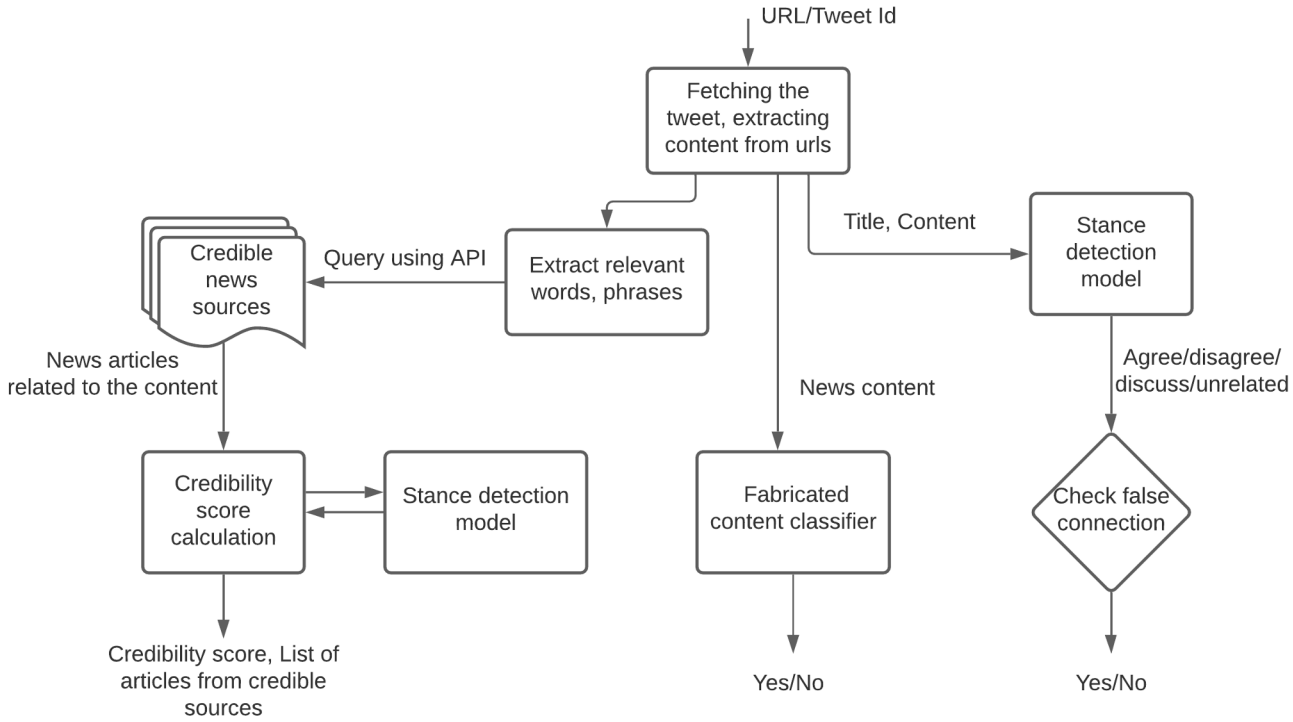

Fig. 1: Design of the solution

- *Discusses:* The news body discuss the same topic as the news headline.
- *Unrelated:* The news body text is not related to the news headline.

B. Fabricated Content Classifier

Fabricated content includes news stories that are completely made up but appear as legitimate news articles. It tries to mimic legitimate news articles but differs from legitimate journalism in writing styles/patterns, structural and other ways. As these are purposely produced for some benefits and not to report facts, they frequently contain biased or offensive language. Thus, to detect fabricated content, it is sensible to utilize linguistic features that extract various writing patterns and sensational captions.

- *Input:* The news text.
- *Output:* Classify the text of the news into one of the two classes:
  - *1:* The news text is fabricated content.
  - *0:* The news text is not fabricated content.

Following section explains the implementation of stance detection model and fabricated content classifier.

V. Implementation

A. Stance Detection Model

1) ***Dataset****:* The Fake News Challenge dataset [15], [8] have been used to train the model. The dataset contains 50,000 stance tuples. Every data item includes the news headline, body text, and the annotated stance that tells whether the headline's claim agrees, disagrees, discusses, or is unrelated to the body text. We divided the dataset into the train (90%), and test (10%) sets for constructing the model.

2) ***Preprocessing****:* The input text is first converted to lower case and removed punctuations, stop words, and non-alphanumeric characters. Natural Language Toolkit (NLTK) library of python is used for these preprocessing tasks.

3) ***Feature Extraction****:* We used the scikit-learn function *TfidfVectorizer* for converting the text data to a matrix of TF-IDF features.

4) ***Training****:*

a) Logistic Regression: The Logistic Regression model is trained with the training data set. The scikit-learn implementation of Logistic Regression is used with parameter *class_weight='balanced'*.

b) Decision Tree Classifier: The scikit-learn implementation of decision tree is used with parameters *criterion='gini'* and *min_samples_split=2*. The model is trained with the training data set.

c) Random Forest Classifier: The scikit-learn implementation of random forest is used with parameters *n_estimators=100*, *criterion='gini'* and *min_samples_split=2*. The model is trained with the training data set.

d) Multinomial Naive Bayes Classifier: The scikit-learn implementation of multinomial Naive Bayes classifier is used with parameters *alpha=0.3* and *fit_prior='True'*. The model is trained with the training data set.

e) Support Vector Machine: The scikit-learn implementation of SVM is used with parameter *kernel='linear'*. The model is trained with the training data set.

B. Fabricated Content Classifier

1) ***Dataset****:* The model is trained using the Fake News dataset [16]. The Fake News dataset is a collection of fake news and truthful articles, collected from different legitimate news sites and sites flagged as unreliable by fact checking websites. These fake news articles include news contents that are completely made up but appear as legitimate news articles. We used the *text* (text of the news article) field in the CSV file for training the model.

2) ***Preprocessing****:* The input text is first converted to lower case and removed punctuations and non-alphanumeric characters. Words are converted to their base form using *WordNetLemmatizer*. Natural Language Toolkit (NLTK) library of python is used for these preprocessing tasks.

3) ***Feature Extraction****:* We map each news text into a high-dimensional vector space using the technique called word embedding. Each word in the corpus is mapped to a real-valued vector in an n-dimensional vector space. *Keras Embedding layer* provides a suitable way to convert words into word embeddings. One-hot encoding has to be done before embedding text data. *Keras* provides the *one_hot()* function that creates efficient integer encoding of each word in the document. The vocabulary size used is 30000. We are using word embeddings as the input to the LSTM layer. It aims to map semantic meaning into a real vector domain and is an improvement over traditional approaches such as *bag-of-word* encoding schemes. We used the *Keras Embedding layer* as the first layer of the network with *input_dim=vocabulary size, output_dim=100,* and *input_length=1000*.

4) ***Training****:*

a) Long Short-Term Memory - Recurrent Neural Network:

LSTMs are very effective solution for sequential input like text input. We used the *Tensorflow Keras* implementation of the LSTM RNN for building the model. We split the dataset into the train (90%) and test (10%) sets.

The first layer is the *Keras Embedding layer* that uses 100-dimensional vectors to encode each word into the real vector domain. The second layer is the LSTM layer with 100 units. The next layer is the dense output layer with one neuron. Since this is a binary classification, we used the *sigmoid activation function* in the output layer to make 0 or 1 predictions for the two classes. For binary classification, *binary_crossentropy* is used as loss function along with *ADAM optimization algorithm*. Figure 2 shows the model architecture plot. Training is done iteratively for ten epochs with batch_size=64 to minimize loss function and to improve accuracy.

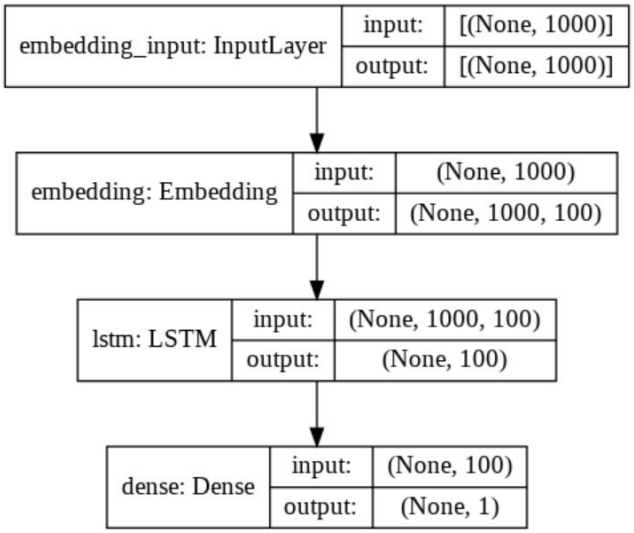

Fig. 2: LSTM Model Architecture Plot

b) Bi-directional Long Short-Term Memory - Recurrent Neural Network:

The first layer is the *Keras Embedding layer* that uses 100-dimensional vectors to encode each word into the real vector domain.

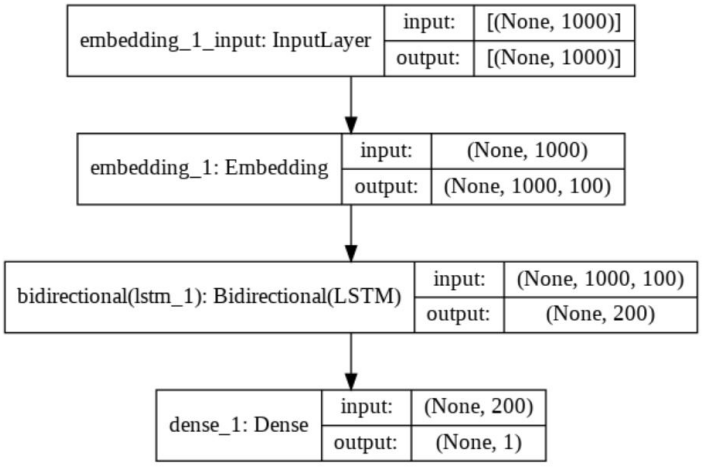

Fig. 3: Bi-directional LSTM Model Architecture Plot

The second layer is obtained by wrapping the LSTM layer (100 smart neurons) with a Bi-directional layer. The next layer is the dense output layer with one neuron. Since this is a binary classification, we used the *sigmoid activation function* in the output layer to make 0 or 1 predictions. For binary classification, *binary_crossentropy* is used as loss function along with the *ADAM optimization algorithm*. Figure 3 shows the model architecture plot. Training is done iteratively for 20 epochs with batch_size=64.

VI. Results and Discussion

A. *Stance Detection Model*

As discussed in Section V-A, we trained models with five different machine learning algorithms using the Fake News Challenge dataset [15], [8]. The testing is done using the test dataset. Below figures shows the confusion matrix of Logistic Regression, Decision Tree, Random Forest, Multinomial Naive Bayes and SVM Classifier.

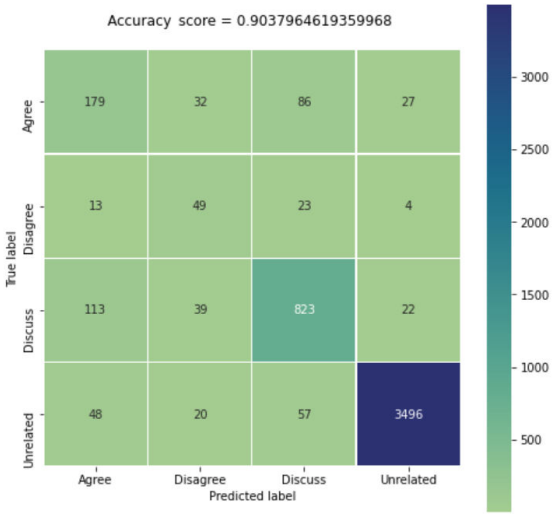

Fig. 4: Confusion Matrix of Logistic Regression

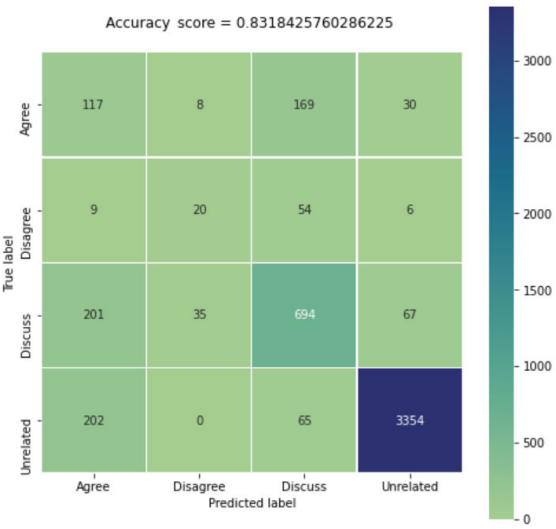

Fig. 5: Confusion Matrix of Decision Tree Classifier

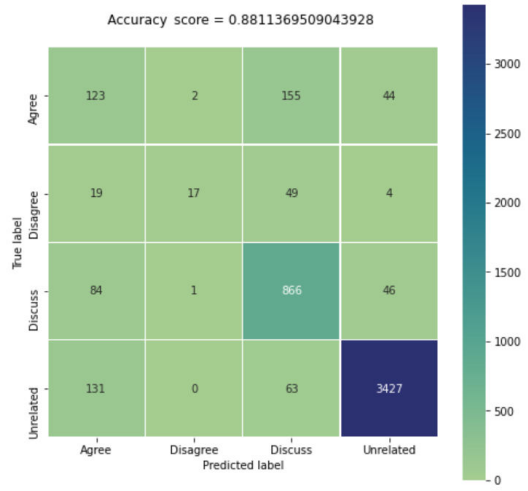

Fig. 6: Confusion Matrix of Random Forest Classifier

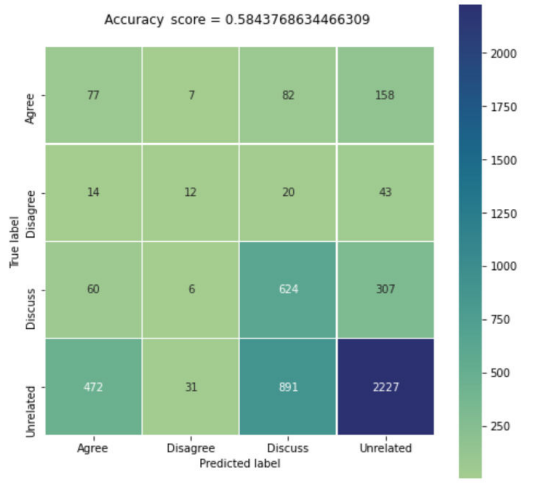

Fig. 7: Confusion Matrix of Multinomial Naive Bayes Classifier

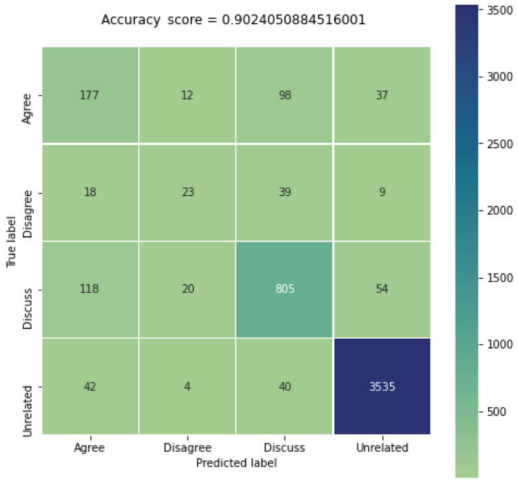

Fig. 8: Confusion Matrix of SVM Classifier

Out of the algorithms tried for stance detection, logistic regression and Support Vector Machine(SVM) show the best performance with accuracy above 90%. Random forest algorithm also performs well, with an accuracy of around 88%. Table I shows the comparison of the algorithms.

Algorithm	Accuracy
Logistic Regression	0.903
Decision Tree	0.831
Random Forest	0.881
Multinomial Naive Bayes	0.584
SVM	0.902

Table I. Comparison of Performance of Different Algorithms

B. Fabricated Content Classifier

As discussed in Section V-B, Long Short-Term Memory Recurrent Neural Network and Bi-directional Long Short-Term Memory Recurrent Neural Network are trained using Fake News dataset.

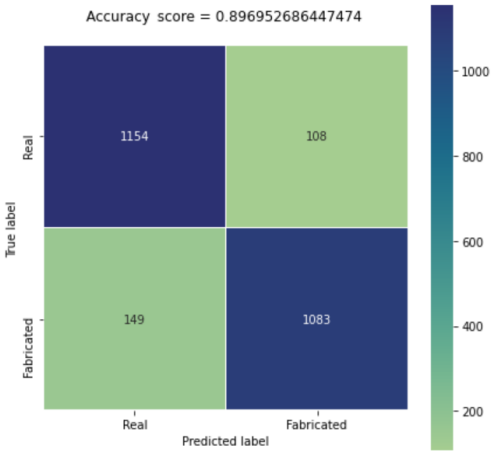

Fig. 9: Confusion Matrix of LSTM

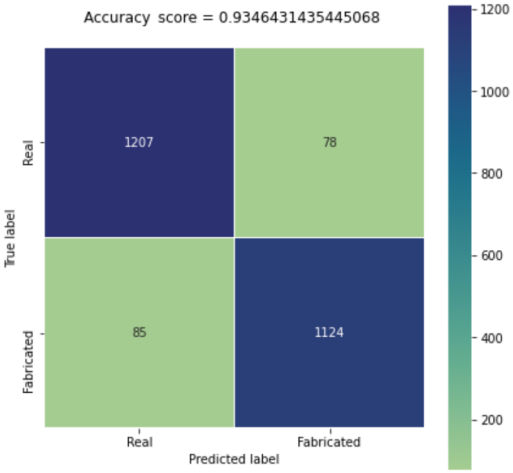

Fig. 10: Confusion Matrix of Bi-directional LSTM

Figure 9 shows the confusion matrix of LSTM. Figure 10 shows the confusion matrix of Bi-directional LSTM. Bi-directional LSTM shows the best performance with an accuracy of 93.46%. Table II shows the comparison of the models.

Model	Accuracy
LSTM RNN	0.896
Bi-directional LSTM RNN	0.934

Table II. Comparison of Performance of Two Models

VII. Conclusion and Future Work

Over the past few years, the use of social media to propagate fake news has increased tremendously. This has numerous negative consequences, and hence there is an urgent need for powerful automatic fake news identification mechanisms. This paper has discussed different characteristics and types of fake news in OSM networks. Also, we have proposed an effective solution to detect fake news, particularly for false connection and fabricated content identification in OSM networks. The stance detection model and the fabricated content classifier are the main two components of the solution. We tried different machine learning algorithms for implementing both models, and the best-performing ones are chosen for building the solution. The stance detection model achieved an accuracy of 90.37% with Logistic Regression, and the fabricated content classifier achieved an accuracy of 93.46% with Bi-directional LSTM.

The proposed solution focuses only on detecting false connection and fabricated content out of the eight types of fake news found in OSM networks. However, as future work, the solution can be improved by developing more machine learning models or techniques for detecting other types of fake news. It might be technically challenging to detect some types while others, like image/video manipulation, can be accurately identified. Also, we can extend the solution to support multiple languages. Here the major challenge is preparing the datasets and building the models as there are no publicly available fake news datasets for most regional languages.

References

[1] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.

[2] Alessandro Bondielli and Francesco Marcelloni. A survey on fake news and rumour detection techniques. *Information Sciences*, 497:38–55, 2019.

[3] Hu Zhang, Zhuohua Fan, Jiaheng Zheng, and Quanming Liu. An improving deception detection method in computer-mediated communication. *Journal of Networks*, 7(11):1811–1816, 2012.

[4] Natali Ruchansky, Sungyong Seo, and Yan Liu. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore*, pages 797–806, 2017.

[5] Benjamin Horne and Sibel Adali. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the International AAAI Conference on Web and Social Media, Montreal, Canada*, volume 11, 2017.

[6] Monther Aldwairi and Ali Alwahedi. Detecting fake news in social media networks. *Procedia Computer Science*, 141:215–222, 01 2018.

[7] Ehesas Mia Mahir, Saima Akhter, Mohammad Rezwanul Huq, et al. Detecting fake news using machine learning and deep learning algorithms. In *2019 7th International Conference on Smart Computing & Communications (ICSCC), Sarawak, Malaysia*, pages 1–5. IEEE, 2019.

[8] Fake News Challenge Stage 1 (FNC-I): Stance Detection. http://www.fakenewschallenge.org/. Online; accessed: 2021-06-12.

[9] Hadeer Ahmed, Issa Traore, and Sherif Saad. Detection of online fake news using n-gram analysis and machine learning techniques. In *International conference on intelligent, secure, and dependable systems in distributed and cloud environments, Vancouver, Canada*, pages 127–138. Springer, 2017.

[10] Jamal Abdul Nasir, Osama Subhani Khan, and Iraklis Varlamis. Fake news detection: A hybrid cnn-rnn based deep learning approach. *International Journal of Information Management Data Insights*, 1(1):100007, 2021.

[11] Xinyi Zhou, Jindi Wu, and Reza Zafarani. SAFE : Similarity-aware multi-modal fake news detection. *Advances in Knowledge Discovery and Data Mining*, 12085:354, 2020.

[12] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*, 2019.

[13] Xichen Zhang and Ali A Ghorbani. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2):102025, 2020.

[14] Fake news. It's complicated. https://firstdraftnews.org/articles/fake-news-complicated/. Online; accessed: 2021-07-25.

[15] FakeNewsChallenge/fnc-1. https://github.com/FakeNewsChallenge/fnc-1. Online; accessed: 2021-06-12.

[16] Fake News Dataset. https://www.kaggle.com/c/fake-news/data. Online; accessed: 2021-06-12.