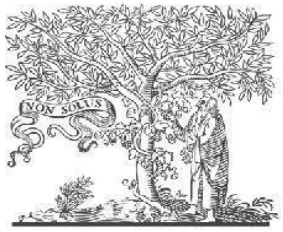
International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

COPY RIGHT



ELSEVIER SSRN

IJIEMR Transactions, online available on 8th Aug 2024. Link

https://ijiemr.org/downloads.php?vol=Volume-13& issue=issue08

DOI: 10.48047/IJIEMR/V13/ISSUE 08/17

Title Detecting Fallacy of Social Media Data by using Content Based Features

Volume 13, ISSUE 08, Pages: 124 - 127

Paper Authors
Dr. G.V. Ramesh Babu , Sampangii Mahesh

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper as Per UGC Guidelines We Are Providing A ElectronicBar code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Detecting Fallacy of Social Media Data by using Content Based Features

Dr. G.V. Ramesh Babu
Associate Professor, Department of Computer Science, Sri Venkateswara University, Tirupati
gvrameshbabu74@gmail.com
Sampangii Mahesh
Master of Computer Applications,
Sri Venkateswara University, Tirupati
rohitmahesh02112000@gmail.com

Abstract
With the rise in popularity of social media, there has been an increase in the amount of information available. This is especially clear in the material about the COVID -19 pandemic. Despite the fact that information about the epidemic is constantly updated, there is still a lot of misunderstanding, with unverified claims thrown into the mix. Misinformation during this time should be closely monitored, as it can cause havoc in society if allowed unchecked. The proposed research conducts a thorough comparison of the performance of various machine learning models for classifying fake news. The study is based on a dataset of 1100 COVID 19-related posts gathered from social media. We test a feature engineering strategy that uses content-based approaches that are meaningful and have the potential to provide new insights and improve machine learning predictions. The content based features are applied to all the classifiers , Neural Network Classifier has given 0.95% on test data.

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. There is major scope for misinformation propagation in today's digital world. "We're facing an infodemic, not just an epidemic," says the author. As WHO Director-General Tedros Adhanom Ghebreyesus stated on the 15th of February 2020 at the Munich Security Conference, [1]. The COVID-19 epidemic is rapidly spreading at an alarming rate over the world. It has also been alleged that misinformation is aiding the spread of COVID-19. [2]. To address this health crisis we need to keep a check on the spread of misinformation related to COVID-19. If not handled it has potentially dangerous consequences. A potential use case in this scenario is to develop classifiers and techniques to stem this flow. A data set containing information related to COVID 19 is collected and different machine learning algorithms are applied and the performance has been evaluated on different classifiers in classifying the fake from original information. The content based features are applied to all the classifiers , Neural Network Classifier has given 0.95% on test data with content based features.

Related Work

This problem of misinformation propagation during this situation could really have adverse effects on the society. 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 overview of the ongoing inventions related to the fake news. In

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

addition, they portrayed the effect of online fakenews, introduced stateof-the-art detection recognition strategies, and talked about the ordinarily utilized datasets utilized to build models to classify fake news.Collins and Erascu [4] also gave an review of different models 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 any one of them. Al Asaad et al. [5] proposed a new model to check the validity of the news using machine learning algorithms. The compare the effectiveness of the models with used different algorithm like Multinomial Naive Bayes and Lagrangian Support Vector Machine classification algorithms. 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 chose half and half highlights from meta information of information records. Wang [7] presented a new dataset for counterfeit news. The dataset comprises of 12.8K physically marked short explanations in different settings from PolitiFact.com. FakeHealth [8] is gathered from medical care data audit site Health News Review, which surveys whether a news is palatable from 10 measures. 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 Center for Disease Control and Prevention (CDC).

Methodology

To start with, we collected over 1,100 news articles and posts on social media related to COVID-19 pandemic from a variety of new sources. We present the steps of our proposed model in the form of an architectural framework. Then again, we talk about our dataset, information preprocessing and highlight extraction and characterization model choice. The steps of our proposed model are displayed in Fig. 1.

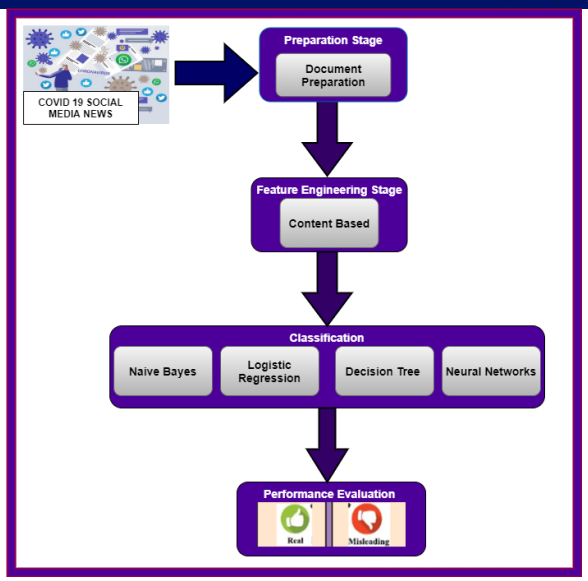


Fig. 1: Architectural Framework

Data Collection and Preprocessing

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 noise removal process is being done during this stage.

Feature Extraction

To classify news articles, raw text data must be transformed into something more usable. This is known as feature extraction. Word counts, n-gram counts, punctuation usage, sentiment analysis, term frequency-inverse document frequency, and many other features can be extracted. The extracted features can then be used to classify the article from which they were extracted. By determining the best features for classifying process can be more accurate. To study the fake and the original news we used a feature engineering technique.

Content based features:

The content based features are extracted from the COVID 19 news data set . The main focus here is the stylistic features that are based on text style, syntax and components related to grammar. This process is carried out using POS tagger which uses NLTK to

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

monitor the tag frequency.. For example, the number of nouns, proper nouns, verbs, determinants, comparatives and superlatives and so on. Along with this, we use the Word Count dictionaries to keep track of the frequencies of negation, belief, surprise, conditional, modal, existential, and interjection words.

Classification: The extracted features will be to train classifiers. The classifiers used here are Naive Bayes, Logistic Regression, decision tree and Neural Network

Evaluation: The performance of each classifier must be assessed. A common simple metric used is accuracy, this is not appropriate for unbalanced datasets. Other methods which will be explored include the confusion matrices, F1 score, precision and recall.

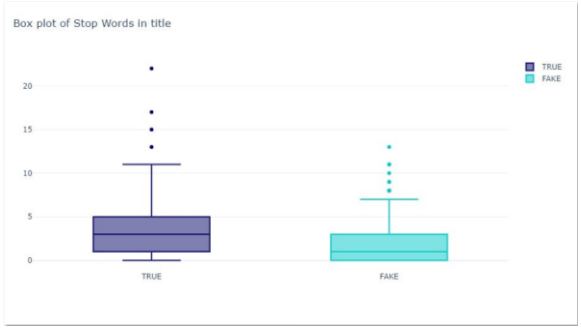
Results and Discussion

Exploratory Data Analysis (EDA): This the data set used for analysis. It consists of 1164 rows in which each of those has 4 attributes. We have a balanced data set because we have both fake and true data equally. Number of fake data: 575 Records containing true data: 584

Feature Engineering:

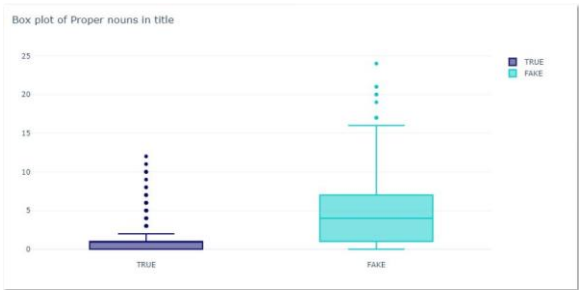
Content based features are formulated using

Stop Words in Title: As the lengths of the articles vary greatly, compute the percentage of stop words in each article body rather than simply counting the number.



Fake news titles have fewer stop-words than those of real news.

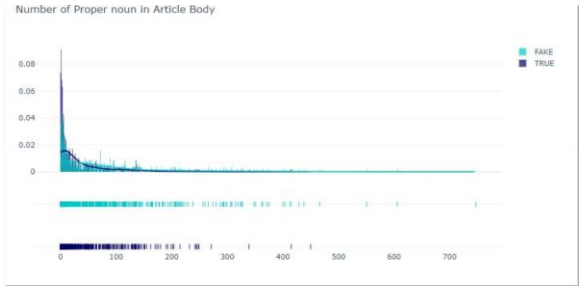
Proper Noun in Title: Count how many proper nouns (NNP) are in each title. Titles for fake news contain more proper nouns. The use of proper nouns in titles appears to be very important in distinguishing fake from real.



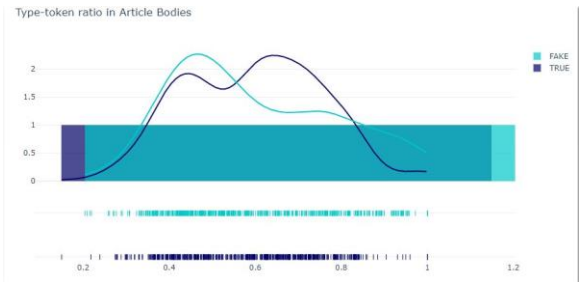
Overall, these findings indicate that fake news writers are attempting to attract attention by using all capitalized words in titles, as well as cramming as much substance into the titles as possible by skipping stop-words and increasing proper nouns.

The stylistic features captured here are:

Like the count on number of nouns



Type-Token Ratio (TTR), is the total number of unique words (types) divided by the total number of words (tokens) in a given segment of language.



Classifiers
The confusion Matrix for Decision Tree Classifier
TP=108; TN=109;FP=10;FN=5;Classifier Accuracy:217/233=0.935

Performance evaluation: Accuracy of all the classifiers

Table 1. 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

The content based features are applied to all the classifiers as shown in Table 1. Neural Network Classifier has given 0.95% on test data with content based features.

Conclusion

In this paper, the methodology in detecting misleading information related to the COVID-19 outbreak is applied. 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 the topic is in the feature engineering stage where an integration of content based is being used to extract the features. The classification algorithms are used and the performance is being assessed. In general, the results demonstrated the validity of our collected ground-truth data and provided useful insights into the performance of various classification algorithms on them. Considering the accuracy neural network classifier could give best results compared to the other three algorithms.

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.
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.