Fakebuster: Fake News Detection System Using Logistic Regression Technique In Machine Learning

Muhammad Syahmi Mokhtar, Yusmadi Yah Jusoh,NoviaAdmodisastro, NorainiChe Pa, AmruYusrinAmruddin

Abstract: The uncontrollable spread of fake news through the net is irresistible in this globalization era. Fake news dissemination cannot be tolerated as the bad impacts of it to the society is really worrying. Furthermore, this will lead to more significant problems and potential threat such as confusion, misconceptions, slandering and luring users to share provocative lies made from fabricated news through their social media to occur. Within Malaysia context, there is lack in platform for fake news detection in Malay language articles and most of Malaysians received news through their social messaging applications. Fake news can be certainly solved by the aid of artificial intelligence which includes machine learning algorithms. The objective of this project is to propose a fake news detection model using Logistic Regression, to evaluate the performance of Logistic Regression as fake news detection model and to develop a web application that allows entry of a news content or news URL. In this study, Logistic Regression was applied in detecting fake news. Model development methodology is referenced and followed in this project. Based on existing studies, Logistic Regression showed a good performance in classification task. In addition, stancedetection approach is added to improve the accuracy of the model performance. Based on analysis made, this model within stance detection approach yields an excellent accuracy using TF-IDF feature in constructing this fake news model. This model is then integrated with web service that accepts input either news URL or news content in text which is then checked for its truth level through "FAKEBUSTER" application.

Keywords: Malay Language, Dissemination, Machine Learning, Logistic Regression, Stance Detection, Fake News

I.INTRODUCTION

Fake news dissemination cannot be compromised as it able to deliver negative effects to the public for a long run. Problematic issues might arise from fake news such as slanders, confusion and misconceptions and provocative lies until up to a level sentiment issues played by irresponsible parties or individuals who love to spread hatred and havoc among one after another.

Muhammad Syahmi Mokhtar, Faculty of Computer Science and Information Technology,Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

Yusmadi Yah Jusoh, Faculty of Computer Science and Information Technology,Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

Novia Admodisastro, Faculty of Computer Science and Information Technology,Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

Noraini Che Pa, Faculty of Computer Science and Information Technology,Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

Amru Yusrin Amruddin, MIMOS BerhadJalanInovasi 3, Technology Park Malaysia, 57000 Kuala Lumpur

A news that has been manipulated or fabricated in its contents with things that are unrelated, fully or partly false is categorized as fake news.Identifying the news is being fake or not fake is really difficult.

However,within the capability of artificial intelligence in machine learning, this is made possible to detect fake news. Several countries have displayed their commitment in dealing with fake news due to its anticipated impact to the society. In other words, fake news is really deceiving and influencing the people to believe on something that is not true and probably have been manipulated.

II.RELATED WORKS

There are some past researches conducted in conjunction with fake news detection in various platform which uses the capability of Logistic Regression to make predictions in text classifications. Below are some existing findings that described the usage of this machine learning algorithm. Below showed past research papers conducting fake news detection.

Based on the data obtained in the table 1, this shows by using Logistic Regression in creating model that detects fake news yields in a good significant of accuracy. It is proven that Logistic Regression is quite good in solving binary classifications due to its predictive power in probability values. Logistic Regression detection model works well in dealing with long and also short input text and the range of accuracy can be achieved is within 79.0% to 89.0% based on the data on the table. By using Term Frequency–Inversed Document Frequency (TF-IDF), it is proven that this is a good feature to be used in text preprocessing task due to its capability. Therefore, TF-IDF is used as feature required for text vectorization so that texts of strings can be converted into correct format which is going to be fit into data trainings and testing in the development.

Retrieval Number: A2633109119/2019©BEIESP
DOI: 10.35940/ijeat.A2633.109119
Journal Website: www.ijeat.org

2407

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Fakebuster: Fake News Detection System Using Logistic Regression Technique In Machine Learning

Table. 1Related works based on research papers conducting fake news detection.

| Author(s)	Input Length	Dataset	Features (Attributes test)	Model Performance		References
				Accuracy	F1-Score	
Samir Bajaj	Long text	Signal Media News	GloVe	N/A	0.65	[1]
Svitlana et. al	Short text	Twitter	Tf-Idf	79.0%	0.89	[2]
Hadeer Ahmad	Long text	PolitiFact	Uni-gram, Tf-Idf	89.0%	N/A	[3]

Note: Accuracy – the ratio of correctly predicted observations to the total of observations.
F1-score – performance metrics.

Definition of Fake News

In understanding fake news might provide various and subjective answers which are acceptable to be understood. By separating the term "fake news" might be a helping hand here. The word "fake" defines as not genuine, counterfeit or a forgery of something, [4] meanwhile the word "news" is defined as a report of current events which has just happened or will occur in the future. The phrase "fake news" refers to false information or statement or propaganda published in resemblance of being authentic news [5]. Based on Malaysian fake news bill launched in April 2018,fake news includes any news, information, data and reports, which is or are wholly or partly false, whether in form of features, visuals or audio recordings or in any other form capable of suggesting words or ideas [6]. Maximum fine of RM500,000 and six year of jails are punished to anyone who break this legislation.

Types of Fake News Detection Approaches

Below are some approaches that have been conducted by past researchers in this particular area of fake news detection.
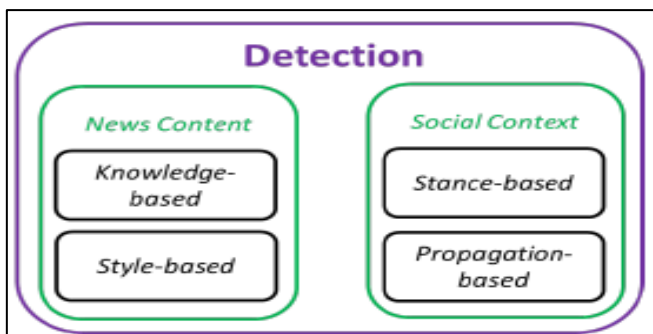

Fig. 1 Types of fake news detection approaches [7]

III.METHODOLOGY

There are various of methodology of software development life cycle (SDLC) can be chosen on as a basis to build a software. Therefore, since this project is about building a fake news detection model, model development methodology is referenced and followed. This project is not the same as developing other typical software system as the focus of it is towards model development in machine learning. Machine learning requires a lot of time for model training and model testing and also a huge and good quality of dataset as well.In other words, the model is counted as good in accuracy if the model produces predicted outcome that is the same with the actual outcome. As an illustration, the model able to predict and classify correct class of news which is either fake or not fake. Below shows flow of diagram regarding distinct phases that involved in model development in machine learning.
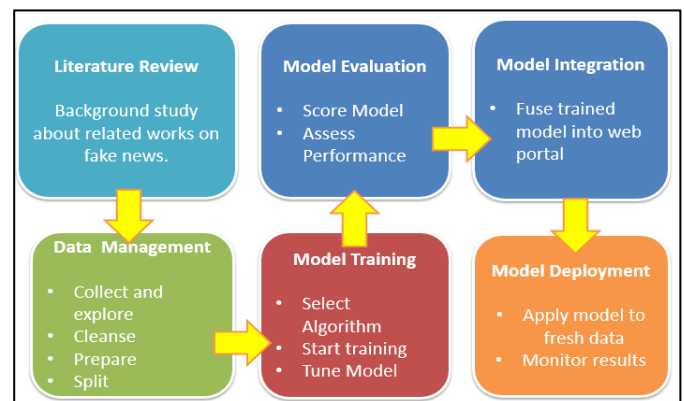
Fig. 2 Adopted machine learning model development activities [8]

Literature Review

First of all, background studies are conducted in order to understand what fake news really means. Researches are made through lots of readings of research papers and understanding regarding the underlying concepts of fake news and artificial intelligence which comprises of Natural Language Processing and Machine Learning as examples. Starting from here, detailed technique can be identified and the main concept can be grasped. The idea and concept of developing machine learning model is carefully understood.

Data Management

In data management phase, a set of data (dataset) is collected which is set of news articles. Once the dataset is collected, the data is explored to get a better knowledge of its structure and meaning. The dataset is then needed to be cleansed which means in order to make it useful, through vectorization process or data pre-processing phase. Term Frequency-Inverse Document Frequency (TF-IDF) is used in this process to convert all those sentences of articles into a structured format. When it is cleansed, the data is going to be prepared for loading into programming environment. Finally, the data is going to be separated into trainings and validation subsets.

Model Training

After the data is properly prepared, the machine learning model is ready to be trained. In this Model Training stage, overall approaches were considered and a learning task is decided which is a prediction task. Available features in the training data set is then studied. Then, appropriate algorithm is selected to train the model. In this case, Logistic Regression is chosen. Dataset is fit into the algorithm for training and testing purposes. The model is then saved as .sav file.

Model Evaluation

In evaluating the model, the output of the model produced is measured. Scoring of the model is conducted using certain performance metrics like F1 score, precision, recall and accuracy rate which is based on confusion matrix report. Some adjustments can be made within the model until satisfaction is achieve in making the model yields in good accuracy of output in order to proceed with the final stage.

Model Integration

Here, the model is integrated into web portal to allow entry of news contents or news URL. The saved model is passed through API service which act as web service to the portal.

Model Deployment

In last stage, the model is applied to new or fresh data and outcomes are monitored. In this phase, every results that are created are taken into close observations and recorded. The model still can be improved if necessary.

IV.RESULT ANALYSIS AND DISCUSSIONS

Result Analysis on Developed Model

Along the journey of development, some interesting findings are found which there is a major difference between stance-based and content-based. These two different approaches yield a very different accuracy from each other and one outperformed another. Below is the accuracy of model achieved by comparing approaches between stance-based and content-based using 3 levels of TF-IDF for text preprocessing in model evaluation. This clearly shows stance-based outperformed content-based. Therefore, the model is saved within stance based.

Content Based		
Accuracy	**Training**	**testing**
word	0.773	0.562
ngram	0.807	0.590
char	0.727	0.530
Stance Based		
Accuracy	**Training**	**testing**
word	0.996	1.000
ngram	0.996	0.996
char	0.992	0.976

Fig. 3Accuracy of Model in Content-Based and Stance-Based

Limitations and Constraints

In developing this fake news detection model, there are some obstacles that are difficult to be experienced with. This model is not an excellent model due to some concrete reasons. To be specific, since this model is a classification model or in other words it is a supervised model which is able to detect news to be fake or not. In this project scenario, the data collected which are the news articles of Malay language is insufficient and not quality enough to fit the model for data trainings. This is because, the model is solely based on past trainings of data for predicting future news. If the dataset is small, it can be trained but it would not be able to predict correctly. Although the accuracy of model performed outstandingly within stance based as shown in Figure 3, but unfortunately it is only good in testing but not in new fresh data. As an illustration, when the model detects fresh news data, it predicts the news label wrongly which is not supposed to be it.

V.CONCLUSION

Vast spreading of fake news through the net will deliver bad impacts to the society. Fake news will mislead readers and deceive them to the ultimate confusion in believing something that is not true to be true. This is the danger of fake news as people nowadays still are unable to differentiate between fake news and real news in their daily life with their naked eyes. However, this problem can be certainly solved by harnessing the power of machine learning to predict news to be fake or not. Within this capability, at least will help people to be more aware on the news they obtained by checking it first via FAKEBUSTER which is embedded with a detection model into it before sharing it. In addition, this system has been completed in this final year project in which certainly needs more improvements in the near future.

REFERENCES

1. Bajaj, S. (2018). "The Pope Has a New Baby!" Fake News Detection Using Deep Learning. [online]1,pp.5-6.Availableat:https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2710385.pdf
2. Volkova, S., Shaffer, K., Yea Jang, J. and Hodas, N. (2017). Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter. [online] pp.3-4. Available at: https://www.cs.jhu.edu/~svitlana/papers/VSYH_ACL2017.pdf [Accessed 26 Jul. 2018].

3. Ahmad, H. (2018). Detecting Opinion Spam and Fake News Using N-gram Analysis and Semantic Similarity. [online] p.44. Available at: http://dspace.library.uvic.ca:8080/bitstream/handle/1828/8796/Ahmed_Hadeer_Masc_2017.pdf?sequence

4. Oxford Dictionaries | English. (2018). *fake | Definition of fake in English by Oxford Dictionaries*. [online] Available at: https://en.oxforddictionaries.com/definition/fake [Accessed 23 Jul. 2018].

5. Stroud, F. (2018). *What Is Fake News? Webopedia Definition*. [online] Webopedia.com. Available at: https://www.webopedia.com/TERM/F/fake-news.html [Accessed 23 Jul. 2018].

6. K. Tee, "Govt tables anti-fake news Bill for first reading |Malay Mail", *Malaymail.com*, 2018. [Online]. Available: https://www.malaymail.com/s/1607407/govt-tables-anti-fake-news-bill-for-first-reading. [Accessed: 11- Dec- 2018].

7. K. Shu, A. Sliva, S. Wang, J. Tang and H. Liu, *Fake News Detection on Social Media: A Data Mining Perspective*. Arizona State: Michigan State University, 2018.

8. M. Chang, "4 Stages of the Machine Learning (ML) Modeling Cycle", *4 Stages of the Machine Learning (ML) Modeling Cycle*, 2018. [Online]. Available: https://www.linkedin.com/pulse/4-stages-machine-learning-ml-modeling-cycle-maurice-chang. [Accessed: 09- Sep- 2018]