Optimized Convolutional Neural Network For Fake News Detection

StanlyJayaPrakash[1], NandhaKumar[2], Ravivarman[3], Santhosh[4], Sivasankar[5]
[1]Professor, Dept of Computer Science
[2, 3, 4, 5]Dept of Computer Science
[1, 2, 3, 4, 5]Mahendra Institute Of Technology, Tamil Nadu, Namakkal DT – 637 503

Abstract- *The advent of the World Wide Web and the rapid adoption of social media platforms (such as Facebook and Twitter) paved the way for information dissemination that has never been witnessed in the human history before. With the current usage of social media platforms, consumers are creating and sharing more information than ever before, some of which are misleading with no relevance to reality. Automated classification of a text article as misinformation or disinformation is a challenging task. Even an expert in a particular domain has to explore multiple aspects before giving a verdict on the truthfulness of an article. In this work, we propose to use machine learning ensemble approach for automated classification of news articles. Our study explores different textual properties that can be used to distinguish fake contents from real. By using those properties, we train a combination of different machine learning algorithms using various logistic regression methods and evaluate their performance on 4 real world datasets. Experimental evaluation confirms the superior performance of our proposed ensemble learner approach in comparison to individual learners. Fake news detection attracts many researchers' attention due to the negative impacts on the society. Most existing fake news detection approaches mainly focus on semantic analysis of news' contents. We propose a novel fake news Logistic regression technique.*

I. INTRODUCTION

The recent proliferation of social media has significantly changed the way in which people acquire information. Nowadays, there are increasingly more people consuming news through social media, which can provide timely and comprehensive multimedia information on the events taking place all over the world. Compared with traditional text news, the news with images and videos can provide a better storytelling and attract more attention from readers. Unfortunately, this is also taken advantage by fake news which usually contain misrepresented or even forged images, to mislead the readers and get rapid dissemination. The dissemination of fake news may cause large-scale negative effects, and sometimes can affect or even manipulate important public events. For example, within the final three months of the 2016 U.S. presidential election, the fake news generated to favor either of the two nominees was believed by many people and was shared by more than 37 million times on Facebook. Therefore, it is in great need of an automatic detector to mitigate the serious negative effects caused by the fake news.Such proliferation of sharing articles online that do not conform to facts has led to many problems not just limited to politics but covering various other domains such as sports, health, and also science. One such area affected by fake news is the financial markets, where a rumor can have disastrous consequences and may bring the market to a halt.Our ability to take a decision relies mostly on the type of information we consume; our world view is shaped on the basis of information we digest. There is increasing evidence that consumers have reacted absurdly to news that later proved to be fake. One recent case is the spread of novel corona virus, where fake reports spread over the Internet about the origin, nature, and behavior of the virus. The situation worsened as more people read about the fake contents online. Identifying such news online is a daunting task.

II. EXISTING SYSTEM

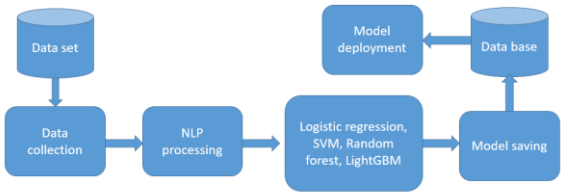
In the existing method, fake news detection multi-task learning (FDML) model has been presented which is based on the following observations.Some certain topics have higher percentages of fake news. Some certain news authors have higher intentions to publish fake news.FDML model investigates the impact of topic labels for the fake news and introduce contextual information of news at the same time to boost the detection performance on the short fake news.

III. PROPSED SYSTEM

Through this project, we have dived deep into the fake news by analyzing the details of social media. In this project, we try to extend the possible end-users by including users and admin.There are a number of computational techniques that can be used to mark certain articles as fake on the basis of their textual content. Majority of these techniques use fact checking websites such as "PolitiFact" and "Snopes.") Here are a number of repositories maintained by researchers

that contain lists of websites that are identified as ambiguous and fake. However, the problem with these resources is that human expertise is required to identify articles/websites as fake. More importantly, the fact checking websites contain articles from particular domains such as politics and are not generalized to identify fake news articles from multiple domains such as entertainment, sports, and technology. In the proposed method, we proposed the Fake news detection technique with logistic regression architecture. For the preprocessing, the Natural Language processing (NLP) processes are perform to extract the information from the text data. After that, Logistic regression is take places in order to perform the classification operations. The proposed architecture is deployed in web based application by Django framework. The World Wide Web contains data in diverse formats such as documents, videos, and audios. News published online in an unstructured format (such as news, articles, videos, and audios) is relatively difficult to detect and classify as this strictly requires human expertise. However, computational techniques such as natural language processing (NLP) can be used to detect anomalies that separate a text article that is deceptive in nature from articles that are based on facts. Other techniques involve the analysis of propagation of fake news in contrast with real news. More specifically, the approach analyzes how a fake news article propagates differently on a network relative to a true article. The response that an article gets can be differentiated at a theoretical level to classify the article as real or fake. A more hybrid approach can also be used to analyze the social response of an article along with exploring the textual features to examine whether an article is deceptive in nature or not.

IV. SYSTEM ARCHITECTURE



ADVANTAGE

- Higher accuracy of the model.
- The performance of the model is high.
- The proposed model has ability to work with different kind of dataset.

V. MODULE DESCRIPTION

DATA COLLECTION

The datasets we used in this study are open source and freely available online. The data includes both fake and truthful news articles from multiple domains. The truthful news articles published contain true description of real world events, while the fake news websites contain claims that are not aligned with facts.The conformity of claims from the politics domain for many of those articles can be manually. We have used three different datasets in this study, a brief description of which is provided as follows. The first dataset is available at Kaggle (hereafter referred to as DS2) which contains a total of 20,386 articles used for training and 5,126 articles used for testing. The dataset is built from multiple sources on the Internet. The articles are not limited to a single domain such as politics as they include both fake and true articles from various other domains. The second dataset is also available at Kaggle (hereafter referred to as DS3); it includes a total of 3,352 articles, both fake and true. The true articles are extracted from trusted online sources such as CNN, Reuters, the New York Times, and various others, while the fake news articles are extracted from untrusted news websites. The domains it covered include sports, entertainment, and politics. A combined dataset is the collection of articles from the three datasets (hereafter referred to as DS4). As the articles vary in nature in each dataset, the fourth dataset is created to evaluate the performance of algorithms on datasets which cover a wide array of domains in a single dataset.

NLP PREPROCESSING

The NLP toolkit is used to extract the text from the data set.In this step, we preprocessed the data that we scraped in order to make it ready for data analysis. It involved tasks such as data cleaning and data integration. Since the description field was paragraph based, we removed all the stop words, punctuations to extract the keywords for NLP tasks. We integrated data because the job posts were scraped from two different websites and there were many duplicate listings as well. Hence we applied entity resolution technique, Jaccard similarity to identify the similar pairs after integration.

MODEL SELECTION

In the proposed method, the Logistic regression is used to build the model.Based on the hyper parameter, the different kind of model is developed and which are shortlisted based on the accuracy of the model. The based model is saved to deploy the model.

MODEL DEVELOPMENT

Flask is a frame work of the python to build the website. The flask is used to deploy our model with user

friendly. Heroku is a cloud platform which is used to host our website for the online users.

VI. WORKING

Machine learning uses two types of techniques: supervised learning, which trains a model on known input and output data so that it can predict future outputs, and unsupervised learning, which finds hidden patterns or intrinsic structures in input data.

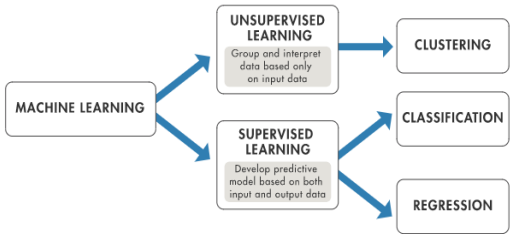


Fig. Machine Learning Techniques Include Both Unsupervised And Supervised Learning.

Types of machine learning

There are various ways to classify machine learning problems. Here, we discuss the most obvious ones.

1) Supervised learning: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs. The training process continues until the model achieves the desired level of accuracy on the training data. Some real-life examples are:

Image Classification: You train with images/labels. Then in the future you give a new image expecting that the computer will recognize the new object.

Market Prediction/Regression: You train the computer with historical market data and ask the computer to predict the new price in the future.

2) Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. It is used for clustering population in different groups. Unsupervised learning can be a goal in itself (discovering hidden patterns in data).

Clustering: You ask the computer to separate similar data into clusters, this is essential in research and science.

High Dimension Visualization: Use the computer to help us visualize high dimension data.

Generative Models: After a model captures the probability distribution of your input data, it will be able to generate more data. This can be very useful to make your classifier more robust.

LOGISTIC REGRESSION

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique.
- It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable.
- Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic regression becomes a classification technique only when a decision threshold is brought into the picture. The setting of the threshold value is a very important aspect of Logistic regression and is dependent on the classification problem itself.

The decision for the value of the threshold value is majorly affected by the values of precision and recall. Ideally, we want both precision and recall to be 1, but this seldom is the case. In the case of a Precision-Recall trade-off, we use the following arguments to decide upon the threshold:-

1.**Low Precision/High Recall**: In applications where we want to reduce the number of false negatives without necessarily reducing the number of false positives, we choose a decision value that has a low value of Precision or a high value of Recall. For example, in a cancer diagnosis application, we do not want any affected patient to be classified as not affected without giving much heed to if the patient is being wrongfully diagnosed with cancer. This is because the absence of cancer can be detected by further medical diseases but the presence of the disease cannot be detected in an already rejected candidate.

2.**High Precision/Low Recall**: In applications where we want to reduce the number of false positives without necessarily reducing the number of false negatives, we choose a decision value that has a high value of Precision or a low value of Recall. For example, if we are classifying customers whether they will react positively or negatively to a personalized advertisement.

Based on the number of categories, Logistic regression can be classified as:

1)**Binomial**: target variable can have only 2 possible types: "0" or "1" which may represent "win" vs "loss", "pass" vs "fail", "dead" vs "alive", etc.

2)**Multinomial**: target variable can have 3 or more possible types which are not ordered (i.e. types have no quantitative significance) like "disease A" vs "disease B" vs "disease C".

3)**Ordinal**: it deals with target variables with ordered categories. For example, a test score can be categorized as: "very poor", "poor", "good", "and very good". Here, each category can be given a score like 0, 1, 2, and 3.

Training data assumptions for logistic regression

Training data that satisfies the below assumptions is usually a good fit for logistic regression.

- The predicted outcome is strictly binary or dichotomous. (This applies to binary logistic regression).
- The factors, or the independent variables, that influence the outcome are independent of each other. In other words there is little or no multicollinearity among the independent variables.
- The independent variables can be linearly related to the log odds.
- Fairly large sample sizes.

Steps in Logistic Regression:

To implement the Logistic Regression using Python, we will use the same steps as we have done in previous topics of Regression. Below are the steps:

- Data Pre-processing step
- Fitting Logistic Regression to the Training set
- Predicting the test result
- Test accuracy of the result (Creation of Confusion matrix)
- Visualizing the test set result

1. Data Pre-processing step: In this step, we will pre-process/prepare the data so that we can use it in our code efficiently. It will be the same as we have done in Data pre-processing

2. Fitting Logistic Regression to the Training set: We have well prepared our dataset, and now we will train the dataset using the training set. For providing training or fitting the model to the training set.

3. Predicting the Test Result: Our model is well trained on the training set, so we will now predict the result by using test set data.

4. Test Accuracy of the result:
Now we will create the confusion matrix here to check the accuracy of the classification. To create it, we need to import the confusion_matrix function of the sklearn library.

5. **Visualizing the training set result:** Finally, we will visualize the training set result. To visualize the result, we will use Listed Colormap class of matplotlib library.

VII. DEPLOYMENT

Component diagrams are used to describe the components and deployment diagrams shows how they are deployed in hardware. UML is mainly designed to

focus on the software artifacts of a system. However, these two diagrams are special diagrams used to focus on software and hardware components.

GOALS

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

VIII. CONCLUSION

The proposed method Logistic regression based system presented higher accuracy while compared with other existing approaches. The accuracy of the model is 97.5% on the test data. The proposed model has ability to perform the classification operation on different datasets. A novel fake news detection multi-task learning (FDML) model based on the following observations: 1) some certain topics have higher percentages of fake news; and 2) some certain news authors have higher intentions to publish fake news. FDML model investigates the impact of topic labels for the fake news and introduce contextual information of news at the same time to boost the detection performance on the short fake news. Specifically, the FDML model consists of representation learning and multi-task learning parts to train the fake news detection task and the news topic classification task, simultaneously. As far as we know, this is the first fake news detection work that integrates the above two tasks. The experiment results show that the FDML model outperforms state-of-the-art methods on real-world fake news dataset.

IX. DECLARATION

Conflicts of Interest

No conflict of interest in this manuscript

Authors Contributions

StanlyJayaPrakash, Sivasankar was involved in data collection, data analysis & manuscript writing.Author, NandhaKumar, Ravivarman, Santhosh wasinvolved in conceptualization, data validation, and critical review of manuscripts.

Acknowledgment

The authors would like to express their gratitude towards Mahendra Institute of Technology (Formerly known as Mahendra University) for successfully carrying out this work.

REFERENCES

[1] N. Shavit, “Data on facebook’s fake news problem,” Jumpshot, 2016.
[2] S. Vosoughi, D. Roy, and S. Aral, “The spread of true and false news online,” Science, vol. 359, no. 6380, pp. 1146–1151, 2018.
[3] A. Verma, “Google engineer beaten to death, 3 critical in indian lynch mob attack fueled by ’kidnapping’ rumor,” thomson reuters, 2018.
[4] D. DiFranzo and M. J. K. Gloria, “Filter bubbles and fake news,” ACM Crossroads, vol. 23, no. 3, pp. 32–35, 2017. [Online]. Available: https://doi.org/10.1145/3055153
[5] K. Shu, D. Mahudeswaran, and H. Liu, “Fakenewstracker: a tool for fake news collection, detection, and visualization,” Computational & Mathematical Organization Theory, vol. 25, no. 1, pp. 60–71, 2019.
[6] S. Ghosh and C. Shah, “Toward automatic fake news classification,” in 52nd Hawaii International Conference on System Sciences, HICSS 2019, Grand Wailea, Maui, Hawaii, USA, January 8-11, 2019, 2019, pp. 1–10.
[7] V. Perez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, “Au- ´ tomatic detection of fake news,” in Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018, 2018, pp. 3391–3401.
[8] C. Castillo, M. Mendoza, and B. Poblete, “Information credibility on twitter,” in Proceedings of the 20th International Conference on World Wide Web, 2011, pp. 675–684.