Ensemble Model for Profiling Fake News Spreaders on Twitter

Notebook for PAN at CLEF 2020

[1]H.L Shashirekha, [2]Anusha M.D, [3]Nitin S Prakash
[1,2]Department of Computer Science, Mangalore University, Mangalore, Karnataka, India
[3]Department of Information and Communication Technology, Manipal Institute of Technlogy, Manipal, Karnataka, India
[1]hlsrekha@gmail.com, [2]anugowda251@gmail.com, [3]nitinsp123@gmail.com

Abstract. Capturing millions of users round the globe, social media has become one of the most popular means of communication and information sharing. Yet, using social media such as Facebook, WhatsApp, Twitter, and Instagram is like a double-edged sword. On one side is the easy accessible, low cost and rapid dissemination of information simultaneously to the large audience while on the other side is the dangerous exposure of fake news. The large spread of fake news has the potential for creating extremely negative impacts on individuals and society. Hence, detecting fake news spreader is the need of the day. To tackle this problem, this paper presents an Ensemble model for profiling fake news spreaders on Twitter which allows using multiple heterogeneous classifiers and combining the results of these classifiers just like a team decision rather than an individual decision. The first step in building the model is to extract three feature sets namely, Unigram TFIDF, N-gram TF and Doc2vec from the PAN 2020 fake news spreader dataset and applying feature reduction on Unigram TFIDF and N-gram TF. Two Linear SVC classifiers and a Logistic Regression classifier are built using three feature sets and are ensembled using majority voting. The results obtained on PAN 2020 test set are encouraging.

1 Introduction

In today's world, social media have become one of the main platforms for exchange of information. Social media such as Facebook, WhatsApp, Twitter, and Instagram, have the advantage of reaching wider audience simultaneously at a much faster rate which has made them popular especially among younger generation. It is very easy and economic to generate news online and disseminate it very fast through social media. However, the information available on social media may be false or faked purposefully with various intentions such as to deceive the readers, insult an individual or a group creating embarrassing situations, impact negatively on individuals as well as on the society and so on. So identifying fake news spreader is an important task in this modern era and is gaining popularity day by day.

Fake news will usually be incomplete, unstructured and noisy [1]. Hence, effective methods are required to extract useful features from the fake news to differentiate

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). CLEF 2020, 22-25 September 2020,Thessaloniki, Greece

whether the news is fake or not. As the news has to be identified as fake or not, it can be modeled as a binary Text Classification (TC) problem with only two labels 'yes'/'1' representing fake news and 'no'/'0' representing genuine news. Text classification is the process of assigning one of the predefined tags/categories/labels to a new or unlabeled text automatically according to its content. It is one of the fundamental tasks in Natural Language Processing (NLP) with broad applications such as fake news detection, sentiment analysis, topic labeling, spam detection, and intent detection. Researchers have explored several algorithms for TC that gives good performance. However, some algorithms perform better on some datasets, but the same algorithms may not give even an average performance on some other datasets. So, it is very difficult to claim or prove that a particular classifier is good for all datasets. As a result, instead of using a single classifier it is better to use a group of classifiers and take a collective decision just like a decision of a team rather than an individual. This approach called as Ensemble approach overcomes the weakness of one classifier by the strength of other classifier and gives better performance than an individual classifier.

In this paper, we propose an Ensemble approach for profiling fake news spreaders on Twitter. The rest of the paper is organized as follows. Section 2 highlights the related work and the proposed Ensemble approach for profiling fake news spreaders on Twitter is described in Section 3. Experiments and results are described in Section 4 and the paper finally concludes in Section 5.

2 Related Works

Researchers have developed several algorithms for profiling fake news spreaders. Few important and relevant ones are highlighted below:

A Machine Learning (ML) model for Spam Identification proposed by Kyumin et. al. [2] uses social honey pots in MySpace and Twitter as fake websites that act as traps to spammers. Since the collected spam data contained signals strongly correlated with observable profile features such as contents, friend information or posting patterns, these features are used to feed various ML classifiers and obtained results in range of 95% to 98% F1 score. A framework for collecting, preprocessing, annotating and analyzing bots in Twitter proposed by Zafar et. al. [3] extracts several features such as the number of likes, retweets, user replies, mentions, URLs, follower friend ratio and it has been found that humans share novel contents compared to bots which rely more on retweets or sharing URL. A bot and gender profiling study submitted to the PAN 2019 by Hamed et. al. [4] have exploited the TF-IDF features to train a model for human-bot detection and then an ensemble voting classifier has been built using three base models on character-level and word-level representations of the documents. The proposed model has achieved accuracies of 83% and 73% for English and Spanish respectively on the test set provided by PAN2019.

An author profiling model submitted to PAN 2017 by Basile et. al. [5] have constructed Ensemble model with voting classifier and dynamic and ad-hoc grid search approach. The model was trained with character-level and word-level N-gram representations of the documents and results aggregated using majority voting

achieved 85% accuracy on the test set. A ML model for gender classification of Twitter using n-grams has been proposed by Burger et. al. [6]. In addition to using user's full name as most informative field with respect to gender, they have used word unigrams, bigrams and character 1 to 5 grams as features and obtained an accuracy of 89.1%. Daneshvar et. al. [7] have proposed a ML model that uses latent semantic analysis on the TF-IDF matrix with a linear SVM classifier for the dataset provided by PAN 2018. Their model reported accuracies of 82.21%, 82% and 80.9% on English, Spanish, and Arabic datasets respectively and was the best-performing model. John et. al. [8] have addressed the bot identification problem from an emotional perspective and evaluated their model on India Election Dataset collected from July 15th 2013 to March 24th 2014. They have also performed a comparative study of classifiers including and excluding sentiment features, and sentiment model achieved an accuracy of 95% which is better than the other models. Tetreault et. al. [9] proposed a ML model using Logistic Regression classifiers to build an Ensemble classifier for Native Language Identification (NLI). A wide range of features such as word and character n-gram, POS, function words, writing quality markers and spelling errors are used to build the classifier. The proposed model achieved an accuracy of 90.1% on International Corpus of Learner English (ICLE) corpus. Filter based feature selection methods for the prediction of risks in Hepatitis disease is proposed by Pinar et. al. [10]. Their work consists of a preprocessing module including standardizing non-standard language expressions such as replacing slang words, contractions, abbreviations, and emoticons by their corresponding normalized language expressions. The proposed feature selection methods helped in improving learning accuracy and learning time reduction and obtained an accuracy of 84% on the dataset collected from UCI machine learning data repository that contains 19 fields with one class attribute.

Emotions play a key role in deceiving the reader. Based on emotions, Bilal et. al. [11] proposed a LSTM neural network model that is emotionally-infused to detect false news. The proposed model obtained accuracies of 80.72% and 64.82% for news articles and Twitter based on a large set of emotions and the results illustrate that false information has different emotional patterns in each of its types. A model based on Low Dimensionality Representation (LDR) proposed by Francisco et. al. [12] for language variety identification was applied on the age and gender identification task in PAN Lab at CLEF and the obtained results are quite competitive with the best model in author profiling task in PAN.

3 Methodology

Our proposed Ensemble approach for profiling fake news spreaders in Twitter is explained in this section. Architecture for building the Ensemble model using the fake news spreader training set provided by PAN 2020 [13] is shown in Figure 1. The proposed approach consists of the following modules:

3.1 Data Preparation

This step combines all the tweets of each user as one text document and assigns the corresponding label as per the data given by PAN 2020 [13].

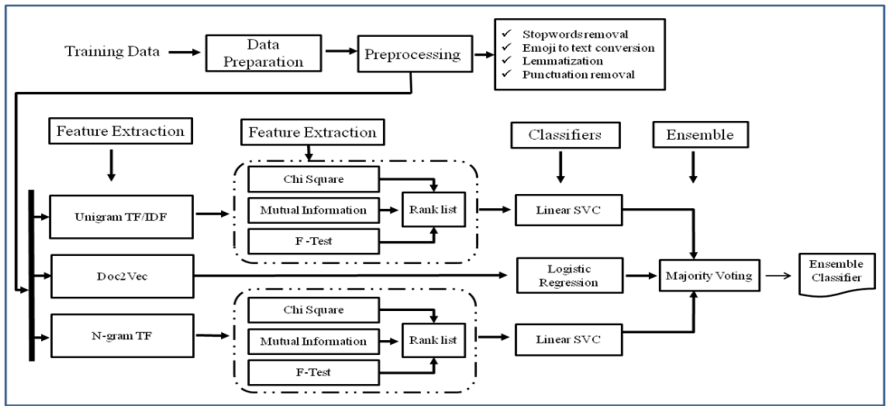

Figure 1. Architecture for building the Ensemble model

3.2 Preprocessing

The first step in preprocessing is demojification which is the process of converting emojis to text. Emojis are visual representation of emotions, object or symbol which can be inserted individually or together to create a string. As these visual representations convey certain meaning they are converted to words conveying the corresponding meaning and used similar to other words in the text. Following this all punctuation symbols and numeric information are removed as they do not contribute to TC. The remaining text is tokenized into words and to reduce number of words, unwanted words including all stopwords, uninformative words, words with length less than three are removed. Further, Stemming and Lemmatization are applied to reduce the words into their root forms. Text data by default is high dimensional. On the average preprocessing reduces the dimension of text by 20-25%. The remaining words are given as input to the feature extraction step to extract features.

3.2 Feature Extraction

Feature extraction module is responsible for extracting the distinguishing features from the given text collection that completely describe the original dataset. The following feature selection algorithms are used in our model:

- **Unigram Term Frequency/Inverse Document Frequency (TF/IDF)** is a weighting scheme often used in Information Retrieval and Text Mining (TM) which represents the relative importance of the word in the document and the entire corpus. It is a common technique used in any text analysis application including TC. A single word is considered as a feature and hence the name Unigram TF/IDF.
- **N_gram** is a sequence of N words called as word N_grams or N characters called as character N_grams which are highly used in TM and NLP. This simple idea has been found to be effective in many applications [14]. N_gram model predicts the occurrence of the word based on the occurrence of its previous N-1 words. A TF of bi-grams, tri-grams and four-grams are used in our model.

- **Doc2Vec** is an NLP tool for representing text documents as a vector and is a generalization of the word2vec method. It is an unsupervised algorithm to generate vectors for variable length pieces of text such as sentences, paragraph, or documents, which has the advantage of capturing semantics in the input texts. The vectors generated by doc2vec can be used for tasks like finding similarity of text at various levels like sentences, paragraph or documents. A vector of size 300 is used in our model.

3.3 Feature Reduction

Text data is by default high dimensional as they consists of words and words are the features for any text analysis applications. Complexity of any algorithm increases with the increase in the number of features. Hence, reducing the number of features yet representing the original dataset becomes very important to reduce the complexity of the algorithms. Feature reduction plays an important role in reducing the number of features by eliminating redundant and irrelevant features thereby improving the performance of the algorithm [15].

Three feature selection algorithms namely Chi Square test, Mutual Information and F-test are used for feature reduction. All these three feature selection algorithms are based on filter approach. In filter approach, a statistical measure is used to assign a score to each feature and the features are ranked by the score. Only top k ranked features will be selected for further processing. Brief descriptions of the feature selection algorithms are given below:

- **Chi-square test** helps to solve the problem of feature selection by testing the relationship between the features. It is based on the difference between the observed and the expected values for each category. Higher Chi-square value shows that feature more dependent on the response variable and can be selected for the model training.
- **Mutual Information** is a powerful feature selection technique that can be used to measure the relationship between the variables including non-linear relationship. It is invariant under transformations in the feature space that are invertible and differentiable, e.g. translations, rotations, and any transformation preserving the order of the original elements of the feature vectors. The main advantage of this method is rapidity off execution.
- **F-test** is a class of statistical tests that calculates the ratio between variances values, such as the variance from two different samples or the explained and unexplained variance by a statistical test, like ANOVA. The scikit-learn machine library provides an implementation of the ANOVA f-test in the f_classif( ) function. It is most often used to compare statistical models that have been fitted to a dataset, in order to identify the model that best fits the population from which the data were sampled.

Feature reduction is applied only for Unigram TF/IDF and N-gram models as the number of features in these two models are very high. The features are ranked using the above methods and top 5000 features are considered in each method. Then, a reduced feature set consisting of the disjoint union of all these features is used for building the classifiers.

3.2 Classifier Construction

Classifiers are the supervised learning models constructed using the training set. Two classifiers namely, Linear SVC and Logistic regression available at Scikit-learn, Machine Learning in Python[1] are used in our work. Brief descriptions of these two classifiers are given below:

- **Linear SVC** is a Linear Support Vector Classifier (SVC) similar to Support Vector Machine classifier with linear kernel available at Scikit-learn. It accepts the training data and returns a "best fit" hyperplane that categorizes the given data. As Linear SVC is efficient, flexible and works fast, it is highly recommended for TC.
- **Logistic regression** is a very effective classification method used on text data. It is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist.

After preprocessing the training set provided by PAN 2020 [13], Unigram TFIDF, N-gram TF and Doc2vec features are extracted. Further, feature reduction is applied to Unigram TFIDF and N-gram TF models only as these two models have more features. All the three features sets are scaled using MaxAbsScaler**.** This scaler automatically scales the data to a [-1, 1] range based on the absolute maximum. It is meant for data that is already centered at zero or sparse data. It does not shift/center the data, and thus does not destroy any sparsity. These scaled feature sets are used to build the learning models.

Two Linear SVC models are constructed using scaled Unigram TFIDF and N-gram TF features and Logistic Regression model is constructed using scaled Doc2vec features of vector size 300.

3.5 Ensemble

There is no classifier which always gives a good result for all the datasets. Hence, instead of using a single classifier, an ensemble of three classifiers mentioned above (Linear SVC using Unigram TFIDF, Linear SVC using N-gram TF features and Logistic Regression using Doc2vec) are used with majority voting [16].

The ensemble classifier accepts the test data from the PAN 2020 organizers and assigns either a ‘1’ or ‘0’ representing ‘fake’ or ‘not fake’ labels respectively as shown in Figure 2. The test data is preprocessed and the three sets of features mentioned above, namely, Unigram TFIDF, N-gram TF and Doc2vec are extracted. Only those features presented in the reduced feature set are input to the ensemble model to assign the suitable label ignoring the rest.

4 Experiment results

The code is implemented in Python using Scikit-learn, Machine Learning in Python. Uncompressed dataset provided by PAN 2020 consists of two folders, one for English and another for Spanish language. Each folder in turn contains an xml file

[1] scikit-learn.org

author (twitter user) with 100 Tweets and name of the XML file corresponds to the unique author id. Author list and ground truth is given in truth.txt file. Details of dataset are given in Table 1. Two separate models are created for two languages English and Spanish.

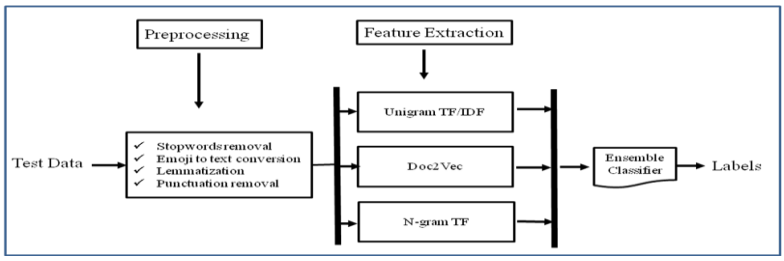

Figure 2. Ensemble model for predicting the class labels of test data

Table. 1 Details of datasets provided by PAN

Language	No. of Authors	No. of tweets / Author	No of fake tweets	No. of genuine tweets
English	100	300	150	150
Spanish	100	300	150	150

The proposed model is evaluated through PAN submission system called TIRA Integrated Research Architecture, which is a modularized platform for shared tasks [17] and the performance of the model is reported by task evaluators. Our model has obtained an accuracy of 73.50% for English language text and 67.50% for Spanish language text and an overall accuracy of 70.50% with a runtime of 00:01:52 (hh:mm:ss).

5 Conclusion

With the increasing popularity of social media, more and more people consume news from social media instead of traditional news media. However, social media has been used to spread fake news which has strong negative impact on individual and the society**.** In this paper, an Ensemble model is built for profiling fake news spreaders on Twitter task in PAN 2020. Ensemble approach uses majority voting of the three (two Linear SVC classifiers and a Logistic Regression) classifiers built using Unigram TF/IDF, N_gram TF and Doc2Vec feature sets. The proposed model obtained 73.50% and 67.50% accuracies on English and Spanish languages respectively.

References

1. Jiliang Tang, Yi Chang, and Huan Liu. “Mining social media with social theories a survey”. ACM SIGKDD Explorations Newsletter, 15(2), pp. 20–29, 2014.
2. Kyumin Lee, Brian David Eoff, and James Caverlee. “Seven months with the devils: A long-term study of content polluters on twitter”. In Fifth International AAAI Conference on Weblogs and Social Media, 2011.
3. Zafar Gilani, Reza Farahbakhsh, Gareth Tyson, Liang Wang, and Jon Crowcroft. “Of bots and humans (on twitter)”. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, pp. 349–354, ACM, 2017.
4. Hamed Babaei Giglou, Mostafa Rahgouy, Taher Rahgooy, Mohammad Karami Sheykhlan and Erfan Mohammadzadeh. “Author profiling: Bot and gender prediction using a multi-aspect ensemble approach”. Notebook for PAN at CLEF 2019”. In Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. 2019.
5. Basile A., Dwyer G., Medvedeva M., Rawee J., Haagsma H. and Nissim M. “N-GrAM: New Groningen Author-profiling Model” - Notebook for PAN at CLEF 2017. In Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Dublin, Ireland, 2017.
6. Burger J.D., Henderson J., Kim G. and Zarrella G. “Discriminating gender on twitter”. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1301–1309. EMNLP’11, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011.
7. Daneshvar S. and Inkpen D. “Gender Identification in Twitter using N-grams and LSA” - Notebook for PAN at CLEF 2018. In Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.), CLEF 2018 Evaluation Labs and Workshop – Working Notes Papers, 10-14 September, Avignon, France, 2018.
8. John P Dickerson, Vadim Kagan and V S Subrahmanian. “Using sentiment to detect bots on twitter: Are humans more opinionated than bots”? In Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 620–627, 2014.
9. Tetreault J., Blanchard D., Cahill A. and Chodorow M. “Native Tongues, Lost and Found: Resources and Empirical Evaluations in Native Language Identification”. In Proceedings of COLING 2012, pp. 2585–2602, 2012.
10. Pinar Yildirim, “Filter Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease”, International Journal of Machine Learning and Computing, Vol. 5(4), 2015.
11. Ghanem Bilal, Paolo Rosso, and Francisco Rangel. “An emotional analysis of false information in social media and news articles”. ACM Transactions on Internet Technology (TOIT) Vol 20, no. 2, pp.1-18, 2020.
12. Rangel Francisco, Marc Franco-Salvador, and Paolo Rosso. “A low dimensionality representation for language variety identification”. In International Conference on Intelligent Text Processing and Computational Linguistics, pp. 156-169, Springer, Cham, 2016.

13. Rangel F., Giachanou A., Ghanem B., and Rosso P. “Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter”. In: L. Cappellato, C. Eickhoff, N. Ferro, and A. Névéol (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings.CEUR-WS.org, 2020.
14. Keselj V., Peng F., Cercone N. and Thomas C. “N-gram-based Author Profiles for Authorship Attribution”. In Proc. of the Conference Pacific Association for Computational Linguistics, 2003.
15. Gao W., Kannan S., Oh S., and Viswanath P. “Estimating mutual information for discrete-continuous mixtures”. arXiv preprint arXiv:1709.06212
16. Tuwe Lofstrom. “On Effectively Creating Ensembles of Classifiers, Studies on Creation Strategies, Diversity and Predicting with Confidence”, Stockholm University, Ph.D. thesis, 2015.
17. Potthast Martin, Tim Gollub, Matti Wiegmann, and Benno Stein. “TIRA integrated research architecture”. In Information Retrieval Evaluation in a Changing World, pp. 123-160. Springer, Cham, 2019.