The Process Of Multi-Class Fake News Dataset Generation

Sajjad Rezaei
Department of Computer Engineering
Ferdowsi University
Mashhad, Iran
sajjad.rezaei@mail.um.ac.ir

Mohsen Kahani
Department of Computer Engineering
Ferdowsi University
Mashhad, Iran
kahani@um.ac.ir

Behshid Behkamal
Department of Computer Engineering
Ferdowsi University
Mashhad, Iran
behkamal@um.ac.ir

***Abstract*— Nowadays, news plays a significant role in everyday life. Due to the increasing usage of social media and the dissemination of news by people who have access to social media, there is a problem that the validation of the news may be questioned, and people may publish fake news for their benefit. Automatic fake news detection is a complex issue. It is necessary to have up-to-date and reliable data to build an efficient model for detection. However, there are very few such datasets available for researchers. In this paper, we proposed a new fake news dataset extracted from three famous and reliable fact-checking websites. Because of the different labels used in each site, an algorithm was developed to integrated these 37 labels into five unified labels. Some experiments were conducted to show the usability and validity of the dataset.**

Keywords*— *fake news; fake news detection; social media; reliable dataset;

I. INTRODUCTION

The impact of fake news on society increases because fake news has far-reaching effects on politics, economics, and public trust. For example, fake news that "Barack Obama wounded in an explosion" spread and stock market value dropped for $ 130 million [1]. Social media can help publish rapid, low-cost, and widespread dissemination of news to the whole world. On the other hand, it has some disadvantages, such as publishing news without validation and using it for personal or group benefits. These cases show that detecting fake news can help people make the right decision. Fake news can be defined as intentionally and verifiably false news published by a news outlet [2].

Identify fake news is a subset of the deception detection field .Deception detection is not a new problem in natural language processing. An Early study focuses on detecting deception goes over opinions in sentiment analysis, using a crowdsourcing method to create training data for the positive class, and then combine with truthful opinions from TripAdvisor [3]. Crowdsourcing fact-checking is a critical approach to create datasets, but with the massive increase of fake news, it became no longer appropriate to use crowdsourcing to build datasets because it was a costly and time-consuming method.

The further we go, the process of writing fake news evolves and gets closer to the real news style, making it more difficult to identify fake news. One way to get more efficient from automatic fake news detection models is to use up-to-date and reliable data to familiarize the model with the new features of fake news and provide higher performance.

For this reason, we tried to create a complete and up-to-date dataset that is highly reliable (because the information has been extracted from reputable sites in the field of fake news), So that researchers can build better models with this reliable dataset.

The rest of paper is organized as follows: Section II reviews the related works, significantly articles that introduced the dataset and used it to detect fake news with different models. In section III, we explain how to extract news from fact-checking websites, reviewing the number and type of labels available on each site, and finally, the completed database is provided. Section IV shows the experiments performed on the prepared dataset to show its validity and utility.

II. RELATED WORK

In this section, we look at the related works that have helped fake news detection using popular datasets. Most of the focus is on the datasets and brief explanations of how they were constructed.

In [4], the focus is on Information credibility on Twitter. It used the Twitter events detected by Twitter Monitor for 2-months and then asked evaluators to assign labels, like *chat* or *news,* for the crawled data. Credibility assessment on this work had four labels: "almost certainly true", "likely to be false", "almost certainly false" and "I can't decide.

In [5] uses Tri-Relationship Exploiting for fake news detection. In this work, they use BuzzFeed[1] dataset that only contains the title and content for each news. It creates two fake news datasets, which both contain news content and social engagement information. The reliable labels are

[1] https://github.com/BuzzFeedNews/2016-10-facebook-factcheck/blob/master/data

978-1-6654-0208-8/21/$31.00 ©2021 IEEE

gathered from journalist experts from BuzzFeed and the well-recognized fact-checking website PolitiFact.

Paper [6] combines content and social signals to detect fake news. They propose a machine learning fake news detection method and implement it using a Facebook Messenger chatbot and rate it with a real-world application, gaining a fake news detection accuracy of 81.7%. FakenewsNet dataset contains posts and likes from Facebook pages belonging in two classes: scientific news sources vs. conspiracy news sources. The resulting dataset comprises 15500 posts, coming from 32 pages with more than 2300000 likes users. 58 posts are hoaxes and 42 are non-hoaxes.

On the other hand, unsupervised fake news detection investigated detect fake news in an unsupervised manner. They treated truths of news and user's credibility as latent random variables and exploited user's engagements on social media to identify their opinions towards the authenticity of the news. They use LIAR [7] and BuzzFeed dataset to build their model [8].

Table I shows published datasets on fake news detection in recent works [9]. It consists of news counts (size), number of classes, modality, source, and data category. The last row belongs to our dataset.

Most of the available datasets have binary labels, and they are collected only from one source; due to a large number of fake news in the world, some news may remain by one source and not have been reviewed. Due to the rapid progress of fake news, another drawback is that the dates of datasets are old.

Our dataset has up-to-date data fake news data, which has been extracted from the top three sites in this field. It is also present in five classes, which is one of the advantages of this dataset because it can be used to identify news that combines real news with fake news to mislead automated models.

III. Dataset development

To develop the dataset, we crawled three expert-based fact-checking websites to collect fake news data from 25 September 1995 up to 16 January 2021 (for Snopes from 25 Sep 1995 to 16 Jan 2021, for Politifact 9 Sep 2009 to 13 Jan 2021, and for Truthorfiction from 17 Mar 2015 to 11 Jan 2021). Those websites are Politifact.com, Snopes.com, and Truthorfiction.com. The data collected is reliable because those three websites are among the most reliable websites in the fake news domain [2]. Any news published will be reviewed by the experts in the field related to each news, and then the news is tagged.

Unfortunately, each site has its own set of labels. We collected 35598 news with 37 different labels from those websites. Due to the existence of different labels in each site, after checking the definitions of each label accurately and investigating the labels of other datasets, we decided to choose labels that are common to all three websites. On the other hand, because we can mention Politifact.com as the most reliable online fact-checking website, five prominent labels have been selected from this site and applied to the labels of other websites. So, five final tags were selected (True, False, Half-True, Mostly-true, and Mostly-false).

Because most published datasets have binary labels, with the many advances made in writing fake news, this news has become very similar to real news.

Table I. Comparison of various fake news detection datasets [9].

Dataset Name	Source	Category	Type	News Counts	Number of Classes
CREDBANK	Twitter	Diverse	Text	60000000	5
FakeNewsCorpus	Opensources.co	Diverse	Text	9400000	10
Fakeddit	Reddit	Diverse	Text-Image	1063106	2,3,6
NELA-GT-2018	194news outlets	Diverse	Text	713000	8
FAKENEWSNET	Twitter	Political	Text	602659	2
FEVER	Wikipedia	Diverse	Text	185455	3
image-verification-corpus	Twitter	Diverse	Text-Image	17806	2
some-like-it-hoax	Facebook	Scientific/Conspiracy	Text	15500	2
LIAR	Politifact	political	Text	12836	6

Today, fake news is published by combining several paragraphs of true and among them, some false (fake) paragraphs, that binary label (True and Fake) can no longer be effective enough to identify fake news. For this purpose, in addition to false and true labels, we considered three other labels to present fake news better. Tables II-III-IV, the labels with bold font are mapped to five selected labels and the rest of them removed.

A. Politifact

Fact-checking in PolitiFact is done by journalists whose main goals are fairness, reporting, and clear writing, independence, and transparency. We collected 17386 news with nine labels: False, Mostly False, Pants-Fire, True, Half True, Full-Flop, Mostly True, Half-Flip, and No-Flip. Table II shows the statistics of the number of collected news with different labels. (You can see the definition of different labels on Politifact website)

In table II, those labels with bold font are used in the final dataset.

TABLE II. STATISTICS OF COLLECTED NEWS FROM POLITIFACT.COM

Labels	Count
False	**4054**
Mosty-False	**2841**
Pants-Fire	2144
True	**2122**
Half-True	**3074**
Full-Flop	154
Mostly-True	**2913**
Half-Flip	60
No-Flip	24

B. Snopes

To find the news that has been verified, should refer to the fact-checks section. We collected 15921 news with 18 labels which are listed in table III, separately. (You can see the definition of different labels on this Snopes website)

We put news with False labels in the database without any changes, but we mapped the news with Mixture labels to half-true and then put it in the dataset.

TABLE III. STATISTICS OF COLLECTED NEWS FROM SNOPES.COM

Labels	Count
Satire	1
False	**5210**
Mostly False	**622**
Mixture	**1201**
True	**1558**
Miscaptioned	436
Correct Attribution	158
Outdated	123
No Rating	19
Unproven	643
Mostly True	**310**
Scam	80
Misattributed	94
Labeled Satire	320
Legend	168
-Blank-	4967
Research In Progress	2
Lost Legend	9

C. Truthorfiction

Truthorfiction is a neutral website where Internet users can receive information about fake news. Truthorfiction is designed to be of value to the ordinary user of the Internet who wants to ensure that an email, post, or story contains information, not misinformation. Our crawler extracted 2291 news from Truthorfiction with ten different labels. Labels included Decontextualized, Unknown, True, Not True, Mixed, This News Has No Tags, -Blank-, To Compute Gains and Losses, Misattributed and Rich Buhler & Staff. see the number of each label in Table IV.

Based on the definitions for labels on this site, it was decided that there would be only True, Mixed, and Not True labels in the final database. Mixed label mapped to Half-true label and Not True mapped to False.

TABLE IV. STATISTICS OF COLLECTED NEWS FROM TRUTHORFICTION.COM

Labels	Count
Decontextualized	246
Unknown	186
True	**472**
Not True	**562**
Mixed	**27**
This News Has No Tags	759
-Blank-	27
To Compute Gains And Losses	5
Misattributed	3

Fig. 1 and Fig. 2, Shows the number of news collected from each fact-checking website and the number of news labels in the dataset, respectively. In Fig. 3, the Title and Content of news word clouds are shown.

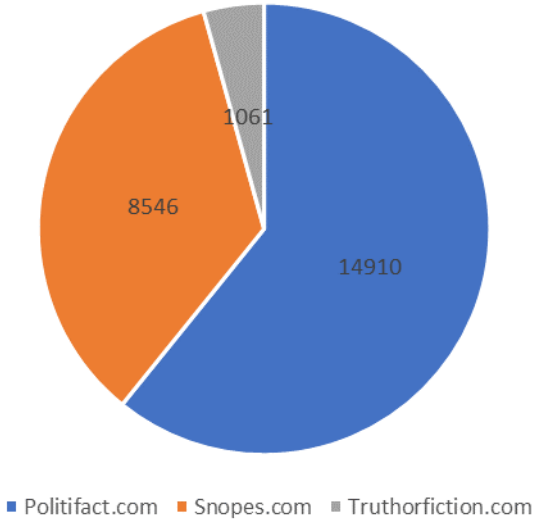


Fig. 1. Number of news collected from each website

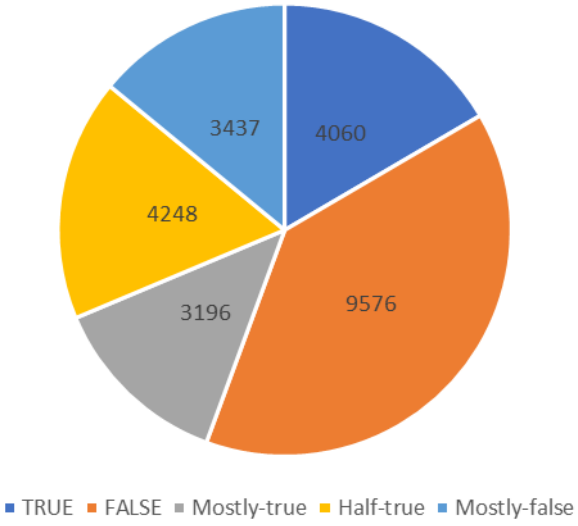
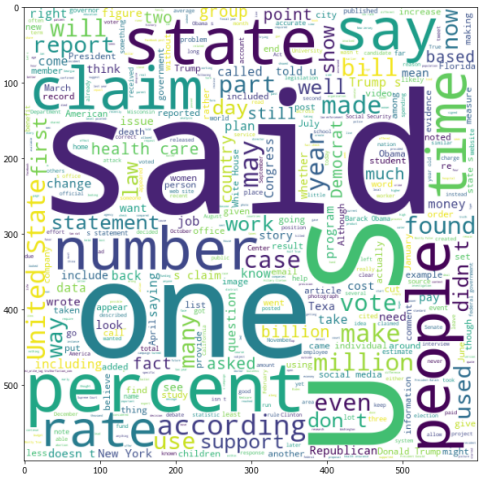
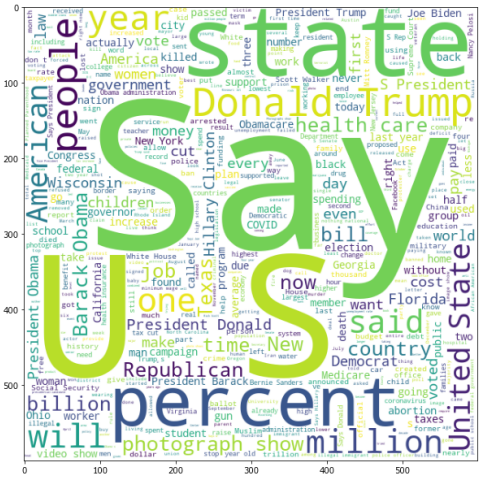


Fig. 2. Number of news label



(a) Content of news



(b) Title of news

Fig. 3. Word cloud generated from title and content of news

IV. Baselines and Results

A. *Pre-processing*

Different sites have different structures. After collecting data from them, pre-processing is required to achieve data with the same format. The news published date format changed from the string to the correct date format. Remove all single characters, substituting multiple spaces with single space, converting to Lowercase, Remove Stop words, and Lemmatization.

B. *Feature extraction*

This research use Term Frequency-Inverse Document Frequency (TF-IDF) with use 2-gram, 3-gram, and 4-gram sequences of words.

C. Data splitting

For building train and test data we used train_test_split from the Sklearn library. Train_size = 0.8 and test_size = 0.2, respectively.

D. Hyperparameter tuning

This step uses GridSearchCV from the Sklearn library, which exhaustively searches over specified parameter values for an estimator. We found the best available parameters for each classification algorithm used and set the best parameters found for each of the algorithms.

1) XGBoost

- Booster = gblinear
- Objective = multi:softmax
- Eval_metric = merror
- Num_class = 5
- Use_label_encoder = False

2) DecisionTree

- Criterion = gini
- Splitter = best
- Max_depth = 10
- Min_samples_leaf = 1
- Min_samples_split = 2

3) RandomForest (RF)

- N_estimators = 100
- Max_depth = 25
- Min_sample_leaf = 1
- Min_sample_split = 2

4) Support Vector Machine (SVM)

- Kernel = rbf
- C = 10
- Gamma = 0.1

E. Machine Learning Algorithms

For classification, we train XGBoost algorithms, Support Vector Machine (SVM), Random Forest (RF), and DecisionTree. The results show in table V. We can see that the best accuracy score of 85% is achieved by XGBoost algorithms. Then the DecisionTree accuracy is closest to XGBoost with 82%. The third place is awarded to RF with 81% accuracy, and SVM is placed in the last position with 79% accuracy.

TABLE V. Accuracy, weighted average Precision (Precision), weighted average Recall (Recall), weighted average F1-score (F1-score) of classification model on the dataset.

Model	Accuracy	Precision	Recall	F1-score
XGBoost	0.85	0.86	0.85	0.85
DecisionTree	0.82	0.83	0.82	0.82
RF	0.81	0.83	0.81	0.80
SVM	0.79	0.81	0.79	0.79

We have 9576 records with False label, 4248 records with Half-true label, 4060 records with True label, 3437 records with Mostly-false label, and 3196 records with Mostly-true label which the classification algorithms train on it. We use stratify method in the Sklearn library for preserving the exact proportions of examples in each class as observed in the original dataset.

LIAR [7] dataset is one of the famous fake news datasets which built-in 2017. According to their report, the accuracy of SVM and LR algorithms is 0.258, 0.257, respectively. This metric in our dataset for SVM algorithms is 0.79.

V. CONCLUSION AND FUTURE WORK

In this paper, we presented a new dataset for fake news detection research. This dataset contains 24517 news with five labels on various topics. We crawled three popular fact-checking websites (Politifact, Snopes, and Truthorfiction). Then, from 37 different labels, according to the definitions of each label and expert opinion, we selected five labels and removed the rest from the dataset. We conducted four experiments with traditional machine learning algorithms such as SVM, RandomForest, XGBoost, and DecisionTree as potential baselines. Among the classification models, XGboost performs the best with 85% accuracy.

Future research can investigate tracking a user's engagement for metadata that can help fake news detection models. Another research that can be defined is extracting new features from fake news.

REFERENCES

[1] K. Rapoza, "Can âĂŸfake newsâĂŹ impact the stock market?(2017)," 2017.

[2] X. Zhou and R. Zafarani, "Fake news: A survey of research, detection methods, and opportunities," *arXiv preprint arXiv:1812.00315,* vol. 2, 2018.

[3] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," *arXiv preprint arXiv:1107.4557,* 2011.

[4] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*, pp. 675-684, 2011.

[5] K. Shu, S. Wang, and H. Liu, "Exploiting tri-relationship for fake news detection," *arXiv preprint arXiv:1712.07709,* vol. 8, 2017.

[6] M. L. Della Vedova, E. Tacchini, S. Moret, G. Ballarin, M. DiPierro, and L. de Alfaro, "Automatic online fake news detection combining content and social signals," in *2018 22nd conference of open innovations association (FRUCT)*, pp. 272-279: IEEE, 2018.

[7] W. Y. Wang, "" liar, liar pants on fire": A new benchmark dataset for fake news detection," *arXiv preprint arXiv:1705.00648,* 2017.

[8] S. Yang, K. Shu, S. Wang, R. Gu, F. Wu, and H. Liu, "Unsupervised fake news detection on social media: A generative approach," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, pp. 5644-5651, 2019.

[9] K. Nakamura, S. Levy, and W. Y. Wang, "r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection," *arXiv preprint arXiv:1911.03854,* 2019.