Improving Classification Efficiency of Fake News using Semi-Supervised Method

Suben Kumer Saha (✉ subenksaha@gmail.com)

American International University Bangladesh <https://orcid.org/0000-0003-0547-9906>

Khandaker Tabin Hasan

American International University Bangladesh

Research Article

Keywords: Fake News, Machine Learning, Semi-supervised, classification

Posted Date: January 4th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1201074/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Improving Classification Efficiency of Fake News using Semi-supervised Method

Suben Kumer Saha^{1*} and Dr Khandaker Tabin Hasan¹

^{1*}Computer Science, American International University-Bangladesh, 408/1, Kuratoli, Dhaka, 1229, Bangladesh.

*Corresponding author(s). E-mail(s): subenksaha@gmail.com;
Contributing authors: tabin@aiub.edu;

Abstract

Online News media which is more accessible, cheaper, and faster to consume, is also of questionable quality as there is less moderation. Anybody with a computing device and internet connection can take part in creating, contributing, and spreading news in online portals. Social media has intensified the problem further. Due to the high volume, velocity, and veracity, online news content is beyond traditional moderation, also known as moderation through human experts. So different machine learning method is being tested and used to spot fake news. One of the main challenges for fake-news classification is getting labeled instances for this high volume of real-time data. In this study, we examined how semi-supervised machine learning can help to decrease the need for labeled instances with an acceptable drop of accuracy. The accuracy difference between the supervised classifier and the semi-supervised classifier is around 0.05 while using only five percent of label instances of the supervised classifier. We tested with logistic regression, SVM, and random forest classifier to prove our hypothesis.

Keywords: Fake News, Machine Learning, Semi-supervised, classification

1 Introduction

As of late, Online news sites and social media platforms played a vital role to disseminate false facts, urban legends, or, more precisely misinformation [kim \(2018\)](#), and it is increasing day by day, mainly because of two reasons: (i) consuming news on social media is affordable and timely compared to television or newspapers; (ii) it is much simpler to additionally comment, share, and talk about the news with other people online [Shu et al \(2017a\)](#). For example, in 2018, Pew Research Center showed that 68 percent of US citizens get at least some sort of news on social media, and it's constantly increasing since 2016. The quality of

news pieces in traditional media, such as newspapers, television is much higher than the news spread on online platforms, because of the absence of the regulatory authority on online media. The news ecosystem is getting imbalanced because of the spreading of fake news on online media, it's affecting both society and individuals. Fake news is also being spread by organizations or individuals for their political gains. For example, it's also reported that fake news had played a vital role in the 2016 US presidential election [Shu et al \(2020\)](#).

2 Problem

The problem of fake news is ubiquitous both online and offline. False or fabricated news is not

a new thing, rather it has been part of history since the printing press has been invented. Before that, rumors have been spread through the human mouth but didn't get any considerable boost. Printing press scaled up the reach of fake news spread. The invention of the internet and social media not only made communication easy, but it also made fake news spread: faster, easier, economically cheap, and scalable. The New York Sun's "Great Moon Hoax" of 1835, which proclaimed there was an alien civilization on the moon, is a notable example of fake news which established The Sun as a profitable and leading newspaper [Soll et al \(2016\)](#). In the past, fake news was used as a weapon to sell more newspapers and tabloids. In the internet era, catchy and fabricated headlines entice people to click on the link and read more which in turn increases advertising revenue for news outlets [Burston et al \(2018\)](#).

Recently, The most notable and widely studied fake news is the 2016 US presidential election. In this time, Social media has grown as an important source of news. At the time of the 2016 presidential election, 14 percent of Americans called social media their "most important" source of news consumption. Analysts found of the known false news stories that appeared in the three months before the election, Trump favored news got shared a total of 30 million times on the contrary Clinton favored news was only 8 million times. Due to confirmation bias, people tend to believe the news that is in favor of their preferred candidate [Allcott and Gentzkow \(2017\)](#).

We have also seen an infodemic at the start of the COVID-19 pandemic which started to spread in late December 2019. Misinformation regarding COVID-19 includes, but is not limited to claiming unproven medical treatment such as gargling with lemon or saltwater and injecting yourself with bleach, false conspiracy theories about the possibility of COVID-19 is bioengineered, or that the 5G cellular network is causing or aggravating symptoms of COVID-19 [van der Linden et al \(2020\)](#).

Such fake news spread much faster due to online media. Online news media publish a large volume of news at any given time (Velocity) with various types of media such as photo, video, audio, etc. (Variety). These 3V of big data make it much harder to deploy traditional moderation

techniques such as expert moderation in the context of online media. Machine learning techniques can eliminate some of the need for expert moderation and make it an efficient process but, to train an effective machine learning model, we need labeled instance which is also expert knowledge dependent and costly to acquire. Without a large dataset from a varied domain, our model won't be good enough to be deployed in the real world. Therefore, we need a more efficient way of getting labeled instances.

3 Related Work

In the literature, a countless attempt was taken to study fake news. Broadly the study of fake news can be categorized as Knowledge-Based, Style based, Propagation based, and Credibility based [Zhou and Zafarani \(2018\)](#).

Knowledge-based study of fake news generally involves analyzing news content by some expert, which is harder to scale, especially in a setting such as a web. Some studies tried to make a knowledge graph which is also sometimes outdated to apply in web context due to the virality of news content on the web. *Style Based* fake news detection is mostly used in deception analysis and detection such as clickbait detection.

Propagation based and *Credibility based* fake news detection are more suitable in social media context as SNS (Social Network Service) provide a lot of propagation signals e.g. like, share, comment, and credibility measure such as user profile. A social network is a great example of a heterogeneous network where news publishers (e.g., Facebook Page), news articles (e.g., Facebook status, Twitter tweet), and news spreaders (i.e., users) create a tri-relationship. For such a network, Shu et al. proposed a hybrid framework with three main components: (I) entity embedding and representation, (II) relation modeling, and (III) semi-supervised learning [Shu et al \(2017b\)](#).

Existing studies on news verification are mostly done using Supervised classification making it a binary classification problem. however, the concerns related to this approach is:

- To find effective features for training classifiers [jin \(2016\)](#).
- No measure of truthfulness or fairness as news content can be partially true or false.

Some early studies, classified social media posts (ie. tweets on Twitter, status on Facebook) individually. Though this approach worked well in many natural language-based approaches, Gupta, Zhao, and Han presented, the propagation-based approach, specifically, they evaluate an event's credibility by propagating credibility values between users, tweets and events [gup \(2012\)](#).

Recent studies such as Kim et al. [kim \(2018\)](#) investigated crowd-powered scalable online stochastic algorithms which dependent on crowd signals such as flagging, which is also prone to adversarial attack.

4 Dataset

The major challenge for supervised machine learning is finding labeled data. Labeling data is costly, time-consuming and one of the major roadblocks to developing a robust model to classify fake news. Another challenge for fake news detection is concept drift. Past experiences don't necessarily predict future occurrences. So we need constant new labels and update on our model to keep it up to date. Minimizing the required label instance trading off some accuracy will build a viable model for the internet-scale problem.

In spite the fact that, there exist several data sets for fake news detection, there are no settled upon benchmark data sets. Some data sets which are publicly available: BuzzFeedNews [Silverman \(2016\)](#), CRED BANK [Shu et al \(2017b\)](#), LIAR [Wang \(2017\)](#), Facebook Hoax [Tacchini et al \(2017\)](#), BS Detector, Buzz Face [san \(2018\)](#), FakeNewsNet [Shu et al \(2020, 2017a,b\)](#).

Among the above dataset, FakeNewsNet is the most complete one. Alongside the textual features, it contains visual and social context, which may further increase classification efficiency. In this study, FakeNewsNet dataset has been used.

5 Methodology

5.1 Definitions

Event: An event is a real-world incident that is public and important enough to be spread among other people who are not just involved in the event. An event can be political, social, or anything. An event sparks discussion and subsequent phases.

News: When an event occurs, some people observe the event. They can discuss the event with other people, some may make a video of it as the mobile camera is fairly available to mass people, a journalist can make content out of the event and publish it in print or online media. All of this can be considered as news in a broad definition. In this study we considered news published in online media only.

Fig. 1 Methodology Overview

5.2 Preprocessing

5.2.1 Case Normalization

Case normalization such as lowercasing or upper casing all words in text help as news corpus is expected to be sparse, big, and has titles and frequent capital words. Normalization makes words consistent (eg. "America" or "america") and solves sparsity issues.

5.2.2 Noise Removal

Noise removal is about removing characters digits and pieces of text that can interfere with text analysis. Data collected from the web has a lot of noise like URLs, special characters eg. hash-tags, ads, non-news content. Pure numbers like phone numbers, a street address has almost no semantic meaning in natural language processing. Noise removal highly depends on the source of text. Structured documents have very little noise which is not the case here. Stemming or lemmatization without noise removal may result in the accurate result as shown in [2](#).

5.2.3 Stop Words Removal

Stop words are a set of commonly used words in a particular language. For example, "a", "the",

Raw word	Lemmatized word
...trouble..	...trouble..
trouble;	trouble;
trouble!	trouble!
<a>trouble	<a>trouble
1.trouble	1.trouble

Fig. 2 Lemmatization without noise removal

Raw word	Cleaned	Lemmatized
...trouble..	trouble	trouble
trouble;	trouble	trouble
trouble!	trouble	trouble
<a>trouble	trouble	trouble
1.trouble	trouble	trouble

Fig. 3 Lemmatization with noise removal

"is", "are", etc. are the most common stop words in English. These words add very little contextual information and may mislead algorithms. Stop word removal can be done using a pre-compiled word list or based on appearance in x% of all documents. The latter method can be more effective in multilingual settings such as the web.

5.2.4 Normalization

The web has a problem of distorted words like "goood" and "gud". Other such example words are "stopwords", "stop-words" and "stop words". Text normalization can transform these words into a canonical (standard) form. Satapathy et. al. showed text normalization strategy for Tweets can improve sentiment classification accuracy by about 4% [sat \(2017\)](#).

5.2.5 Lemmatization

Lemmatization/stemming is a similar approach where the goal is to map a word to its root form like stripping prefix/suffix. Stemming is a more crude way that uses a regular expression to identify a pattern and chop things off. Lemmatization is smarter than stemming as it can map good, better, best to the same word "good" which reduces dimensionality a lot without losing so much meaning. in this study, lemmatization was used and WordNet was used as knowledge for lemmatizer.

5.3 Feature Extraction

Machine learning algorithms can not process text inherently. So text documents need to be mapped

Original Word	Lemmatized word
trouble	trouble
troubled	trouble
goose	goose
geese	goose

Fig. 4 Effects of Lemmatization of WordNet

to a feature space, practically in numeric space. in this study, *term frequency* and *term frequency-inverted document frequency* used for feature extraction. These methods are described below:

5.3.1 Term Frequency

TF counts words that appear in the document to find similarities between documents. Each document is represented by an equal length vector that contains the word counts. Then each vector is normalized so that they sum up to 1. This is done for conversion to probability which is a value between 0 and 1.

Let D denote a corpus, or set of documents. Let d denote a document, $d \in D$; we define a document as a set of words w . Let $n_w(d)$ denote the number of times the word w appears in document d . Hence, the size of document d is $|d| = \sum_{w \in d} n_w(d)$

The normalized TF for word w to document d is defined as follows:

$$TF(w)_d = \frac{n_w(d)}{|d|} \quad (1)$$

5.3.2 Term frequency-inverted document frequency

Historically TF-IDF is used in information retrieval in NLP. It is a statistical metric used to measure how important a term is to a document in a data set. The more a word occurs in a document the more its importance increases for the document but if the word is present in all the documents it may not so important for that particular document. The base intuition of TF-IDF is the rarer the word in the corpus the more semantic meaning it has.

TF-IDF for a word w with respect to document d and corpus D is calculated as follows:

$$TF - IDF(w)_{d,D} = TF(w)_d * IDF(w)_D \quad (2)$$

where IDF is calculated as follows:

$$IDF(w)_D = 1 + \log\left(\frac{\|D\|}{\|d : D\|w\epsilon d\|}\right) \quad (3)$$

5.4 Dimensionality Reduction(PCA)

Principle Component Analysis(PCA) is the most popular dimensionality reduction algorithm. it first identifies the hyperplane that lies closest to data and then projects the data onto it. PCA identifies the axis that accounts for the largest amount of variance in the training set.

For higher-dimensional data, PCA finds as many axes as the dimension of data and chose the one which keeps the highest variance. Under the hood, PCA uses a standard matrix factorization technique called Singular Value Decomposition (SVD) that can decompose the training set matrix X into the matrix multiplication of three matrices $U \sum V^T$, where V contains the unit vectors that define all the principal components that we are looking for, as shown as follows:

$$V = \begin{pmatrix} \parallel & \parallel & \parallel \\ c_1 & c_2 & c_3 \\ \parallel & \parallel & \parallel \end{pmatrix} \quad (4)$$

Fig. 5 Principal components matrix

To project the training set onto the hyperplane and obtain a reduced dataset X_{d-proj} of dimensionality d , compute the matrix multiplication of the training set matrix X by the matrix W_d s, defined as the matrix containing the first d columns of V , as shown in 6.

$$X_{d-proj} = XW_d \quad (5)$$

Fig. 6 Projecting the training set down to d dimensions

5.5 Clustering

As different news media report the same event in the different article yet semantically same, consisting of same actors, places, etc., clustering would put 'news' of same 'event' in the same group and

reduce the required number of labeled data in large.

For clustering, the choice of the algorithm was k -means algorithm. though it's hard to know the number of clusters beforehand, a sufficiently large number of clusters (still very few compared to total examples) though may divide the same event news into multiple groups but still increase classification efficiency. At the same time, k -means is the fastest clustering algorithm suitable for big data such as web news datasets. k means has also online variant as described in [zho \(2005\)](#)[lib \(2016\)](#) which can tackle the velocity and scale of online news.

The way k -means algorithm work is as follows:

1. Specify number of clusters k .
2. Initialize centroids by first shuffling the dataset and then randomly selecting k data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e the assignment of data points to clusters isn't changing.

k -means algorithm uses **Expectation Maximization** strategy for clustering. The E-step is assigning the data points to the closest cluster. The M-step is computing the centroid of each cluster. The objective function is:

$$J = \sum_{i=1}^m \sum_{k=1}^K w_{ik} x_i - \mu_k^2 \quad (6)$$

where $w_{ik} = 1$ for data point x_i if it belongs to cluster k ; otherwise, $w_{ik}=0$. Also, μ_k is the centroid of x_i 's cluster.

5.6 Semi-Supervised Learning

After finding the clusters and cluster center aka. centroids, label those instances. Labeling these instances is highly effective as they are representative of the whole cluster. Though instances at the boundary may be clustered wrongly still it's good enough for the task.

5.7 Classification

Labeled centroids were then used to train the classifier. In this paper, we explored Linear Regression, SVM, and Random Forest classifiers. Results are discussed and analyzed in Results Section 6

6 Results

Results has been demonstrated on FakeNewsNet [Shu et al \(2020\)](#). After the train and test split, supervised classifiers were trained using all the train instances and tested on the test dataset. Then same classifiers were trained only using the centroid instances found from k means clustering and tested again on the same train split dataset. Empirically, k was initialized as 5% of instance count. The results are presented here in table 1:

From the table, it could be observed that, despite labeling 5 out of 100 instances only, the accuracy didn't drop much. We achieved around 5% less accuracy for semi-supervised logistic, SVM, and Random forest classifiers in comparison to a supervised classifier.

Fig. 7 Effect of k value over clustering inertia and classification efficiency

The effect of changing value of k also has been observed through inertia and accuracy. The k vs. Inertia and k vs Accuracy (SVM) graph

Method	label	Supervised Result				Semi supervised result			
		precision	recall	f1	accuracy	precision	recall	f1	accuracy
Logistic	Fake	0.77	0.34	0.47	0.81	0.83	0.08	0.15	0.77
	Real	0.82	0.97	0.88		0.77	0.99	0.87	
SVM	Fake	0.71	0.57	0.63	0.84	0.75	0.22	0.34	0.79
	Real	0.87	0.92	0.89		0.79	0.98	0.87	
Random Forest	Fake	0.72	0.41	0.52	0.81	0.70	0.13	0.22	0.77
	Real	0.83	0.95	0.88		0.77	0.98	0.87	

Table 1 Classification comparison using confusion matrix

has showed above to demonstrate how changing k value affect clustering inertia and classification accuracy.

7 Conclusion

Classifying news needs a huge number of labeled instances to work equally well in varied domains. In this study, we examined how we can decrease the need for labeled instances with a slight increase in inaccuracy. To achieve these, we applied a semi-supervised method to cluster similar news, then label the centroid instance to get the highest quality labels. We took five percent of this centroid instance and trained a classifier based on those which gave only five (5) percent lower accuracy than using all instances. This lower accuracy is justifiable as in the context of fake news classification, labeling some true positive instances as false negative is more acceptable than labeling true negative instances as false positive. In the latter case, true news may fail to reach the users' news feed which is not desired at all. As labeling news is costly, our approach can benefit big tech companies with a high volume of the news stream.

Declarations

- Funding: The authors did not receive support from any organization for the submitted work.
- Conflict of interest/Competing interests: The authors have no relevant financial or non-financial interests to disclose.

References

- (2005) Efficient online spherical k-means clustering, Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005., vol 5, IEEE
- (2012) Evaluating event credibility on twitter, Proceedings of the 2012 SIAM International Conference on Data Mining, SIAM
- (2016) An algorithm for online k-means clustering, 2016 Proceedings of the eighteenth workshop on algorithm engineering and experiments (ALENEX), SIAM
- (2016) News verification by exploiting conflicting social viewpoints in microblogs, Thirtieth AAAI conference on artificial intelligence
- (2017) Phonetic-based microtext normalization for twitter sentiment analysis, 2017 IEEE International Conference on Data Mining Workshops (ICDMW), IEEE
- (2018) Buzzface: A news veracity dataset with facebook user commentary and egos, Twelfth International AAAI Conference on Web and Social Media
- (2018) Leveraging the crowd to detect and reduce the spread of fake news and misinformation, Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining
- Allcott H, Gentzkow M (2017) Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31(2):211–36. <https://doi.org/10.1257/jep.31.2.211>, URL <https://www.aeaweb.org/articles?id=10.1257/jep.31.2.211>
- Burston A, Barrios JC, Gomez D, et al (2018) A brief history of fake news. URL <https://www.cits.ucsb.edu/fake-news/brief-history>
- van der Linden S, Roozenbeek J, Compton J (2020) Inoculating against fake news about covid-19. URL <https://www.frontiersin.org/articles/10.3389/fpsyg.2020.566790/full>
- Shu K, Sliva A, Wang S, et al (2017a) Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter* 19(1):22–36
- Shu K, Wang S, Liu H (2017b) Exploiting tri-relationship for fake news detection. *arXiv preprint arXiv:171207709* 8
- Shu K, Mahudeswaran D, Wang S, et al (2020) Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data* 8(3):171–188
- Silverman C (2016) This analysis shows how viral fake election news stories outperformed real news on facebook. *BuzzFeed news* 16

Soll J, White JB, Sitrin SS, et al (2016) The long and brutal history of fake news. URL <https://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535/>

Tacchini E, Ballarin G, Della Vedova ML, et al (2017) Some like it hoax: Automated fake news detection in social networks. arXiv preprint arXiv:170407506

Wang WY (2017) ” liar, liar pants on fire”: A new benchmark dataset for fake news detection. arXiv preprint arXiv:170500648

Zhou X, Zafarani R (2018) Fake news: A survey of research, detection methods, and opportunities. arXiv preprint arXiv:181200315