International Journal of Language, Literature and Culture (IJLLC)
ISSN: 2582-9823
Vol-3, Issue-1, Jan-Feb 2023
Journal DOI: 10.22161/ijllc
Article CrossRef DOI: https://dx.doi.org/10.22161/ijllc.3.1.3
Peer-Reviewed Journal

Article

Supervised Machine Learning Applications for Detecting Internet Research Agency Misinformation

Thomas Wiese[1], Jessica Wiese

[1]SUNY Empire State College, New York
*Corresponding Author

Article Info

Received: 17 Jan 2023,

Received in revised form: 11 Feb 2023,

Accepted: 19 Feb 2023,

Available online: 28 Feb 2023

Keywords— ***Information Security, Machine Learning, Artificial Intelligence, Misinformation, Fake News, Natural Language Processing, Data, Technology, Analytics, Warfare***

Abstract

Misinformation has shifted political narratives across the globe. Because information shared over social media platforms lack traditional publishers and editors, the public is more susceptible to consuming information that is untrue. During the 2016 U.S. presidential election, the Russian government sponsored information operatives to spread misleading and/or false claims through social media. This study defines a method for automated detection of misinformation on social media using machine learning.

I. INTRODUCTION

Misinformation has shifted the political narrative across the globe. The ability to rapidly share information over social media has in turn enabled vast audiences to consume large quantities of publicly available information for free. However, information shared over social media platforms lacks traditional gatekeepers whose reputations can be damaged if the news they publish is found to be incorrect [1]. This has created an entirely new paradigm wherein information seekers can exist in information bubbles. These niche groups shape the collective reality of millions of people, whether the information shared is true or not [2].

Propaganda has been the tool of choice for authoritarian regimes throughout history [3]. It has been used to control and distort current events to meet their agendas. Contemporary society at large is not immune to these tactics. Rather than hanging posters and using drop leaflets, it is now possible to use social media to spread sympathetic narratives. The public is often unaware that they are consuming propaganda and are thus unable to discern a meaningful difference between traditional and reputable information and state-sponsored propaganda. The modern battlefield is no longer limited to soldiers and submarines; through the internet, warfare has been extended to include the hearts and minds of the citizenry.

II. MATERIALS AND METHODS

Python is the primary programming language used for natural language processing in this study. Recent studies have choosen CNNs, LTSMs, and RNNs as their deep learning frameworks. Keras and Tensorflow libraries as well as conventional sklearn NLP methods are employed during this study [4]. The results and ROC AUC score are measured using a relative scale to determine the maximum detectability of state-sponsored propaganda [5].

All known tweets identified by U.S. intelligence agencies from the 2016 presidential election cycle were analyzed [6]. A machine learning algorithm was written and trained on existing curated rumor detection datasets.

The ROC analysis was developed based on the vectorized tweet content. If AUC was not sufficient, it considered metadata such as the publisher, timestamp, and

network medium. The predicted AUC indicated how well the vectorized text predicted the truthfulness of a given tweet.

More than 1.2 million Russian misinformation tweets were used to test the sentiment analysis algorithm, which was trained on known reputable curated rumor detection datasets [7]. Wang presents the LIAR dataset, cited in previous works reviewed in this dissertation. The LIAR dataset curators collected 10 years of manually labeled short pieces of text from the website PolitiFact.com. PolitiFact provides detailed sourcing and labeling for every specific case determination. The LIAR dataset curation team designed a hybrid, surface-level linguistic neural network model that integrates the metadata described with the associated text [8].

The necessary data were obtained by reviewing the work of Darren L. Linvill and
Patrick L. Warren, which they explained as follows:

Our research employed a data set of 9.03 million tweets released by Twitter on
October 17, 2018 (Gadde & Roth, 2018). These tweets came from 3,661
accounts, which are a subset of the 3,841 accounts given by Twitter to Congress.
A list of these account handles was released on June 18, 2018 by the U.S.
House Intelligence Committee (Permanent Select Committee on Intelligence,
2018). The Twitter release included hashed/de-identified versions of account
handles for accounts with fewer than 5000 followers. We used an alternate
version of the Tweets we collected for an earlier draft of this project to re-identify
most of the accounts.

Linville and Warren distilled the dataset down to 3 million tweets and added features related to the political leanings of tweet authors within the dataset [6]. Their dataset was selected for this research because the additional features were useful for understanding the context of the data. The news organization "Five Thirty Eight" made the large dataset publicly available in a Github repository for researchers and analysts. The researcher's footnote read "add data update from Clemson U. researchers [7]."

The researchers employed a combination of quantitative and qualitative methods to compile this data. To interpret and summarize emergent themes within the body of text, they conducted axial coding [7].

After training the support vector classifier using the LIAR dataset, the testing set was replaced with the IRA tweet data [7, 8]. To allow the classifier to evaluate the accuracy of the prediction, a new column was created within the data. This column contained a true/false category, enumerated as 1 or 5, and all known IRA tweets were set to "1" or false. Internet Research Agency tweets are classified as false based on the following assumptions:

1. The list of IRA twitter handles submitted to the U.S. House Intelligence Committee
in 2018 was correct.
2. Tweets associated with the IRA authors are captured and represented within the
dataset published by Linville and Warren.
3. All tweets published by the IRA are misinformation. Since the IRA mounted an
active and well-documented misinformation campaign, this research assumes that
all information published by them was not credible.

The machine learning models were trained using the LIAR dataset to detect misinformation [8]. The models selected for testing were chosen based on the research of Ries et al., as outlined in Section 3.2. When Ries et al. used supervised learning for fake news detection [8], they calculated the following summary statistics for each model:

Table 1. Results obtained for different classifiers w.r.t AUC and F1 score.

Classifier	AUC	F1
KNN	0.80±0.009	0.75±0.008
NB	0.72±0.009	0.75±0.001
RF	**0.85±0.007**	**0.81±0.008**
SVM	0.79±0.030	0.76±0.019
XGB	**0.86±0.006**	**0.81±0.011**

RF and XGB performed best.

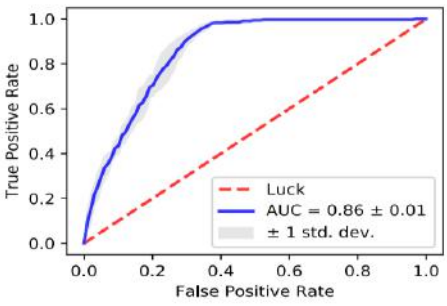

Fig.1. Results Obtained For Different Classifiers By Wang.

"Classifier" indicates the machine learning model used, and F1 represents the f-score comparison of true positive rates determined by the classifier. First developed by Fix and Hodges in 1951, the K Nearest Neighbors (KNN)[9] classifies data into clusters using predefined labels in supervised learning [10]. When making a prediction, KNN determines what cluster a new datapoint falls within based on its nearest neighbors and classifies the point accordingly.

III. RESULTS

A KNN classifier was trained using the LIAR dataset and was tested on the IRA tweet dataset to determine how accurately it could predict if a tweet was true or false. This yielded the following result: 0.8066870643436842

```
[[2180676  522573]
 [      0       0]]
              precision    recall  f1-score   support

           1       1.00      0.81      0.89   2703249
           5       0.00      0.00      0.00         0

    accuracy                           0.81   2703249
   macro avg       0.50      0.40      0.45   2703249
weighted avg       1.00      0.81      0.89   2703249
```

Fig.2. KNN Classifier Results

The result indicates a prediction accuracy of 80.6% and an f1 score of .89 for labeling IRA tweets as false. This accuracy was higher than the accuracy achieved when LIAR was tested on itself.

Naive Bayes classifiers are based on Thomas Bayes' theorem describing the probability of an event based on the conditions that led to that event. In multinomial Bayes, each independent variable is assigned a probability for how likely the dependent variable is to occur and is classified thusly [11]. A multinomial Naive Bayes classifier was trained using LIAR data and tested on IRA tweets.

```
0.9651918857641305
[[2609154   94095]
 [      0       0]]
              precision    recall  f1-score   support

           1       1.00      0.97      0.98   2703249
           5       0.00      0.00      0.00         0

    accuracy                           0.97   2703249
   macro avg       0.50      0.
weighted avg       1.00      0.97      0.98   2703249
```

Figure 3. Bayes Classifier Results.

Fig.4. Bayes Classifier Results.

This calculation predicted that an IRA tweet was false with 96.5% accuracy and an f1 score of 00.98. Random Forest classifiers produce a probabilistic decision tree and predict a target based on the likelihood of a series of binary classifiers. It chooses from multiple layers of nodes until the most likely target is selected based on the data features [12].

A Random Forest classifier was trained using the LIAR dataset and tested on the IRA tweet dataset to determine how accurately it could predict whether a tweet was true or false.

The calculation yielded the result:

0.9536801826246861

[[2578035 125214]

[0 0]]

	precision	recall	f1-score	support
1	1.00	0.95	0.98	2703249
5	0.00	0.00	0.00	0
accuracy			0.95	2703249
macro avg	0.50	0.48	0.49	2703249
weighted avg	1.00	0.95	0.98	2703249

Fig.4. Random Forest Classifier Results.

The calculation predicted that an IRA Tweet was false with 95.3% accuracy and an f1 score of 00.98.

A Support Vector Classifier was trained using the LIAR dataset and tested on the IRA tweet dataset to determine how accurately it could predict whether a tweet was true or false. This is the classifier used in 5.4.1 to validate LIAR's intrinsic accuracy [8]. The calculation yielded the following result:

0.7997113843378838

[[2161819 541430]

[0 0]]

	precision	recall	f1-score	support
1	1.00	0.80	0.89	2703249
5	0.00	0.00	0.00	0
accuracy			0.80	2703249
macro avg	0.50	0.40	0.44	2703249
weighted avg	1.00	0.80	0.89	2703249

Fig.5. Support Vector Classifier Results.

The calculation predicted that an IRA Tweet was false with 79.9% accuracy and an f1 score of 00.89.

An XGBoost was trained using the LIAR dataset and tested on the IRA tweet dataset to determine how accurately it could predict whether a tweet was true or false [8].

The calculation yielded the following result:

0.9973136030014254

[[2695987 7262]

[0 0]]

	precision	recall	f1-score	support
1	1.00	1.00	1.00	2703249
5	0.00	0.00	0.00	0
accuracy			1.00	2703249
macro avg	0.50	0.50	0.50	2703249
weighted avg	1.00	1.00	1.00	2703249

Fig.6. XGBoost Results

The calculation predicted that an IRA Tweet was false at 99.7% accuracy and an f1 score of 1.00. This was the most accurate classifier for this prediction and is consistent with Ries et al.'s finding that XGBoost was the best classifier for fake news detection.

Finally, since true tweets were not represented in the IRA dataset, 200 news tweets from the same period were pulled from 12 news organizations on Twitter and placed in a "full truth" dataset. The goal was to determine if they could be accurately classified as "true" when combined with the IRA tweets. Two thousand four hundred "true" tweets from well-known news organizations were preprocessed and shuffled in with the population of IRA tweets [7].

The news organization tweets in the "full truth" dataset included tweets from the following twitter accounts:

1. Associated Press - "@AP"
2. Thomas Reuters - "@Rueters"
3. PBS News - "@PBS"
4. CBS News - "@CBSNews"
5. ABC News - "@ABC"
6. NPR News - "@NPR"
7. Bloomberg News - "@Bloomberg"
8. USA Today - "@USAToday"
9. BBC News - "@BBC"
10. The Wall Street Journal - "@WSJ"
11. Financial Times - "@FinancialTimes
12. The Hill - "@TheHill"

s.

Fig.7. News Twitter Handles

The full truth dataset was created, preprocessed, and vectorized using the same methodology as the tweets in the IRA dataset. It was then appended and shuffled with the IRA tweets. The LIAR dataset was not used to train the machine learning model during this test. Rather, the train-test-split method was used to split the combined IRA tweet/full truth dataset. In total, 80% were used for training and 20% were used for testing accuracy. The train-test-split method automatically ensures that the dependent variables represented in each group are normally distributed.

An SVM algorithm was applied to the combined IRA tweet/full truth dataset, which yielded the following result:

0.9998509292283432

[[650 86]

[35 810924]]

	precision	recall	f1-score	support
0	0.95	0.88	0.91	736
1	1.00	1.00	1.00	810959
accuracy			1.00	811695
macro avg	0.97	0.94	0.96	811695
weighted avg	1.00	1.00	1.00	811695

Fig.8 SVM Algorithm

In the above table, "0" represents a tweet classified as "true," and "1" represents a tweet classified as "false." Notably, even with a relatively small number of "true" tweets in the dataset, the model was able to correctly distinguish between an IRA tweet and a tweet from a major news outlet with 99.985% accuracy.

The high "true positive rate" and prediction is illustrated against the naive prediction in the logistic ROC AUC of 0.997 and associated ROC curve for the model:

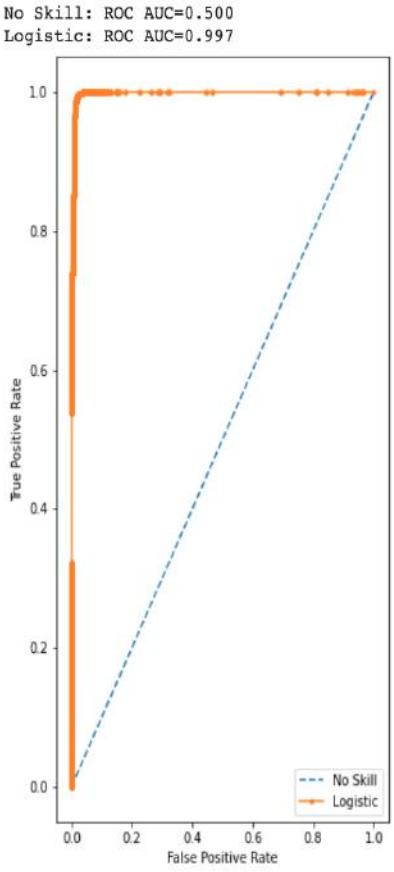

Fig.9 ROC curve.

Therefore, the IRA tweets themselves would be an excellent predictor of future state-sponsored active misinformation campaigns on social media when used to train an SVM learning model.

Discussion

The testing results are summarized in Figure 10 as follows:

Classifier	Accuracy	F1 score
NN	0.806	00.890
B	0.965	00.980
F	0.953	00.980
VM	0.799	00.890
GB	**0.997**	**01.000**

Fig.10. Results By Classifier for IRA Tweet Prediction.

However, since the LIAR dataset is comprised of statements rather than tweets, a limitation of the method could be that the LIAR dataset is biasing the model to classify language style. Language typically used in "tweets" could be assumed false, when statements from speeches that use proper sentence structure, rather than slang or abbreviations, are assumed true [8].

Therefore, an additional test was performed where the IRA tweets themselves were mixed with a curated "full truth" dataset. This was done to determine how accurately the IRA tweets could detect current and potentially future misinformation campaigns on twitter or other social media platforms [7].

The answers to the following research questions were obtained:

RQ 1: Will a machine learning model trained using the LIAR misinformation detection dataset be able to correctly classify an IRA tweet as false with >0.80 accuracy?

H1: The IRA misinformation twitter posts will be detectable to a level of accuracy > 0.90 using the curated dataset

RQ 2: Is the magnitude of sentiment of a false tweet larger than the magnitude of sentiment of a true tweet?

H0: The magnitude of sentiment of a false tweet is less than a true tweet.

RQ3: Will a machine learning model trained using the IRA tweet dataset be able to correctly classify an IRA tweet as false with >0.80 accuracy?

H1: The IRA misinformation twitter posts will be detectable to a level of accuracy > 0.90 using the curated dataset.

RQ 4: Will the machine learning algorithm be able to differentiate between tweets from well-known and reputable news organizations and IRA tweets with >0.80 accuracy?

H1: The machine learning model will be able to differentiate between well-known and reputable news organizations and IRA tweets with >0.80 accuracy.

The SVM model used to predict whether a tweet was "true" or "false" achieved 99.99% accuracy when trained and tested on the combined IRA tweet and full truth datasets. This verified the hypothesis that machine learning could be used to automate the detection of state-sponsored misinformation campaigns on social media under these conditions.

The development of the full truth benchmark dataset and utilization of that dataset to achieve a ROC AUC of 0.997 in detecting Russian Federation-sponsored misinformation on Twitter is the most important finding of this study.

IV. CONCLUSIONS

State-sponsored misinformation campaigns are not new. Indeed, they have been used to influence the opinions of society for at least the past century. The currency of the information age is the ability to influence public opinion. This power ranges from micro-targeted online advertisements influencing purchases, to governments cultivating sympathetic or divisive narratives to meet their own ends.

Misinformation can spread globally, limited only by what platforms and languages are available in specific nations or regions. Concerningly, false information tends to spread more widely, and penetrate more deeply, within

digital networks than truthful information [12]. From Operation INFEKTION in India to COVID-19 misinformation cultivation in 2021, misinformation has been a weapon used to tear apart the social fabric of global rivals [13].

The continually decreasing cost of computing power, combined with the increase in computer memory and capacity, has created an environment where researchers can easily and inexpensively perform complex statistical modeling on extremely large data sets. Complex statistical models are being employed by high memory, high-powered personal computers. They predict outcomes while having access to the whole population of data within a particular domain, rather than within a single sample. This further simplifies the interpretation of results because results are reported in simple descriptive statistics [14].

During this research, a "big data" set containing all tweets by the Internet Research Agency, a Russian state-sponsored misinformation organization, was analyzed to determine whether these tweets could be detected using contemporary machine learning methods [6]. The results indicate that these methods are a viable mechanism for filtering misinformation. This process could be used by fact-checkers to identify and review potentially false tweets for misinformation by exception.

AVAILABILITY OF DATA AND MATERIALS

The data utilized is available in a publicly accessible repository that does not issue DOIs. Publicly available datasets were analyzed in this study. This data can be found here: [https://github.com/tfs4/liar_dataset, https://github.com/fivethirtyeight/russian-troll-tweets/].

AUTHOR CONTRIBUTIONS

TW designed the research, ran experiments, analyzed results, and wrote the manuscript. JW contributed to manuscript preparation.

REFERENCES

[1] Shoemaker, Pamela J., and Tim P. Vos. *Gatekeeping Theory*. New York: Routledge, 2009.

[2] Hunt Allcott, Matthew Gentzkow, and Chuan Yu. Trends in the Diffusion of Misinformation on Social Media. *Research & Politics* 6, no. 2 (2019): 205316801984855, https://doi.org/10.1177/2053168019848554

[3] Barghoorn, Frederick Charles. *Soviet Foreign Propaganda.* Princeton, New Jersey: Princeton University Press, 1964. https://doi.org/10.1515/9781400874590.

[4] Luhn, Hans Peter "A Statistical Approach to Mechanized Encoding and Searching of Literary Information." *IBM Journal of Research and Development* 1, no. 4 (October 1957): 309–17. https://doi.org/10.1147/rd.14.0309.

[5] Manning, Christopher D., and Hinrich Schütze. "Foundations of Statistical Natural Language Processing." NLP Stanford. February 22, 2015. https://nlp.stanford.edu/fsnlp/.

[6] Mueller, Robert. *Report on the Investigation into Russian Interference in the 2016 Presidential Election. Volume II of II.* Washington DC: U.S. Department of Justice, 2019. https://www.justice.gov/storage/report_volume2.pdf.

[7] Linvill, Darren L., and Patrick L. Warren. "Troll Factories: Manufacturing Specialized Disinformation on Twitter." *Political Communication* 37, no. 4 (2020): 447–467. https://doi.org/10.1080/10584609.2020.1718257.

[8] Wang, William Yang. "'Liar, Liar Pants on Fire': A New Benchmark Dataset for Fake News Detection." In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vancouver, July 2017,* 422-426. Vancouver: Association for Computational Linguistics. https://doi.org/10.18653/v1/p17-2067.

[9] Fix, Evelyn, and J. L., Jr., Hodges. "Discriminatory Analysis - Nonparametric Discrimination: Consistency Properties." PsycEXTRA Dataset, 1951. https://doi.org/10.1037/e471672008-001.

[10] Guacho, Gisel Bastidas, Sara Abdali, Neil Shah, and Evangelos E. Papalexakis. "Semi- Supervised Content-Based Detection of Misinformation via Tensor Embeddings." *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Barcelona, 2018: 322–325. https://doi.org/10.1109/asonam.2018.8508241.

[11] Dale, Andrew I. "Thomas Bayes, An Essay towards Solving a Problem in the Doctrine of Chances (1764)." In *Landmark Writings in Western Mathematics 1640-1940*, edited by Ivor Grattan- Guinness, Roger Cooke, Leo Corry, Pierre Crépel and Niccolo Guicciardini, 199–207. Amsterdam, Boston: Elsevier, 2005. https://doi.org/10.1016/b978-044450871-3/50096-6.

[12] Soroush Vosoughi, Deb Roy, and Sinan Aral, "The Spread of True and False News Online," Science 359, no. 6380 (August 2018): pp. 1146-1151, https://doi.org/10.1126/science.aap9559.

[13] Boghardt, Thomas (December 2009). "Soviet Bloc Intelligence and Its AIDS misinformation Campaign (Operation INFEKTION)" (PDF). Studies in Intelligence. 53 (4).

[14] Lukoianova, Tatiana, and Victoria L. Rubin. "Veracity Roadmap: Is Big Data Objective, Truthful and Credible?" *Advances in Classification Research Online* 24, no. 1 (2014): 4-15. https://doi.org/10.7152/acro.v24i1.14671.