A MACHINE LEARNING APPROACH ON COVID-19 TWEETS ANALYSIS

Jinbo Liu[†]
Computer Science Department
Troy University
Troy, AL, US
+1 334 372 9735
jliu183346@troy.edu

Yongchen Zhou[†]
Computer Science Department
Troy University
Troy, AL, US
+1 334 372 5760
yzhou183452@troy.edu

Yuchen Shen[†]
Computer Science Department
Troy University
Troy, AL, US
+1 334 372 8317
yshen@troy.edu

Jiling Zhong
Computer Science Department
Troy University
Troy, AL, US
jzhong@troy.edu

ABSTRACT

More people rely on social media to gather news and other information. [1] However, with it comes an increasing amount of false information. Since the COVID-19 pandemic, social media including Twitter and Facebook, among others, have been filled with misinformation and disinformation about the transmission, prevention, and treatment of the disease. [2] Therefore, it is necessary to detect such misinformation and disinformation and prevent them from spreading. In this work, we compare several traditional machine learning methods when applied to detecting misinformation/disinformation. Further, we experimented with adding new features to the original models. Our experimental results show that these added features significantly improve the performance of these models.

Keywords

COVID-19; fake news detection; machine learning; tweets.

1. INTRODUCTION

With the development of the Internet in recent years, social media has become an effective way for information dissemination [1]. In a BBC survey, of the 18-to-24-year-olds surveyed, 28% cited social media as their main news source, compared with 24% for TV [1]. In just a few seconds, news and information from one place can travel across the ocean and reach every corner of the world. While social media brings convenience to our lives, many problems also arise. In the past, communication methods such as newspapers and radio were controlled by the relevant government agencies (for example, if there were any problems with newspapers or broadcasts, the government agencies could easily hold the relevant media accountable); with social media, however, participants, in many cases, are small group of people, or even individuals, that are not associated with any traditional news organizations. As these ad hoc journalists have various level of trainings in journalism and come from all kinds of social and economic background, there is a higher chance that misinformation and disinformation may occur. To make things worse,

From an application perspective, many websites dedicated to social media are among the most popular—Wikipedia (collective knowledge generation), MySpace and Facebook (social networking), YouTube (social networking and multimedia content sharing), Digg and Delicious (social browsing, news ranking, and bookmarking), Second Life (virtual reality), and Twitter (social networking and microblogging), to name just a few [3]. due to this year's COVID-19 pandemic, many social media are spreading relevant information about the pandemic every day. Reports show 70% of people stated ‘isolation’ and ‘loneliness’ have impacted their mental health, a cause of some COVID-19 enforcement strategies. A 2016 study comparing the USA with high-income countries showed emotional distress such as ‘anxiety’ was found in 26% of participants, nearly 10% more than the UK at 17% [4]. More than 50% of the respondents were afraid from COVID-19 and 64% had stigma towards infected people and their contact during the COVID-19 pandemic [5]. The virus is spreading fast and scary, but fake news on these platforms is just as fast and scary.

2. RELATED WORK

Social media outlets, such as Facebook, Twitter, YouTube, etc., emerged as major information seeking and sharing channels during the pandemic. As a result, the use of social media platforms increased by 20–87% around the world [6]. In recent years, with the large-scale epidemic, there have been more and more reports and news about COVID-19. Inevitably, much related fake news has also been created.

Glazokva et al.'s developed an inheritance-based model in which a vote-to-vote strategy was used, resulting in a weighted accuracy score of 98.69% [7]. In addition, Patawa's team performs classification comparison tests by extracting word frequency features from tweets and using multiple and their learning ?, such as decision trees, gradient boosting, and SV. Their final weighted score was 93.32% [8]. Felber et al. employed several features such as n-grams, sentiment, its use, and readability to cons. They experimented with the device, and in the end, they achieved accuracy score of nearly 95.7% [9].

Additionally, Garcia-Gasulla et al. analyzed the impact of COVID-19 on social mobility, health, and social and economic behavior. They aim to use the results of these analyses to make practical market adaptations and decisions for the private sector. Using a BERT deep learning model, their research found that as COVID-19 broke out and spread, most message recipients reflected fear, sadness, anger, and anticipation [10]. Rosenfeld et al. propose nine more misinformation detections, and with the help of graph kernels, they extract topological information from Twitter cascades. They then experimented with predictive models

[†] These authors contributed equally.

without user identity, language, and temporal information, demonstrating that diffusion patterns play a decisive role in the authenticity of the information. They finally show that through the corresponding aggregation, the collective sharing patterns of crowds can explain whether the information transmitted among them is true or false [11].

Feature selection (FS) is a method that aims to improve classification accuracy and performance by selecting and removing inferior features (redundant or weakly related to the target class). Feature selection prepares datasets for processing by performing dimensionality reduction [13]. This is correct but should it be place here, related work?

3. METHODOLOGY

For a given tweet set $T = \{T_1, T_2, ..., T_n\}$, the goal is to assign one of the labels in the label set $L = \{Real, Fake\}$ to each tweet T_i. To fulfill this task, various tweet features are extracted for T_i and then the feature vector is passed to the data mining component for classification.

3.1 Datasets

To test the effectiveness of our proposed method for detecting COVID-19 fake news, we conduct experiments on a real-world social media dataset [14]. The dataset consists of 8,560 real and fake news tweets about COVID-19. There are 4,480 real news and 4,080 fake news, as shown in figure 1.

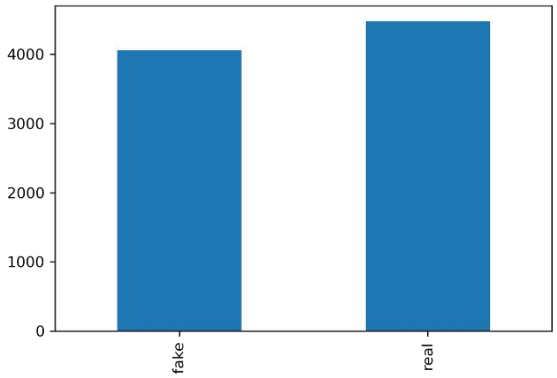

Figure 1. The number of fake news and real news.

Figure 2 and Figure 3 show the top 20 highest frequency words of real news and fake news, respectively.

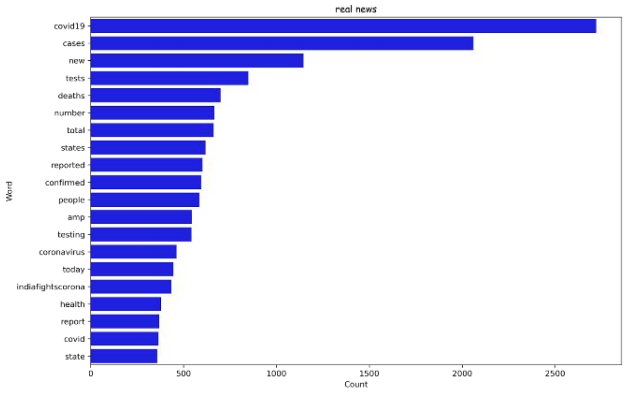

Figure 2. Top20 words of real news.

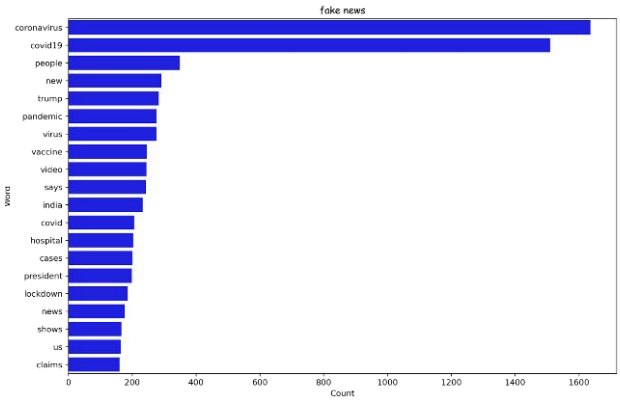

Figure 3. Top20 words of fake news.

We also make two words clouds of real news (see Figure 4) and fake news (see Figure 5).

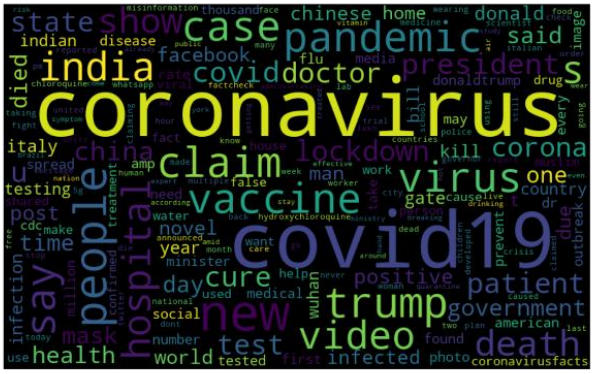

Figure 4. Word cloud of real news.

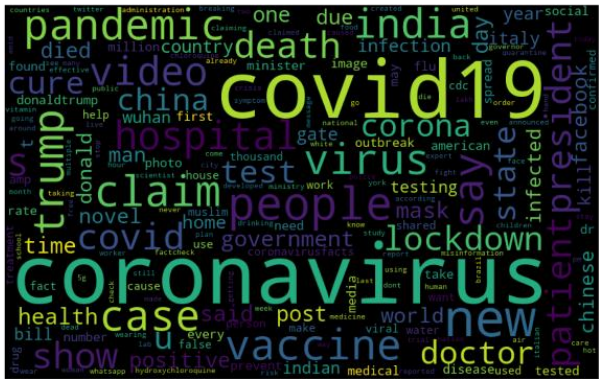

Figure 5. Word cloud of fake news.

3.2 VADER sentiment analysis

VADER (Valence Aware Dictionary and Sentiment Reasoner) is a social media sentiment analysis tool with an integrated lexical database and rule-based framework. It has four scores. The ratios for the percentages of text that fall into each category are the pos, neu, and neg scores. Each word's valence score is added together, modified in accordance with the guidelines, and then normalized to fall between -1 (the most negative) and +1 (the most positive) to get the compound score [15].

† These authors contributed equally.

3.3 Word count feature

We added word count as a new feature. The reason why we add this feature is that there is a big difference, between fake and real news, in the number of words each news contains. On the one hand, real news usually contain some real elements and details, such as emotional words, fact descriptions, and proof of opinions. On the other hand, fake news usually just state opinions or misrepresented facts without convincing proof. In order to show the difference between true and false news more clearly, we make a set of histograms about the number of tweets and word count (see Figure 6 and Figure 7).

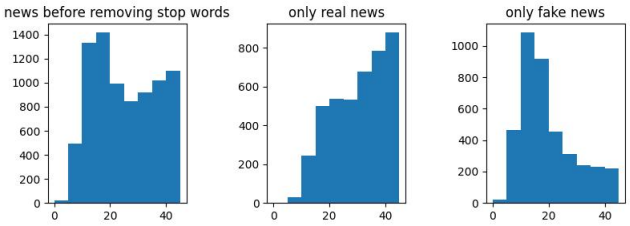

Figure 6. Word count of news before removing stop words.

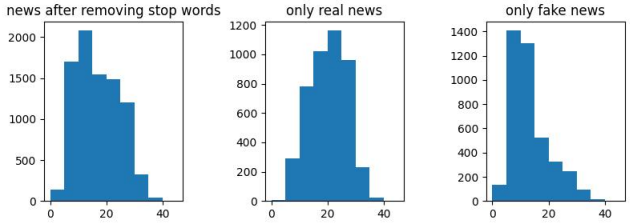

Figure 7. Word count of news after removing stop words.

In both situations (before and after removing stop words), there is a remarkable difference in word count between real and fake news. At the same time, most fake news contain fewer words and the word counts distribute mostly within 10-20, while most real news contains more words and the word counts distribute mostly with 20-40.

3.3.1 Word count threshold

We set a minimum threshold for the word count, for each tweet, if the number of words is less than this threshold, the word count feature will be 0. However, if a tweet's word count is more than this threshold, the feature of it will be 1.

3.3.2 Word count after feature scaling

In this way, the size of this feature of each tweet directly depends on their word counts. The larger their word counts are, the larger this feature will be.

3.3.2.1 Removing outliers

Before adding this feature, we need to remove outliers in the dataset by tweets' word count. In statistics, an outlier is a data point that differs significantly from other observations [16][17]. An outlier can cause serious problems in statistical analyses [18]. The reason why it is necessary to remove outliers is that there may be some tweets that contain hundreds of words. In this situation, these tweets may have a big effect on the whole word count feature. We use the definition of outliers in Statistics to choose our outlier data:

If Q_1 and Q_3 are the lower and upper quartiles, respectively, then one could define an outlier to be any observation outside the range:

$$[Q_1 - 1.5 * (Q_3 - Q_1), Q_1 - 1.5 * (Q_3 - Q_1)][18]$$

3.3.2.2 Feature scaling

Feature scaling is a method to unify self-variables or feature ranges in data [19]. We tried two ways of feature scaling. They are Standardization Scaling and Min-max Scaling. The Standardization Scaling is defined as:

$$x^{'} = \frac{x - \bar{x}}{\sigma}$$

Where x is the original feature vector, $\bar{x} = average(x)$ is the mean of that feature vector, and sigma is its standard deviation [20].

The Min-max Scaling is defined as:

$$x^{'} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Where x is an original value, $\bar{x}$ is the normalized value [20].

4. EXPERIMENT

4.1 Traditional ML methods

We divide the dataset into three parts: training set, validation set, and test set with a proportion of 6:2:2. Firstly, after removing stop words, we make up a dictionary and use it to transform every tweet into vectors. Then, we use TfidfTransformer [21] to make feature selection. However, in each vector, there are too many zeros, and that could decrease accuracy, so we use SciPy [22] to transform the sparse matrix to a lil_matrix. lil_matrix uses two lists to store non-0 elements. Data saves the non-zero elements in each row, and rows save the column where the non-zero elements are located. This format is also suitable for adding elements one by one and can quickly obtain row-related data [23]. Table 1 shows the result of four classic classifiers, which are Decision Tree, Random Forest, SVM Linear, and KNN.

Table 1. Result of classical classifiers.

Estimator	Accuracy	Precision	Recall	F1
Decision Tree	78.69%	69.90%	94.82%	80.47%
Random Forest	88.52%	82.95%	94.69%	88.43%
SVM Linear	93.21%	92.78%	92.54%	92.66%
KNN	87.65%	80.92%	95.95%	87.80%

4.2 Adding VADER sentiment feature

There are four kinds of sentiment features which are positive, negative, neutral, and compound. We test four combinations: adding all of the four features, adding only compound scores, adding positive scores and negative scores, and adding only positive scores (Table 2, Table 3, Table 4 and Table 5 show their results, respectively.).

4.2.1 All of the four scores

Table 2. Result after adding all of the four scores.

Estimator	Accuracy	Precision	Recall	F1
Decision Tree	79.22%	70.72%	94.06%	80.74%
Random Forest	88.41%	82.69%	94.82%	88.34%

† These authors contributed equally.

SVM Linear	93.50%	93.15%	92.79%	92.97%
KNN	87.00%	80.30%	95.32%	87.17%

4.2.2 Only compound scores

Table 3. Result after adding only compound scores.

Estimator	Accuracy	Precision	Recall	F1
Decision Tree	79.51%	75.67%	82.17%	78.79%
Random Forest	89.40%	84.58%	94.31%	89.18%
SVM Linear	93.38%	92.91%	92.79%	92.85%
KNN	87.12%	80.15%	95.95%	87.34%

4.2.3 Positive scores and negative scores

Table 4. Result after adding positive and negative scores.

Estimator	Accuracy	Precision	Recall	F1
Decision Tree	79.57%	71.17%	93.93%	80.98%
Random Forest	89.34%	84.02%	95.07%	89.21%
SVM Linear	93.68%	93.28%	93.05%	93.16%
KNN	88.35%	81.69%	96.46%	88.46%

4.2.4 Only positive scores

Table 5. Result after adding only positive scores.

Estimator	Accuracy	Precision	Recall	F1
Decision Tree	79.04%	70.29%	94.82%	80.73%
Random Forest	88.58%	83.26%	94.31%	88.44%
SVM Linear	93.27%	92.78%	92.67%	92.73%
KNN	87.30%	80.21%	96.33%	87.54%

4.3 Adding word count feature

First of all, we remove outliers in the dataset. After that, we add the feature of word count threshold and the feature of word count after feature scaling, respectively, and compare the results of them to choose the best one.

4.3.1 Word count threshold

We choose integers from 10 to 39 as the threshold. Table 6 and Figure 8 show the results after adding this feature. In Figure 8, the dotted line is the accuracy rate without this feature, and the solid line is the accuracy rate with this feature added.

Table 6. Result after adding word count threshold feature.

	N=10	…	N=14	…	N=39
DecisionTree	78.75%	…	83.02%	…	79.45%
RandomForest	89.17%	…	89.05%	…	88.17%
SVM Linear	94.15%	…	94.09%	…	93.56%
KNN	88.52%	...	88.17%	…	88.35%

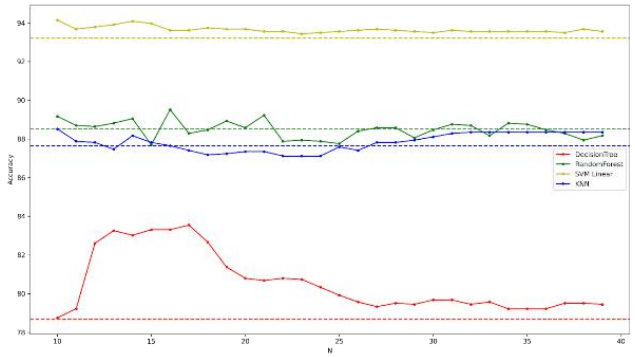

Figure 8. Accuracies of classifiers.

The result of N=14 is one of the highest accuracies. The result is in Table 7.

Table 7. Result when N=14.

Estimator	Accuracy	Precision	Recall	F1
Decision Tree	83.02%	78.05%	88.12%	82.78%
Random Forest	89.05%	83.41%	95.32%	88.97%
SVM Linear	94.09%	92.80%	94.56%	93.68%
KNN	88.17%	81.43%	96.46%	88.31%

4.3.2 Word count after feature scaling

We also add the word count after standard scaling as a feature after adding sentiment features. Table 8 and Table 9 are the results after adding standardization and min-max normalization word count feature, respectively.

Table 8. Result after adding standardization scaled word count feature.

Estimator	Accuracy	Precision	Recall	F1
Decision Tree	84.07%	81.15%	85.46%	83.25%
Random Forest	89.70%	85.06%	94.31%	89.45%
SVM Linear	94.03%	93.22%	93.93%	93.58%
KNN	88.99%	83.84%	94.44%	88.82%

Table 9. Result after adding Min-max normalization scaled word count feature.

Estimator	Accuracy	Precision	Recall	F1
Decision Tree	84.13%	82.02%	84.20%	83.09%
Random	89.70%	85.14%	94.18%	89.44%

[†] These authors contributed equally.

Forest				
SVM Linear	94.03%	93.55%	93.55%	93.55%
KNN	90.11%	84.40%	96.46%	90.03%

We find that the result after adding the Min-max normalization scaled word count feature is better than the result after adding the standardization scaled word count feature or word count threshold. Therefore, we prefer to append the Min-max normalization feature.

5. CONCLUSIONS AND FUTURE WORK

To conclude, first of all, we applied several traditional machine learning methods, including Decision Tree, Random Forest, SVM Linear, and KNN classifiers, respectively, to detect fake news about 8000 tweets and get 78.69%, 88.52%, 93.21%, and 87.65% accuracy respectively. After that, we added another feature, which is the VADER sentimental feature. When adding this feature, we have four choices: all of the four features, only compound feature, positive and negative features, and only positive feature. Adding all of the four features or compound scores can lead to overfitting, resulting in lower accuracy; while adding only positive scores will ignore the negative score, and therefore the accuracy is low. Therefore, we believe that adding positive scores and negative scores is the most effective way to improve. The accuracy of adding positive and negative features is 79.57%, 89.34%, 93.68%, and 88.35%, respectively. In addition, after analyzing the word count distribution of all tweets in the data set, we found that there was a significant difference between the word count of real tweets and fake tweets, and therefore we chose to add the word number feature. There are two kinds of methods for adding word count features. The first one is the word count threshold, and the other one is word count after feature scaling. After comparing the final results, we found that the method of word count after feature scaling is better than the other one, so we chose this method to add new features. Finally, after adding the VADER sentimental feature and word count features, the accuracies of the four classifiers have been significantly improved, which are 84.13%, 89.70%, 94.03%, and 90.11%.

In the future, we plan to add some new useful features to increase accuracy. Besides, we will try other models, such as CNN (Convolutional Neural Network), to analyze fake news.

6. REFERENCES

[1] Jane Wakefield. 2016. Social Media 'outstrips TV' as news source for young people. (June 2016). Retrieved July 19, 2022 from https://www.bbc.com/news/uk-36528256

[2] Elia Gabarron, Sunday Oluwafemi Oyeyemi, and Rolf Wynn. 2021. COVID-19-related misinformation on social media: A systematic review. Bulletin of the World Health Organization 99, 6 (2021). DOI:http://dx.doi.org/10.2471/blt.20.276782

[3] Daniel Zeng, Hsinchun Chen, Robert Lusch, and Shu-Hsing Li. 2010. Social Media Analytics and Intelligence. IEEE Intelligent Systems 25, 6 (2010), 13-16. DOI:https://doi.org/10.1109/mis.2010.

[4] Mental Health America. (2021) The State of Mental Health in America. Available at: https://mhanational.org/issues/state-mental-healthamerica [Accessed 10 June 2021]

[5] Sawsan Abuhammad, Karem H Alzoubi, and Omar Khabour. 2020. Fear of COVID‐19 and stigmatization towards infected people among Jordanian people. International Journal of Clinical Practice 75, 4 (2020). DOI:https://doi.org/10.1111/ijcp.13899

[6] Bruno Kessler Foundation COVID-19 and Fake News in the Social Media (10 March 2020) [(accessed on 8 July 2021)]; Available online: https://www.fbk.eu/en/press-releases/covid-19-and-fake-news-in-the-social-media/

[7] Anna Glazkova, Maksim Glazkov, and Timofey Trifonov. 2021. g2tmn at Constraint@AAAI2021: Exploiting CT-BERT and Ensembling Learning for COVID-19 Fake News Detection. Combating Online Hostile Posts in Regional Languages during Emergency Situation 1402, (April 2021), 116-127. DOI:https://doi.org/10.1007/978-3-030-73696-5_12

[8] Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2021. Fighting an Infodemic: COVID-19 Fake News Dataset. Combating Online Hostile Posts in Regional Languages during Emergency Situation 1402, (April 2021), 21-29. DOI:https://doi.org/10.1007/978-3-030-73696-5_3

[9] Thomas Felber. 2021. Constraint 2021: Machine Learning Models for COVID-19 Fake News Detection Shared Task. DOI:https://doi.org/10.48550/ARXIV.2101.03717

[10] Dario Garcia-Gasulla, Sergio Alvarez Napagao, Irene Li, Hiroshi Maruyama, Hiroki Kanezashi, Raquel P'erez-Arnal, Kunihiko Miyoshi, Euma Ishii, Keita Suzuki, Sayaka Shiba, Mariko Kurokawa, Yuta Kanzawa, Naomi Nakagawa, Masatoshi Hanai, Yixin Li and Tianxiao Li. 2020. Global Data Science Project for COVID-19 Summary Report. DOI:https://doi.org/10.48550/ARXIV.2006.05573

[11] Nir Rosenfeld, Aron Szanto, and David C. Parkes. 2020. A Kernel of Truth: Determining Rumor Veracity on Twitter by Diffusion Pattern Alone. Proceedings of The Web Conference 2020 (April 2020), 1018-1028. DOI:https://doi.org/10.1145/3366423.3380180

[12] Nir Rosenfeld, Aron Szanto, and David C. Parkes. 2020. A Kernel of Truth: Determining Rumor Veracity on Twitter by Diffusion Pattern Alone. Proceedings of The Web Conference 2020 (April 2020), 1018-1028. DOI:https://doi.org/10.1145/3366423.3380180

[13] Girish Chandrashekar and Ferat Sahin. 2014. A survey on feature selection methods. Computers & Electrical Engineering 40, 1 (2014), 16-28. DOI:https://doi.org/10.1016/j.compeleceng.2013.11.024

[14] Parth Patwa. 2021. Fighting an Infodemic: COVID-19 Fake News Dataset. Combating Online Hostile Posts in Regional Languages during Emergency Situation (2021), 21-29. DOI:https://doi.org/10.1007/978-3-030-73696-5_3

[15] C.J. Hutto. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Eighth International AAAI Conference on Weblogs and Social Media 8, 216-225.

[16] Frank E. Grubbs. 1969. Procedures for Detecting Outlying Observations in Samples. Technometrics 11, 1 (1969), 1-21. DOI:https://doi.org/10.1080/00401706.1969.10490657

† These authors contributed equally.

[17] G. S. Maddala. 1992. Introduction to Econometrics. Macmillan College. pp. 89. ISBN 978-0-02-374545-4. An outlier is an observation that is far removed from the rest of the observations.

[18] Wikipedia. 2022. Outlier. (July 2022). Retrieved July 19, 2022 from https://en.wikipedia.org/wiki/Outlier#cite_note-1

[19] Xing Wan. 2019. Influence of feature scaling on convergence of gradient iterative algorithm. Journal of Physics: Conference Series 1213. DOI:https://doi.org/10.1088/1742-6596/1213/3/032021

[20] Wikipedia. 2022. Feature scaling. (April 2022). Retrieved July 19, 2022 from https://en.wikipedia.org/wiki/Feature_scaling

[21] Sklearn. Sklearn.feature_extraction.text.TfidfTransformer. Retrieved July 19, 2022 from https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html

[22] Scipy. Scipy. Retrieved July 19, 2022 from https://scipy.org/

[23] Scipy.sparse.lil_matrix — SciPy v1.8.1 Manual. Retrieved July 26, 2022 from https://docs.scipy.org/doc/scipy/reference/generated/scipy.sparse.lil_matrix.html

† These authors contributed equally.