

IndoPolicyStats: sentiment analyzer for public policy issues

Muhammad Noor Fakhruzzaman¹, Sa'idah Zahrotul Jannah², Sie Wildan Gunawan¹,

Angga Iryanto Pratama¹, Denise Arne Ardanty¹

¹Department of Advanced Technology, Faculty of Advanced Technology and Multidiscipline, Universitas Airlangga, Surabaya, Indonesia

²Department of Mathematics, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia

Article Info

Article history:

Received Nov 11, 2022

Revised Jul 14, 2023

Accepted Aug 2, 2023

Keywords:

Covid-19

Pandemic

Public policy

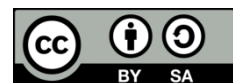
Sentiment analysis

Vaccination

ABSTRACT

The government requires some vaccination for public health. This has led to a debate in recent years, especially during the Covid-19 pandemic. This research aims to analyze the two sentiments of the public regarding the vaccination policy. This would be helpful to ensure the acceptance of the government campaign about vaccination. The data used was text data obtained from Twitter when Indonesia was facing the second wave of the Covid-19 pandemic. The data were pre-processed by removing noise data, case folding, stemming, and tokenizing. Then, the data were classified with random forest, Naïve Bayes, and XGBoost. The results showed that all classifiers exhibit satisfying performance but XGBoost performs slightly better in accuracy value. This method can be deployed to be an automatic sentiment analyzer to help the government understand public feedback about its policies. This would be given by proper pre-processing and enough datasets.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Sa'idah Zahrotul Jannah

Department of Mathematics, Faculty of Science and Technology, Universitas Airlangga

Kampus Merr C, Jl. Dr. Ir. H. Soekarno, Mulyorejo, Surabaya, Indonesia

Email: s.zahrotul.jannah@fst.unair.ac.id

1. INTRODUCTION

Vaccination requirements issued by the government have always created debate among the people throughout decades [1]–[3]. The various responses ranged from positive sentiments about the government's good intent, to negative conspiracy theories [4]. It is important to understand what are the aspects that influence people's sentiments, as they can be used to strengthen the power of the next public health campaign [5], [6].

To ensure the acceptance of the government's campaign, especially public health-related campaigns, we must first understand the socio-cultural background of the people, or the audience [7]–[9]. During the Covid-19 pandemic in Indonesia, there are notably two sides of the spectrum: the conspiracy theorists and the positivists. Hence, it is important to analyze the two sentiments of the public regarding the vaccination policy.

Considering the problems with Indonesians, social media, and how it has such a strong influence on the public's general opinion, this research aims to establish a formal analysis of the public discourse about vaccination policy. Focusing on Indonesia, this study is trying to identify the various sentiment about vaccination policy, specifically to understand what aspects and topics underlie those sentiments. As mentioned in [5]–[7] understanding the theme that drives the various sentiments could give information for better future vaccination campaigns, as similar topics can be used for new campaign material.

This research utilized well-tested baseline models for classifying the sentiments, such as Naive Bayes, XGBoost, and random forest classifiers. Also using term frequency-inverse document frequency (TF-IDF) to measure each word's importance in each tweet and the whole dataset. TF-IDF has been used by

many studies and has been shown to be reliable, especially to represent the knowledge within a corpus. It is also found to be performing well in the smaller dataset, running faster, and using less computing power compared to other feature extraction algorithms [10]–[16].

2. METHOD

Twitter has been an enormous source of data about public responses. The experience of using Twitter made the public feel convenient about sharing their thoughts about anything on the platform, including their reaction to public policy, in this case, the Covid vaccination policy [13], [17]–[20]. Previous studies used social media data, specifically Tweets, to analyze public discourses and sentiments. It has shown reliable results, and the finding could be used as campaign material to better target the audience [5]–[7]. One of the best cases is about analyzing sentiments of human papillomavirus (HPV) vaccination requirement which created ripples of refusal in the US. Using a combination of automated text analysis and qualitative coding, the vaccine promoter can tailor their campaign content to be more relatable to their audience, resulting in better acceptance in future campaigns.

Moreover, studying public sentiment as a whole and comparing it to the key opinion leader's narrative could give us insight into the communication structure in social media. In the traditional sense, targeting the key opinion leaders to follow the government's narrative could easily drive the masses' opinion, as the key opinion leader is often viewed as more trustworthy than the government itself. For example, on the Madura Island of Indonesia, the opinion of their local Kyai (religious leader) is often viewed as divine and must be carried on [21], [22]. Meanwhile, on the internet, the opinion of overly famous YouTubers and social media influencers are also perceived as divine by their followers [23], [24]. Therefore, analyzing the tweets of identified key opinion leaders and their impact on society may also bring insight into the research.

The data used is text data obtained from Twitter containing the word "Covid-19 vaccination". The data was taken from May 15, 2021 to July 12, 2021, when Indonesia was facing the second wave of the Covid-19 pandemic [25]. The 6% of tweets per day were sampled to be analyzed. The tweets were labeled with positive and negative sentiments. The data obtained were preprocessed with several steps, such as removing noise and duplicate, cleaning, case folding, tokenizing, stemming, and feature extraction TF-IDF. The flow of preprocessing to the classification of the data is depicted in Figure 1.

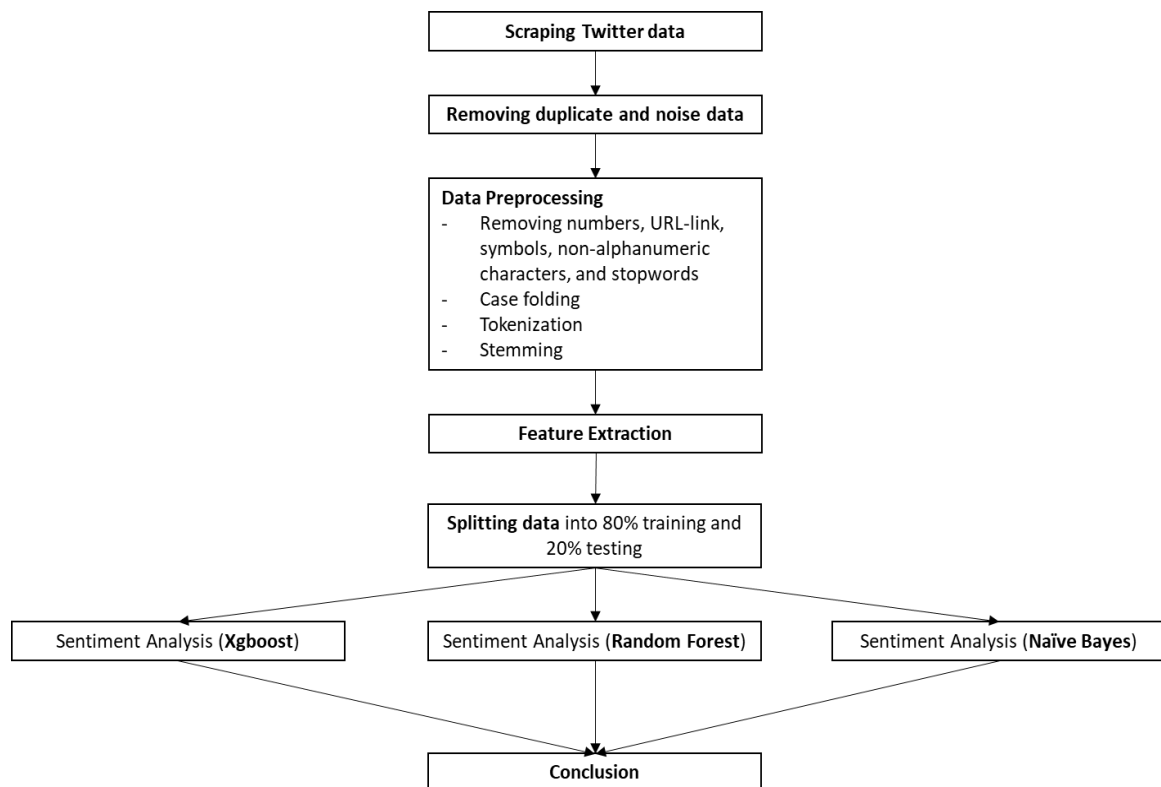


Figure 1. Research flow

Firstly, the data was obtained from Twitter. Then, duplicates were removed from the dataset. The duplicates may have emerged from Twitter's automated bots, and buzzers, which are typically sourced from a news feeder. Thus, it is noise for the dataset. After cleaning and preprocessing, the data used are 873 tweets. Before representing the text into weights, a word cloud for each sentiment group is formed to ensure that the most frequent words are not the query word, which is "Covid-19" and "vaccination". If those words appeared dominant on the word cloud, it needs to be removed for it has not provided any useful information for the classification. Then, the tokens were assigned weights based on their frequency and relevance in the documents [12]. The number generated from that process is the representation of the token in the document and can provide value for the classifier. After the weighting process, each tweet has become a series of weights from each word comprising the tweet.

Moreover, the tokenization process did not only produce unigrams. To better capture the semantic meaning in the whole sentence, the tokenization process also made bigrams and trigrams [26], [27]. These separate forms of tokens were then modeled separately. Afterward, the dataset is split for testing the classifier. Only 80% of the tweets are used for training while the remaining tweets are acting as a control. Finally, each classifier was trained and tested, then each classifier's performance was measured. In the end, the analysis and conclusion are provided.

3. RESULTS AND DISCUSSION

The initial scrape resulted in a total of 58,165 tweets but considering the trends of the tweet on vaccination topics, it was resampled to 6% of each day's matching total tweets. Figure 2 shows the trends of vaccination-related tweets during the scraping session. Specifically, it started to jump on July 4th, 2021, which coincides with the start of the second wave pandemic in Indonesia. People started to aware of the benefits of vaccination and started to talk about it on Twitter.

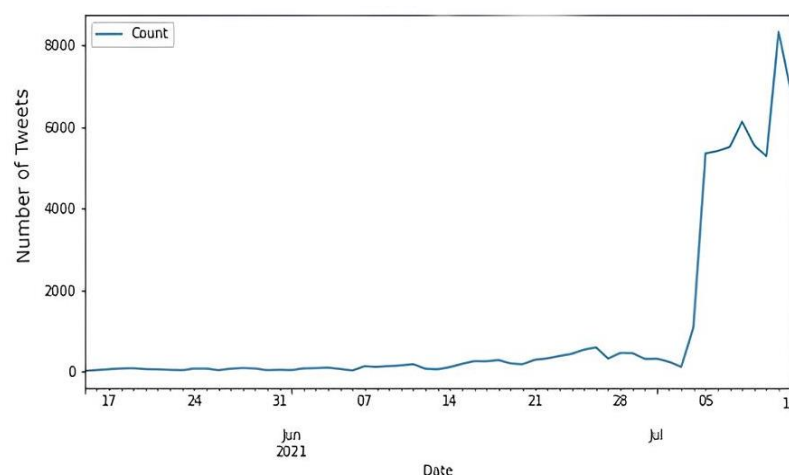


Figure 2. Vaccination-related tweet trend

The second wave in Indonesia started a week after the Eid holiday, which was marked by massive hometown visits of nearly every person in Indonesia. Other than the broken social restriction, a new Covid variant (Delta) was also present during that time. Those massive turn of events incites online debate about the government's vaccination policy, which only imported Chinese vaccines at the time, and its effectiveness compared to other big pharma brands.

Before further preprocessing was done, the tweets were labeled manually first. Most of the tweets have neutral sentiments (45%), which is perfectly normal in any online conversation since people have to be aware to not take sides and lose their online reputation. Furthermore, 39% of the tweets have positive sentiment, and the remainder (16%) expressed negative sentiment.

Figure 3 shows the top 10 words of every tweet in their respective sentiment class, Figure 3(a) positive class and Figure 3(b) negative class. Note that the tweets in each class have been preprocessed, and stopwords, including search query words, were removed. In the positive sentiment class, the word "free", "I hope", and "Thanks God" expressed joy and hope for a better future, which is an optimistic view of the

vaccination policy and showing endorsement for the government's policy. Free vaccination, as promoted by the Indonesian Government, was positively responded to by the people and became the main promotion narrative for the vaccination campaign. Another top word in the positive class, "healthy", shows that people relate vaccines to their health, and by receiving the vaccine, they will become immune from Covid.

In the negative sentiment class, "selling", "paid", and "sold" appeared as some of the top words. This sentiment is assumed to be stemmed from the notion that the government has made a back-alley deal with vaccine manufacturers, and most likely the tweet came from buzzers or anti-government agents. Because clearly, the Government has declared that the vaccination will be free to all citizens of Indonesia. This kind of debate is normal on an online platform, and tweets with negative sentiments are often mucky with hoaxes and blatant slander.

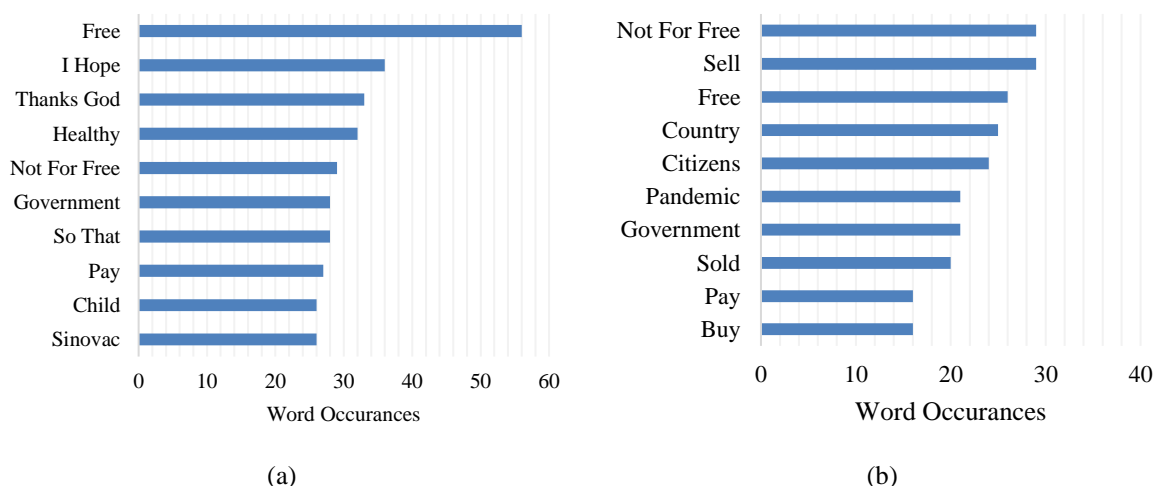


Figure 3. Top 10 unigrams of each sentiment class; (a) positive sentiment and (b) negative sentiment

Bigrams and trigrams forms of the tweets were also generated, and their respective top occurrences were also calculated. Figure 4 shows the Bigrams for the Figure 4(a) positive and Figure 4(b) negative classes. The top occurring bigrams in Figure 4 in positive sentiment class include "health protocol", "herd immunity", and "wearing mask" which portrays citizens' optimism toward the vaccination during the second wave of the pandemic. The people are optimistic to reach herd immunity as soon as the vaccine is administered widely to the public and say that the health protocol must be maintained even after vaccination. From the bigrams, positive sentiment narratives about optimism toward a better post-pandemic future are still consistent with the unigram form, the difference is that from the bigram, the collective effort of the Indonesian people can be seen, and clearly shows Indonesian collectivism as a culture. While in the negative sentiment bigrams as shown in Figure 4(b), "selling out the people", "making profit", and "mutual assistance", likely refer to the name of the vaccination campaign program, dominated the frequent bigrams. The negative sentiment bigram narratives also still focus on how vaccination policy is a kind of governmental fraud, aiming only to enrich big pharma and include shady back-alley deals, all of which are assumed to be generated by buzzers and fake accounts.

Figure 5 shows its trigrams form, also for the Figure 5(a) positive and Figure 5(b) negative classes. There is not much difference in top occurring trigrams compared to the bigrams as shown in Figure 5. On average, the narratives for each respective class still stand. Moreover, the preprocessed tweets are plugged into several machine learning classifiers to be further used for automatic sentiment analysis, specifically related to Indonesian governmental policy responses. The automatic sentiment analyzer will help the Indonesian government "test the water" and see public responses almost instantly.

The classifier algorithms used are XGBoost, random forest, and Naive Bayes. Before being plugged into the classifier, the TF-IDF values for each word are extracted. Each classifier has evaluation metrics value such as accuracy, precision, recall, and f1-score. Table 1 shows the performance comparison based on those metrics of the aforementioned classifiers.

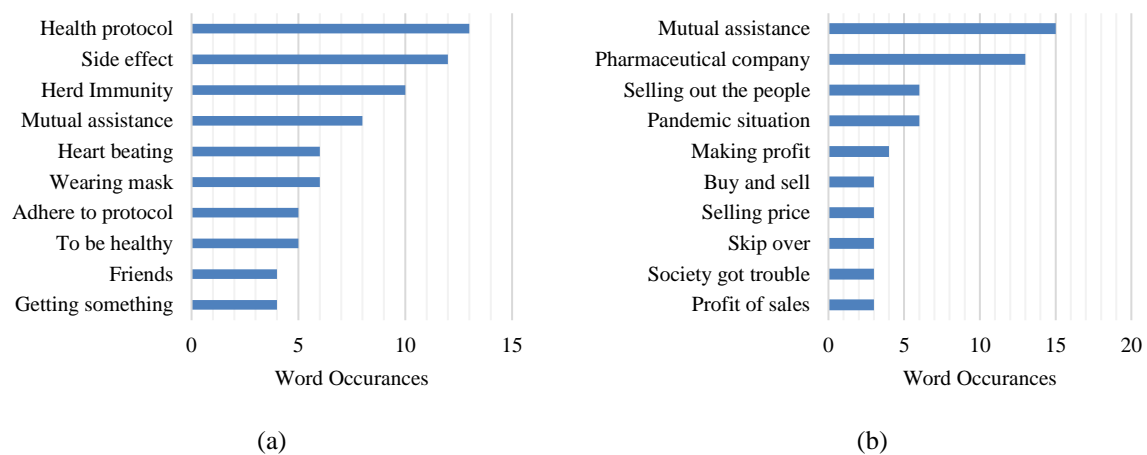


Figure 4. Top 10 bigrams in each sentiment; (a) positive sentiment and (b) negative sentiment

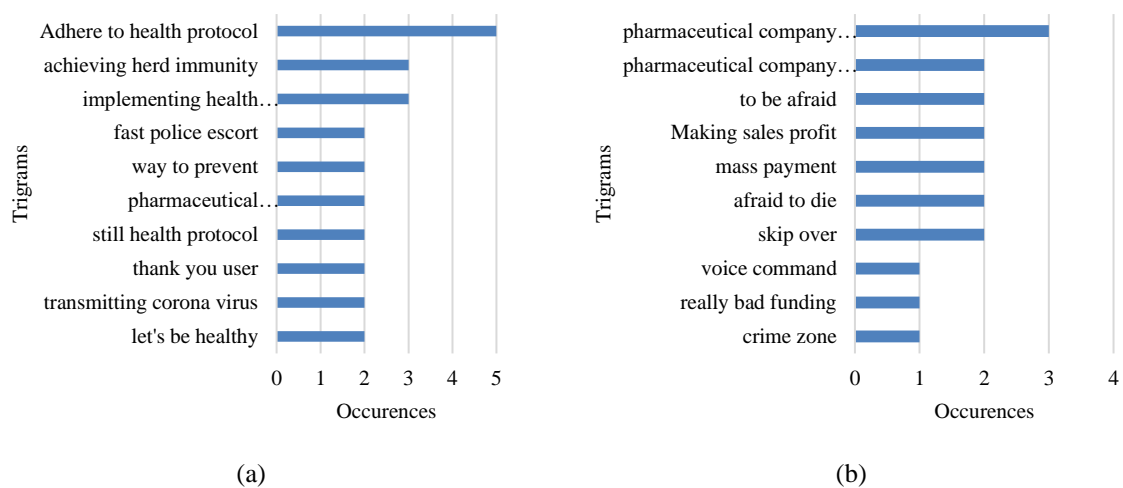


Figure 5. Top 10 trigrams for each sentiment; (a) positive sentiment and (b) negative sentiment

Table 1. Classifier performance

Classifiers	Accuracy (%)	Precision	Recall	F1-score
Random forest	73.6	0.75	0.63	0.65
Naive Bayes	72	0.76	0.55	0.55
XGBoost	75.9	0.64	0.63	0.63

From the classifier performance comparison, in terms of accuracy, XGBoost performs slightly well, although not so much compared to other baselines, much simpler and interpretable classifiers. Due to the dataset imbalance between the positive and negative sentiment tweets, the classifiers predict heavily into positive class. However, additional input of data into the negative class should solve this problem. In any case, TF-IDF as a feature extraction method coupled with several proven machine learning classifiers can be a quite reliable automatic sentiment classifier model. Overall, all classifiers exhibit satisfying performance.

3.1. Discussion

From the results, there are clear differences in narratives between the positive and negative sentiments about Indonesia's vaccination policy. The positive sentiments rely on hope and optimism as their driving factors, which are expected from Indonesian due to their cultural values. Contrary to the positive sentiments, the negative sentiments mostly expressed distrust and suspicion toward the Indonesian government and vaccine manufacturers or big pharma companies. These negative narratives are likely to be influenced by conspiracy theories and anti-government forum discussions.

However, it can also stem from the growing disappointment and skepticism toward the Indonesian government, due to their early pandemic anti-science stance. Understandably, people have negative sentiments toward the vaccination campaign because the Indonesian government initially denies covid as a serious national threat at the beginning of the pandemic, then suddenly promotes a science-backed campaign of mass vaccination. For machine learning modeling, the overall performance of the models is good, judging by the accuracy. However, due to the imbalanced class of positive and negative sentiments, the model tends to overgeneralize the positive sentiment. The models need more data in each respective class, but Indonesian tweets are noisy with many colloquialisms and mixed languages.

3.2. Limitations and future research

This research is limited by the availability of the corpus. As seen in the tweet trends, the discourse about vaccination policy only peaked in 2 days, then died down exponentially. That leads to an imbalanced class for positive and negative sentiments. This research also did not focus on neutral sentiment tweets. The main limitation is that Indonesian Twitter users liked to shorten their typing. A text normalization dictionary is available to use to preprocess those texts, but most of the time, Indonesian Twitter users still invent new shortened words, and it is hard to contain all of the possible shortened words in a dictionary. Future research can address this problem by developing a machine-learning model to predict the full word from shortened words.

Another limitation is that Indonesian Twitter users often use colloquial words and mixed languages, usually a mix of pure Indonesian and their ethnic-regional language. Although the use of code-switching is widespread in all regions of Indonesia, the corpus needed to standardize the language into Indonesian remains a challenge in the natural language processing field. Future research needs to address this problem by proposing a language standardizer by considering colloquial terms and code-switched text.

4. CONCLUSION

Governmental policy is often responded to with various sentiments. Twitter as an online platform enables people to directly express their concerns toward public policy, this allows for a healthier democracy at a national level. Machine learning models with a proper natural language processing method facilitate automating sentiment analysis so that the government can quickly receive feedback on their policy issues. Based on the accuracy value of the three methods used, XGBoost performs better than the other two methods. Previously, at the pre-processing stage, descriptive analysis was conducted on unigram, bigram, and trigram. Bigram analysis shows consistency with unigram in both positive and negative sentiments. Trigram analysis also shows the same context as bigrams and unigrams, so that unigrams are used in the classification analysis stage. It shows that with proper pre-processing and enough datasets, automatic public response sentiment analysis toward national policy is possible to be deployed, which eventually leads to direct communication between the government and the people.

ACKNOWLEDGEMENTS

This research is funded through Universitas Airlangga in the 2021 Novice Lecturer Research funding program number 274/UN3.1.17/PT/2021. We thank them for their support.




REFERENCES

- [1] A. Ninkov and L. Vaughan, "A webometric analysis of the online vaccination debate," *Journal of the Association for Information Science and Technology*, vol. 68, no. 5, pp. 1285–1294, May 2017, doi: 10.1002/asi.23758.
- [2] H. O. Wittman and B. J. Zikmund-Fisher, "The defining characteristics of Web 2.0 and their potential influence in the online vaccination debate," *Vaccine*, vol. 30, no. 25, pp. 3734–3740, May 2012, doi: 10.1016/j.vaccine.2011.12.039.
- [3] N. Calvert, F. Cutts, E. Miller, D. Brown, and J. Munro, "Measles in secondary school children: implications for vaccination policy," *Communicable disease report. CDR review*, vol. 4, no. 6, pp. R70-3, 1994.
- [4] V. Raghupathi, J. Ren, and W. Raghupathi, "Studying Public Perception about Vaccination: A Sentiment Analysis of Tweets," *International Journal of Environmental Research and Public Health*, vol. 17, no. 10, p. 3464, May 2020, doi: 10.3390/ijerph17103464.
- [5] J. Du, J. Xu, H.-Y. Song, and C. Tao, "Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with Twitter data," *BMC Medical Informatics and Decision Making*, vol. 17, no. S2, p. 69, Jul. 2017, doi: 10.1186/s12911-017-0469-6.
- [6] S. Myneni, N. K. Cobb, and T. Cohen, "Finding meaning in social media: content-based social network analysis of QuitNet to identify new opportunities for health promotion," *MEDINFO 2013*, pp. 807–811, 2013, doi: 10.3233/978-1-61499-289-9-807.
- [7] L. K. Larkey and M. Hecht, "A Model of Effects of Narrative as Culture-Centric Health Promotion," *Journal of Health Communication*, vol. 15, no. 2, pp. 114–135, Mar. 2010, doi: 10.1080/10810730903528017.
- [8] G. Hastings and A. Haywood, "Social marketing and communication in health promotion," *Health Promotion International*, vol. 6, no. 2, pp. 135–145, 1991, doi: 10.1093/heapro/6.2.135.
- [9] N. Corcoran, *Communicating health: strategies for health promotion*. Sage, 2013.




- [10] B. Trstenjak, S. Mikac, and D. Donko, "KNN with TF-IDF based Framework for Text Categorization," *Procedia Engineering*, vol. 69, pp. 1356–1364, 2014, doi: 10.1016/j.proeng.2014.03.129.
- [11] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries," *Proceedings of the first instructional conference on machine learning*, vol. 242, no. 1, pp. 29–48, 2003.
- [12] S. Qaiser and R. Ali, "Text mining: use of TF-IDF to examine the relevance of words to documents," *International Journal of Computer Applications*, vol. 181, no. 1, pp. 25–29, 2018.
- [13] V. D. Antonio, S. Efendi, and H. Mawengkang, "Sentiment analysis for Covid-19 in Indonesia on Twitter with TF-IDF featured extraction and stochastic gradient descent," *International Journal of Nonlinear Analysis and Applications*, vol. 13, no. 1, pp. 1367–1373, 2022, doi: 10.22075/IJNAA.2021.5735.
- [14] A. Patil, "Word Significance Analysis in Documents for Information Retrieval by LSA and TF-IDF using Kubeflow," 2022, pp. 335–348.
- [15] A. Chaturvedi, S. Yadav, M. A. M. H. Ansari, and M. Kanojia, "Comparative Multinomial Text Classification Analysis of Naïve Bayes and XGBoost with SMOTE on Imbalanced Dataset," 2022, pp. 339–349.
- [16] Y. Pandey, M. Sharma, M. K. Siddiqui, and S. S. Yadav, "Hate Speech Detection Model Using Bag of Words and Naïve Bayes," 2022, pp. 457–470.
- [17] E. Chen, K. Lerman, and E. Ferrara, "Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set," *JMIR Public Health and Surveillance*, vol. 6, no. 2, p. e19273, May 2020, doi: 10.2196/19273.
- [18] C. Suratnoaji, N. Nurhadi, and I. D. Arianto, "Public opinion on lockdown (PSBB) policy in overcoming Covid-19 pandemic in indonesia: Analysis based on big data twitter," *Asian Journal for Public Opinion Research*, vol. 8, no. 3, pp. 393–406, 2020, doi: 10.15206/ajpor.2020.8.3.393.
- [19] P. H. Prastyo, A. S. Sumi, A. W. Dian, and A. E. Permanasari, "Tweets Responding to the Indonesian Government's Handling of COVID-19: Sentiment Analysis Using SVM with Normalized Poly Kernel," *Journal of Information Systems Engineering and Business Intelligence*, vol. 6, no. 2, p. 112, Oct. 2020, doi: 10.20473/jisebi.6.2.112-122.
- [20] J. H. Jaman, R. Abdulrohman, A. Suharso, N. Sulistiowati, and I. P. Dewi, "Sentiment Analysis on Utilizing Online Transportation of Indonesian Customers Using Tweets in the Normal Era and the Pandemic Covid-19 Era with Support Vector Machine," *Advances in Science, Technology and Engineering Systems Journal*, vol. 5, no. 5, pp. 389–394, 2020, doi: 10.25046/aj050549.
- [21] I. A. Mansurnoor, "Local Initiative and Government Plans: Ulama" and Rural Development in Madura, Indonesia," *Sojourn: Journal of Social Issues in Southeast Asia*, pp. 69–94, 1992.
- [22] Y. Pribadi, "Islam and politics in Madura: ulama and other local leaders in search of influence (1990-2010)," *Leiden University*, 2013.
- [23] M. J. G. Criollo and A. V. V. Benavides, "YouTubers and its Digital Influence. Case Study: Ecuador and Colombia," in *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*, Jun. 2019, pp. 1–6, doi: 10.23919/CISTI.2019.8760719.
- [24] M. Nouri, "The power of influence: traditional celebrity vs social media influencer," *Santa Clara University Scholar Commons*, 2018.
- [25] N. Gamalliel, D. Saminarsih, and A. Taher, "Indonesia's second wave crisis: medical doctors' political role is needed more than ever," *The Lancet*, vol. 398, no. 10303, pp. 839–840, Sep. 2021, doi: 10.1016/S0140-6736(21)01807-9.
- [26] I. E. Tiffani, "Optimization of Naïve Bayes Classifier By Implemented Unigram, Bigram, Trigram for Sentiment Analysis of Hotel Review," *Journal of Soft Computing Exploration*, vol. 1, no. 1, Sep. 2020, doi: 10.52465/josce.v1i1.4.
- [27] S. I. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 2012, pp. 90–94.

BIOGRAPHIES OF AUTHORS






Muhammad Noor Fakhruzzaman    was born and raised in Surabaya. He holds an interdisciplinary master's degree in Human-Computer Interaction and Journalism & Mass Communication from Iowa State University. His current research interests fall between Data Science and Mass Communication, mainly automated media monitoring using natural language processing. He currently teaches at Data Science Technology Study Program at Universitas Airlangga. His research includes brain-computer interface, Indonesian natural language processing, and media monitoring. He is a member of Kappa Tau Alpha, an honor society of Journalism and Mass Communication studies. He also loves to train in grappling sports: Wrestling, Brazilian Jiu-jitsu (2nd-degree blue belt), and fanatically watch pro-wrestling shows. He can be contacted at email: ruzza@ftmm.unair.ac.id and ruzza@alumni.iastate.edu.






Sa'idah Zahrotul Jannah    was born and raised in Surabaya. She got her bachelor's and master's degree in Statistics from Institut Teknologi Sepuluh Nopember, Indonesia. Currently, she is a lecturer at Universitas Airlangga, Indonesia. Her research interest is data mining, multivariate analysis, and natural language processing. She can be contacted at email: s.zahrotul.jannah@fst.unair.ac.id.






Sie Wildan Gunawan    was born and raised in Samarinda. He is now an undergraduate student at Universitas Airlangga pursuing a degree in Robotics and Artificial Intelligence. His current research interests are computer vision and natural language processing. He can be contacted at email: sie.wildan.gunawan-2020@ftmm.unair.ac.id.



Angga Iryanto Pratama    was born in Malang. He currently studies in Data Science Technology Study Program at Universitas Airlangga, Surabaya. He is interested in data and front-end web development. He has data science experience across multiple courses, internships, and individual small projects using Python. Currently, he is the secretary of the Student Executive Body at the Faculty of Advanced Technology and Multidiscipline. He is a member of the Cinematography Universitas Airlangga Student Unit. He also loves graphic design and cinematography. He can be contacted at email: angga.yanto.pratama-2020@ftmm.unair.ac.id.



Denise Arne Ardanty    was born in Madiun. She currently studies in Data Science Technology Study Program at Airlangga University Surabaya. She is a member of Badan Legislatif Mahasiswa Faculty of Advanced Technology and Multidiscipline secretary. Highly interested in the environmental community and external communication. She can be contacted at email: denise.arne.ardanty-2020@ftmm.unair.ac.id.