Available online at: http://proceeding.rsfpress.com/index.php/ess/index
LPPM UPN "Veteran" Yogyakarta Conference Series
Proceeding on Engineering and Science Series (ESS)
Volume 1 Number 1 (2020): 253-261

Design Text Mining for Anxiety Detection using Machine Learning based-on Social Media Data during COVID-19 pandemic

Yuli Fauziah[1] , Shoffan Saifullah[1] , Agus Sasmito Aribowo[2]

[1]Universitas Pembangunan Nasional Veteran Yogyakarta
[2]Universiti Teknikal Malaysia Melaka, Malaysia
E-mail address yuli.fauziah@upnyk.ac.id; E-mail address shoffans@upnyk.ac.id; E-mail address sasmito.skom@upnyk.ac.id

Abstract

The COVID-19 pandemic has a profound impact on all groups, including governments, agencies, and individuals. It can make anxiety have a bad effect. So it is necessary to detect the existence of anxiety from the government to suppress and improve the community's psychology. This research aims to design text mining to detect anxiety during a pandemic by applying machine learning technology. Two methods of machine learning are designed, namely, random forest and xgboost. This design uses a sample of data from YouTube comments with a total of 4862 consisting of 3211 for negative data and 1651 for positive data. Negative data identify anxiety, while positive data identifies hope (not worry). The design of the application of this method was carried out by preliminary testing with three calculations, namely accuracy, precision, and recall. The accuracy of the Random Forest and XGBOOST methods is 83% and 73%. Meanwhile, precision and recall have an inversely proportional value. Random Forest has a precision value greater than 45% compared to xgboost. Whereas Recall, XGBOOST is bigger than ten compared to Random Forest. Random Forest can reference machine learning methods to detect someone's anxiety based on data from social media.

Keywords: anxiety detection, COVID-19, machine learning, random forest, xgboost

I. INTRODUCTION

Currently, the world is shocked by the widespread pandemic, namely coronavirus disease 2019 (COVID-19) (World Health Organization (WHO), 2020a). Previously, it was identified by the 2019-novel coronavirus (2019-nCoV) report on (World Health Organization (WHO), 2020b). This virus spread rapidly throughout the world, originating from Wuhan, China (Zhu *et al.*, 2020). Likewise, in Indonesia, which is experiencing the virus's spread very quickly (Setiati and Azwar, 2020). This spread caused anxiety and fear for everyone, including the Indonesian government. The government plays a very important role in handling this pandemic. The government has made various efforts to neutralize and stop the pandemic with various programs (Somawati *et al.*, 2020).

It has an impact on society. Some of these programs include assisting (Thaha, 2020), free electricity (Silalahi and Ginting, 2020), medical equipment supply (Nanur, Halu, and Juita, 2020).

Every government effort that has been made has pros and cons. It can be seen from news and social media. However, it is very visible from social media that can make comments and feedback regarding the shared news. In these comments, everyone from all walks of life can convey what they feel and express. Comments related to COVID-19 (details of the spread and death rate) and the government's efforts to deal with it have a massive impact on the community's psyche, namely anxiety and even panic (Saputra, 2020). Thus, the panic epidemic on social media spreads faster than the epidemic of disease, which is a challenge to control this pandemic (Ahmad and Murad, 2020).

Much anxiety detection has been carried out based on the concept of sentiment analysis (Aladağ *et al.*, 2018; Ragini, Anand, and Bhaskar, 2018; Yadav *et al.*, 2018). This study explains the anxiety that can be detected using sentiment analysis based on one's social media (Osadchiy, Mills, and Eleswarapu, 2020). It has not explained the concept of text mining in its identification. Besides, related studies explain that the detection of anxiety due to the COVID-19 pandemic can be analyzed through social media (Hamzah *et al.*, 2020; Ni *et al.*, 2020). The detection can be carried out based on the analysis of a psychologist. However, the sophistication of the concept of text mining can detect it quickly and can be trained. In supporting this, we use the concept of machine learning in the anxiety detection process based on an analysis of a person's sentiment during this pandemic. The text analyzed was taken from YouTube comments. While the machine learning methods used are Random Forest and xgboost.

II. LITERATURE REVIEW

Research on the detection of human emotions in social media has not satisfactory. There are two kinds of emoticons, namely, hate and like. Some divide it into three types: strong hate, weak hate, and like (Del Vigna et al., 2017). Besides, research is related to emotion analysis in detecting various fanaticism in text documents, namely Non-fanatic, Code Attitude fanaticism, and Code Red Fanaticism. Classification is carried out using the method Case-Based Reasoning (CBR) and the Naïve Bayes Classification. Accuracy reaches 77% (Almonayyes, 2006). The research was continued by using the method of machine learning (Random Forest). The resulting accuracy is 72% (Kléma and Almonayyes, 2006). But this research was carried out on Arabic language newspapers, including Al-Jazeera, Al-Arabia, Al-Watan, and Al-Qabas. Other methods of machine learning used are C4.5, RIPPER, and PART. The classification accuracy has reached 92% in structured domains (Almonayyes, 2016).

Further research is to combine the algorithm TF-IDF, Support Vector Machine, and Naïve Bayesian with an accuracy of 82.1% (Almonayyes, 2017). Research on sentiment analysis for hate speech detection in an unstructured environment, namely on social media Facebook (Del Vigna et al., 2017) in 2017. The method used is Support Vector Machines (SVM) and a particular Recurrent Neural Network (Long Short Term Memory (LSTM)). Both methods were tested for their performance to identify positive and negative emotions. The result is the level of effectiveness of both methods in the Italian language domain. Research conducted in 2015 used the lexicon method by combining a sentiment dictionary and a sentiment corpus. The result is a model for sentiment detection in English, with a maximum of 73% (Gitari et al., 2015). Another research is sentiment detection, which combines two methods: a two-step method for hate speech detection and technique embeddings to train a binary classifier to separate hate from neutral comments (Djuric et al., 2015).

Other studies detect emotions on social media that detect public sentiment in online media. Negative sentiment is defined as an unpleasant speech targeting certain religions, gender, and ethnicity (Warner and Hirschberg, 2012). There is also research to detect sentiment by applying a two-step method for sentiment detection (Djuric et al., 2015). Analytical sentiment on certain objects are oriented towards race, nationality, and religion (Gitari et al., 2015). Another research is the detection of analytical sentiment automatically using natural language processing (Schmidt and Wiegand, 2017) and the detection of analytical sentiment on Facebook (Del Vigna et al., 2017).

Research (Calderón-Monge, 2017) detects netizens' expectations in the realm of politics using social media Twitter. Besides, the detection of anxiety and fear was studied due to earthquakes (Vo and Collier, 2013), political (Chin, Zappone, and Zhao, 2016), and YouTube video comments (Chen, Chang, and Yeh, 2017). However, the results are not optimal because the machine learning used does not use the ensemble method as the Random Forest and XGBOOST methods.

III. RESEARCH METHODOLOGY

The methodology of this article consists of four main stages, namely, data collection, system analysis and concepts, and system design. Data collection has two ways, namely, literature study and observation data collection. Literature studies get references and resources on machine learning research and text mining. This concept aims to adapt and develop solutions for anxiety detection during a pandemic based on social media. The social media data used are comments and feedback from users (Saifullah, 2019). Besides, data collection uses observation on comments from YouTube videos (Rabbimov *et al.*, 2020). The data used video commentary on the official Indonesian government announcement about free electricity in Indonesia. These comments have emotions in the form of expectations and anxiety.

This research concept applies systems development technology using prototyping techniques to model the detection of sentiment and emotion opinions from social media (Bhati, 2020; Giannakis et al., 2020). The prototyping stage produced a model using three stages: preprocessing, emotion detection (sentiment analysis), and testing with cross-validation. Preprocessing performs retrieval data from social media. The data collected was preprocessed for normalization and detection. Preprocessing includes the process of tokenizing, filtering, stemming, tagging, and other methods by data conditions. The preprocessing results are stored in the database, according to Figure 1.

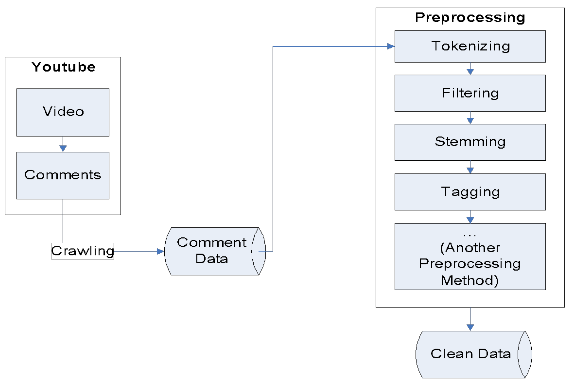


Figure 1. Preprocessing YouTube comments related to the Government Electricity Program

Figure 1 shows the preprocessing of public comments on government programs resulting from clean opinion data and processed by machine learning. The machine learning process is used to detect the sentiment (anxiety) from comment. The next process is feature extraction (Count

Vectorization, TF-IDF, and Hashing), application of machine learning methods, and cross-validation (Figure 2).

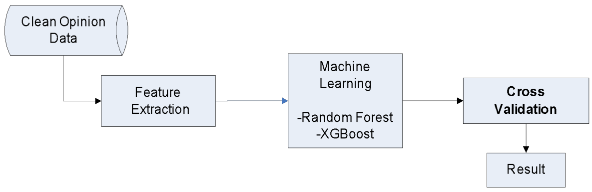
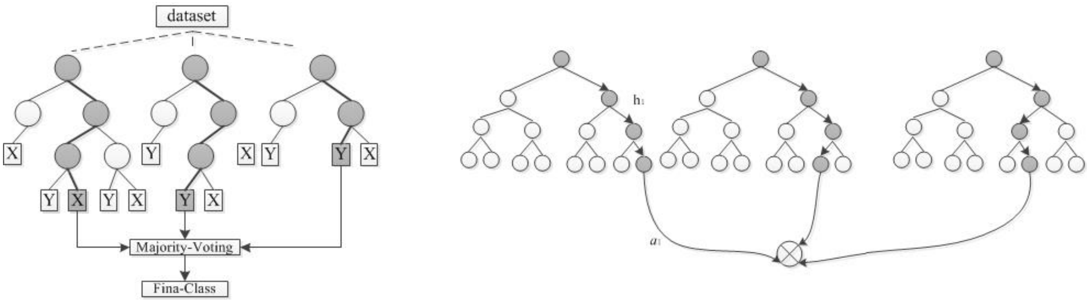


Figure 2. Stages of the Machine Learning Process and Result



(a) (b)

Figure 3. Schematic tree of (a) Random Forest and (b) xgboost

Machine learning uses Random Forest and XGBOOST methods. Random Forest (RF) is an algorithm used in the classification of sentiment analysis (Srujan *et al.*, 2018; Kumar, Yadava and Roy, 2019) and emotional analysis (Gokulakrishnan *et al.*, 2012; Jayalekshmi and Mathew, 2017) in the amount of data the big one. Random forest classification combine tree structure (Figure 3.(a)) to conduct training on sample data.

Extreme Gradient Boosting (XG-Boost) is the development of tree-based (Figure 3.(b)) classification algorithms (Georganos *et al.*, 2018). Extreme Gradient Enhancement mimics the algorithm of random forest creating trees (random forest with gradient descent/increase combination). Gradient Boosting is a machine learning technique for regression and classification problems using a predictive model in a weak ensemble prediction model (Zhang and Haghani, 2015). xgboost is a more efficient and scalable version of GBM. xgboost is capable of performing various functions such as regression, classification, and ranking.

Cross-validation in this study uses a confusion matrix, which is a method of measuring performance in the classification method. This method compares the results of the classification system with the real result. Performance measurement in *the confusion matrix* is shown in Table 1.

Table 1. Confusion Matrix.

Class	**Classified Positive**	**Classified Negative**
Positive	TP (True Positive)	FN (False Negative)
Negative	FP (False Positive)	TN (True Negative) The

The testing uses three calculations of the confusion matrix, namely accuracy, precision, and recall. Accuracy is used to calculate the closeness of the measurement results to the true value (1). Meanwhile, precision and recall are used to evaluate the performance of text mining and text analysis. Precision is used to measure accuracy (2), while recall is a measure of completeness (3).

Precision is the number of opinion samples labeled "true" as a positive sentiment divided by the total number of samples classified as a positive sentiment.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$precision = \frac{TP}{TP + FP} \quad (2)$$

$$recall = \frac{TP}{TP + FN} \quad (3)$$

IV. FINDING AND DISCUSSION

The results obtained and discussions related to this study are described in this section. In detail, this section describes the data set, results, and discussion based on the experiments that have been carried out. In addition, there is a cross-validation test.

IV.1. Data used and labeling.

We use data crawled from YouTube comments regarding the free electricity program from the Indonesian government. Data consists of 4862 data with an integer type. The data is classified into two parts, positive and negative. Positive comments indicate that the state who commented has anxiety = 0 or positive expectations. Meanwhile, negative comments indicate an anxious state. The data classification can be seen in Figure 4. The data for each group is 3211 for the negative group and 1651 for the positive group.

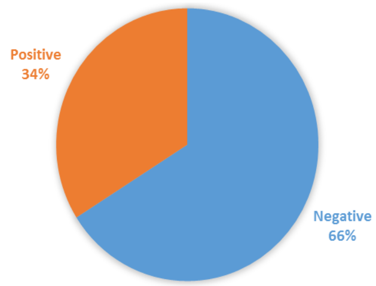


Figure 4. Percentage of Initial Data Classification for Anxiety Identification

Based on Figure 4, the sample sentences processed are Indonesian sentences. Sample sentences can be seen in Table 2. From all available data, Table 2 shows each of the five samples of negative and positive comments.

IV.2. Results and Discussion

Based on classified data, there are two groups, namely negative and positive. The process carried out is based on Figure 1 and Figure 2 using Python and its explanation. The results of data crawling from YouTube comments are presented in Table 2 as a sample.

The initial process is reading data and then preprocessing to clean data to have no noise (standard). This process is carried out in several methods, including tokenizing, filtering (converting slank-words and deleting numbers stopwords, images, duplicates), stemming, and emoticons conversion. In the conversion process, emoticons have rules to improve data cleanliness. There are additional rules, as follows:

a. A single figure word add the word "support" to get a positive sentiment and emotional trust = 1
b. An exclamation mark means the emotion of trust = 2

c. Capital letter sentence are angry emotions = 1
d. An angry emotion is an exclamation mark = 2

Table 2. Sample comments are categorized as negative and positive.

No	Comment Text	Category
1	450 b kok masih bayar	Negative
2	ini penyegelan banyak sekali tetangga sebelah.	Negative
3	Tidak semua yg memakai kwh 900 v rim itu mampu pak ,saya hidup ngontrak sedangkan kwh yg ada d sini 900 v RIM , bagaimana cra bayar nya kalau gk kerja gara2 dampak covid 19	Negative
4	Bagaimana dengan ri 900 v yg tidak bersubsidi , tidak dapat potongan dong ,sedangkan efek dari kasus covid 19 kan dampak nya ksemua org termasuk suami saya ,sudah tidak bekerja karna d lockdown ,sedangkan pembayaran listrik terus melonjak mahal,	Negative
5	Di sini udah jelas untung pelangan R1 450 VA digeratiskan dan untuk pelangan R1 900 VA di diskon 50% Tapi pada nyatanya pembayaran masih tetap sama engga add perubahan...	Negative
6	@Titian Devinta kayaknya harus banyak sabar dan hemat dlm situasi spt ini..	Positive
7	Ya gk pp di syukuri aj ya	Positive
8	@Toto Margono iya bener, hampir semua jenis usaha sedang mengalami lesu... Termasuk saya, dibidang Wedding Organizer semuanya dicancel total dikarenakan tidak diperbolehkan perkumpulan dengan melibatkan banyak masa. Sedangkan yg namanya pernikahan, resepsi ya pasti banyak orang makannya dicancel semua... Bener mas harus bersabar, semoga semua cepat kembali normal seperti biasa 😱	Positive
9	Ayo Yamaha dan Honda kamu pasti bisa menggratiskan juga kok 😂	Positive
10	Sabar nggeh, insyaallah rezekinya ada terus. Aamiin	Positive

The conversion result is a complete sentence without emoticons (Python coding). The emoticon is converted into a sentence to become the sentence ".... face_with_tears_of_joy".

```
kalimat=(conv_emoticon("Subsidi listrik saya dicabut. R1 jadi R1M...Bpjs saya
 kena pengurangan.Auto kaya SAYA!! Dimata pak presiden pastinya😂😂😂"))
for kata in kalimat.split(): print(kata)
```

Preprocessing is code by Python, as shown in Figure 5. The preprocessing results are removing the noise data (useless), so the sentence of the "text" column turns into a "clean_text" column. This preprocessing used all datasets, and the results were the core data from the comments. This data is used for the anxiety identification process so that in detail, it got better results.

```
clean_text2=pd.DataFrame({'clean_text2':hasillistrikgratis})
listrikgratis19fix=pd.concat([listrikgratis19,clean_text2],axis=1)
listrikgratis19fix.head(3347)
```

	No	Text	Sentiment	clean_text2
0	1	450 b kok masih bayar	Negative	bayar
1	2	ini penyegelan banyak sekali tetangga sebelah.	Negative	penyegelan tetanga sebelah
2	3	Tidak semua yg memakai kwh 900 v rim itu mampu...	Negative	tidak memakai rim hidup ngontrak ada rim cr...
3	4	Bagaimana dengan ri 900 v yg tidak bersubsidi ...	Negative	ri tidak bersubsidi tidak dapat potongan efek...
4	5	Di sini udah jelas untung pelangan R1 450 VA d...	Negative	untung pelangan r digeratiskan pelangan r d...

Figure 5. Preprocessing Process Results that Have Clean

The preprocessing data were processed by feature extraction using three vectorization methods, namely Count, TF-IDF, and Hashing. This method is used to obtain feature extraction results, which are then processed by machine learning (Random Forest and xgboost). Random Forest and XGBOOST were tried using testing based on test data and training data. There are two kinds of test data, namely x_test and y_test, with each amounting to 973, meanwhile, for training data using data as much as 3889.

Based on the results of the Python code design, each method is checked for validation. Checking is carried out using three methods with calculations using formulas (1), (2), and (3). The results of the implementation validation in python coding show that the Random Forest and XGBOOST methods are shown in Table 3. Based on the level of accuracy of the two methods, Random Forest is higher than xgboost. Likewise, with precision, Random Forest is higher than xgboost. However, Recall XGBOOST is more than ten times higher than the Random Forest.

Table 3. Results of the calculation of cross-validation from the Random Forest and XGBOOST methods.

Cross-Validation	Random Forest	xgboost
Accuracy	83.04%	73.17%
Precision	73.84%	27.32%
Recall	77.20%	89.52%

V. CONCLUSION AND FURTHER RESEARCH

Based on the results and discussion that has been done, the recommended machine learning method is the Random Forest method, which has an accuracy of more than 83% compared to XGBoost, which only has 73% accuracy. Apart from accuracy, precision and recall are other factors, where precision and recall for both methods have inversely proportional value. Random Forest has a precision value greater than 45% compared to xgboost. Whereas Recall, XGBOOST is more significant than 10% ten compared to Random Forest. Thus, Random Forest can reference machine learning methods to detect anxiety based on social media data.

REFERENCES

Ahmad, A. R., and Murad, H. R. (2020) 'The Impact of Social Media on Panic During the COVID-19 Pandemic in Iraqi Kurdistan: Online Questionnaire Study', *Journal of Medical Internet Research*, 22(5), p. e19556. DOI: 10.2196/19556.

Aladağ, A. E. *et al.* (2018) 'Detecting Suicidal Ideation on Forums: Proof-of-Concept Study,' *Journal of Medical Internet Research*, 20(6), p. e215. DOI: 10.2196/jmir.9840.

Almonayyes, A. (2006) 'Multiple Explanations Driven Naive Bayes Classifier.', *Journal of Universal Computer Science*, 12(2), pp. 127–139.

Almonayyes, A. (2016) 'Classifying Documents By Integrating Contextual Knowledge With Boosting,' in *International Conference on Artificial Intelligence and Computer Science*, pp. 28–29.

Almonayyes, A. (2017) 'Tweets Classification Using Contextual Knowledge And Boosting', *International Journal of Advances in Electronics and Computer Science*, (4), pp. 87–92.

Bhati, R. (2020) 'Sentiment analysis a deep survey on methods and approaches.' Science & Engineering Research Support society.

Calderón-Monge, E. (2017) 'Twitter to Manage Emotions in Political Marketing,' *Journal of Promotion Management*. Taylor & Francis, 23(3), pp. 359–371. DOI: 10.1080/10496491.2017.1294870.

Chen, Y. L., Chang, C. L. and Yeh, C. S. (2017) 'Emotion Classification of YouTube Videos,' *Decision Support Systems*, 101, pp. 40–50. DOI: 10.1016/j.dss.2017.05.014.

Chin, D., Zappone, A. and Zhao, J. (2016) 'Analyzing Twitter Sentiment of the 2016 Presidential Candidates', *Applied Informatics and Technology Innovation Conference (AITIC 2016)*.

Djuric, N. *et al.* (2015) 'Hate Speech Detection with Comment Embeddings,' pp. 29–30.

Georganos, S. *et al.* (2018) 'Very High-Resolution Object-Based Land Use–Land Cover Urban Classification Using Extreme Gradient Boosting,' *IEEE Geoscience and Remote Sensing Letters*, 15(4), pp. 607–611. DOI: 10.1109/LGRS.2018.2803259.

Giannakis, M. *et al.* (2020) 'Social media and sensemaking patterns in new product development: demystifying the customer sentiment,' *Annals of Operations Research*. DOI: 10.1007/s10479-020-03775-6.

Gitari, N. D. *et al.* (2015) 'A Lexicon-based Approach for Hate Speech Detection,' *International Journal of Multimedia and Ubiquitous Engineering*, 10(4), pp. 215–230. DOI: 10.14257/ijmue.2015.10.4.21.

Gokulakrishnan, B. *et al.* (2012) 'Opinion mining and sentiment analysis on a Twitter data stream,' in *International Conference on Advances in ICT for Emerging Regions (ICTer2012)*. IEEE, pp. 182–188. DOI: 10.1109/ICTer.2012.6423033.

Hamzah, F. A. B. *et al.* (2020) 'CoronaTracker: worldwide COVID-19 outbreak data analysis and prediction', *Bull World Health Organ*, 1, p. 32.

Jayalekshmi, J. and Mathew, T. (2017) 'Facial expression recognition and emotion classification system for sentiment analysis,' in *2017 International Conference on Networks & Advances in Computational Technologies (NetACT)*. IEEE, pp. 1–8. DOI: 10.1109/NETACT.2017.8076732.

Kléma, J. and Almonayyes, A. (2006) 'Automatic Categorization of Fanatic Text Using random Forests,' *Kuwait Journal of Science and Engineering*, 33(2), pp. 1–18.

Kumar, S., Yadava, M. and Roy, P. P. (2019) 'Fusion of EEG response and sentiment analysis of products review to predict customer satisfaction,' *Information Fusion*, 52, pp. 41–52. DOI: 10.1016/j.inffus.2018.11.001.

Nanur, F. N., Halu, S. A. N. and Juita, E. (2020) 'EVALUASI KETERSEDIAAN FASILITAS KESEHATAN YANG MEMADAI TERHADAP PENCAPAIAN REVOLUSI KIA DI MANGGARAI', *Jurnal Kebidanan*, 12(01), pp. 80–92.

Ni, M. Y. *et al.* (2020) 'Mental Health, Risk Factors, and Social Media Use During the COVID-19 Epidemic and Cordon Sanitaire Among the Community and Health Professionals in Wuhan, China: Cross-Sectional Survey', *JMIR Mental Health*, 7(5), p. e19009. DOI: 10.2196/19009.

Osadchiy, V., Mills, J. N., and Eleswarapu, S. V. (2020) 'Understanding Patient Anxieties in the Social Media Era: Qualitative Analysis and Natural Language Processing of an Online Male Infertility Community,' *Journal of Medical Internet Research*, 22(3), p. e16728. DOI: 10.2196/16728.

Rabbimov, I. *et al.* (2020) 'Investigating the Effect of Emoji in Opinion Classification of Uzbek Movie Review Comments,' *arXiv preprint arXiv:2008.00482*.

Ragini, J. R., Anand, P. M. R. and Bhaskar, V. (2018) 'Big data analytics for disaster response and recovery through sentiment analysis,' *International Journal of Information Management*, 42, pp. 13–24. DOI: 10.1016/j.ijinfomgt.2018.05.004.

Saifullah, S. (2019) 'Fuzzy-AHP approach using Normalized Decision Matrix on Tourism Trend Ranking based-on Social Media,' *Jurnal Informatika*, 13(2), p. 16. DOI: 10.26555/jifo.v13i2.a15268.

Saputra, T. A. (2020) 'BENTUK KECEMASAN DAN RESILIENSI MAHASISWA PASCASARJANA ACEH-YOGYAKARTA DALAM MENGHADAPI PANDEMI COVID-19', *JURNAL BIMBINGAN DAN KONSELING AR-RAHMAN*, 6(1), pp. 55–61.

Schmidt, A., and Wiegand, M. (2017) 'A Survey on Hate Speech Detection using Natural Language Processing,' in *Fifth International Workshop on Natural Language Processing for Social Media*, pp. 1–10. DOI: 10.18653/v1/W17-1101.

Setiati, S. and Azwar, M. K. (2020) 'COVID-19 and Indonesia', *Acta Medica Indonesiana*, 52(1), pp. 84–89. Available at: https://www.scopus.com/record/display.uri?eid=2-s2.0-85083414691&origin=inward.

Silalahi, D. E. and Ginting, R. R. (2020) 'Strategi Kebijakan Fiskal Pemerintah Indonesia Untuk Mengatur Penerimaan dan Pengeluaran Negara Dalam Menghadapi Pandemi Covid-19', *Jesya (Jurnal Ekonomi & Ekonomi Syariah)*, 3(2), pp. 156–167.

Somawati, A. V. *et al.* (2020) *Bali vs COVID-19: Book Chapters*. Nilacakra.

Srujan, K. S. *et al.* (2018) 'Classification of Amazon Book Reviews Based on Sentiment Analysis,' in, pp. 401–411. DOI: 10.1007/978-981-10-7512-4_40.

Thaha, A. F. (2020) ‘Dampak covid-19 terhadap UMKM di Indonesia’, *BRAND Jurnal Ilmiah Manajemen Pemasaran*, 2(1), pp. 147–153.

Del Vigna, F. *et al.* (2017) 'Hate me, hate me not: Hate speech detection on Facebook,' in *First Italian Conference on Cybersecurity (ITASEC17)*, pp. 86–95. DOI: 10.1051/matecconf/201712502035.

Vo, B.-K. H. and Collier, N. (2013) 'Twitter Emotion Analysis in Earthquake Situations,' *International Journal of Computational Linguistics and Applications*, 4(1), pp. 159–173.

Warner, W. and Hirschberg, J. (2012) 'Detecting Hate Speech on the World Wide Web,' in *Workshop on Language in Social Media (LSM 2012)*, pp. 19–26.

World Health Organisation (WHO) (2020a) *Novel Coronavirus(2019-nCoV) Situation Report-22, 11 February 2020*. Available at: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200211-sitrep-22-ncov.pdf?sfvrsn=fb6d49b1_2.

World Health Organisation (WHO) (2020b) *Novel Coronavirus (2019-nCoV), Situation Report-1, 21 January 2020.* Available at: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200121-sitrep-1-2019-ncov.pdf?sfvrsn=20a99c10_4.

Yadav, S. *et al.* (2018) 'Medical sentiment analysis using social media: towards building a patient assisted system,' in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Zhang, Y. and Haghani, A. (2015) 'A gradient boosting method to improve travel time prediction,' *Transportation Research Part C: Emerging Technologies*, 58, pp. 308–324. DOI: 10.1016/j.trc.2015.02.019.

Zhu, N. *et al.* (2020) ‘A Novel Coronavirus from Patients with Pneumonia in China, 2019’, *New England Journal of Medicine*, 382(8), pp. 727–733. DOI: 10.1056/NEJMoa2001017.