Sentiment Analysis Using Machine Learning Approach Based on Feature Extraction for Anxiety Detection

Shoffan Saifullah[1,2][0000-0001-6799-3834], Rafał Dreżewski[1][0000-0001-8607-3478], Felix Andika Dwiyanto[1][0000-0002-7431-493X], Agus Sasmito Aribowo[2][0000-0003-3279-1268], and Yuli Fauziah[2][0000-0002-3745-6189]

[1] Institute of Computer Science, AGH University of Science and Technology,
Kraków, Poland
{saifulla,drezew,dwiyanto}@agh.edu.pl

[2] Department of Informatics, Universitas Pembangunan Nasional Veteran
Yogyakarta, Yogyakarta, Indonesia
{shoffans,sasmito.skom,yuli.fauziah}@upnyk.ac.id

Abstract. In this study, selected machine learning (ML) approaches were used to detect anxiety in Indonesian-language YouTube video comments about COVID-19 and the government's program. The dataset consisted of 9706 comments categorized as positive and negative. The study utilized ML approaches, such as KNN (K-Nearest Neighbors), SVM (Support Vector Machine), DT (Decision Tree), Naïve Bayes (NB), Random Forest (RF), and XG-Boost, to analyze and classify comments as anxious or not anxious. The data was preprocessed by tokenizing, filtering, stemming, tagging, and emoticon conversion. Feature extraction (FE) is performed by CV (count-vectorization), TF-IDF (term frequency–inverse document frequency), Word2Vec (Word Embedding), and HV (Hashing-Vectorizer) algorithms. The 24 of the ML and FE algorithms combinations were used to achieve the best performance in anxiety detection. The combination of RF and CV obtained the best accuracy of 98.4%, which is 14.3 percentage points better than the previous research. In addition, the other ML methods accuracy was above 92% for CV, TF-IDF, and HV, while KNN obtained the lowest accuracy.

Keywords: Anxiety Detection · Machine Learning · Sentiment Analysis · Text Feature Extraction · Text Mining · Model Performance

1 Introduction

Anxiety is a mental disorder [2] related to the nervous system, with characteristics such as considerable and persistent anxiety, excitation of autonomic nervous activity, and excessive alertness. The types of anxiety disorders include, among others, generalized anxiety disorder, panic disorder, and social anxiety disorder [8]. The COVID-19 pandemic caused anxiety and stress for the Indonesian people and government, who implemented programs to reduce the pandemic,

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_38

with a waiver of medical supplies, community assistance, and a waiver of electricity bills. However, the program had many pros and cons comments on social media, which indicated the growth of public anxiety and eventually could cause a global anxiety pandemic if not addressed correctly by future government programs [1].

While psychologists can analyze a person's anxiety, using computer technology we can quickly analyze a large amount of social media data. In this study, the artificial intelligence (AI) and sentiment analysis algorithms are used to detect anxiety [20] based on text processing [19,17] of comments shared on social media concerning COVID-19 [18]. In the proposed approach, data consisting of YouTube comments on Indonesian government COVID-19 programs is processed in sequential steps using machine learning methods such as K-NN, NB, DT, SVM, RF, and XG-boost, and feature extraction methods, such as CV, TF-IDF, HV, and Word2Vec.

This paper consists of five main sections. After the introduction, Section 2 presents the related research works. Section 3 explains the proposed method and the research steps. The conducted experiments and obtained results are discussed in Section 4. The conclusions based on experimental results are presented in Section 5.

2 Related Research

Emotion detection can be performed using the data science approach, text mining algorithms and sentiment analysis methods on text data from social media such as Twitter, YouTube, and Facebook. Negative sentiment on social media around topics such as gender, ethnicity, and religion can be used as input data to detect emotions.

An approach to detecting hate speech in text documents using 2 or 3 labels, and Case-Based Reasoning (CBR) and Naïve Bayes (NB) classification methods was applied to detect emotions [9] and bigotry, achieving an accuracy of over 77% [3]. Several researchers analyzed Arabic language sentiment using Random Forest (RF) and got low accuracy of 72% [15]. However, C4.5, RIPPER, and PART methods increased sentiment classification accuracy to 96% and Support Vector Machine (SVM) and Naïve Bayes (NB) with term frequency–inverse document frequency (TF-IDF) have an accuracy of 82.1% in detecting sentiment analysis [3]. Sentiment analysis has been used to recognize emotions and hate speech in Facebook comments in Italian [9] and English [14] with ML methods, such as the SVM, Recurrent Neural Network (RNN), and Long Short Term Memory (LSTM). A lexicon approach based on a dictionary and sentiment corpus has been used to obtain 73% accuracy in [14]. A binary classifier to distinguish between neutral and hate speech was applied in [10].

Different techniques, like paragraph2vec, Continuous Bag of Words (CBOW), and embedding-binary classifier are also used to perform sentiment analysis [10]. Natural Language Processing (NLP) is utilized [23] to automatically detect emotions about nation, religion, and race [14] in sentiment analysis of Facebook comments [9]. Twitter and YouTube comments can also be used to detect user expectations and identify mass anxiety and fear, such as natural disasters (e.g., earthquakes) and political battles [7,6].

To improve the results of anxiety detection, in this paper we applied several machine learning methods such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT), Naïve Bayes (NB), Random Forest (RF), and XG-Boost, together with selected feature extraction (FE) methods like count-vectorization (CV), term frequency–inverse document frequency (TF-IDF), Word Embedding (Word2Vec), and Hashing-Vectorizer (HV). Moreover, the ensemble concept was used to ensure the optimal results.

3 Proposed Method

The research presented in this paper builds on previous results using 6 ML methods and 4 FE methods to detect anxiety based on sentiment and emotional analysis of YouTube text comments. The proposed approach involves four main steps (data collection, preprocessing, feature extraction, and classification) to identify anxiety levels based on sentiment analysis of Indonesian language YouTube comments. This study also adopts a prototyping method with system modeling to identify sentiments in emotional data from social media [4,13]. As shown in Figure 1, the process involves preprocessing, emotion detection based on sentiment analysis, and cross-validation testing.

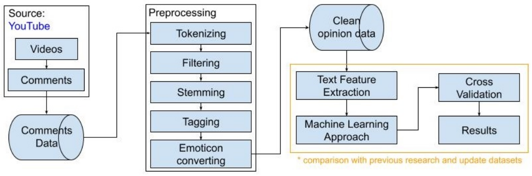


Fig. 1. The sentiment analysis flow using ML algorithms on YouTube comments.

ML methods such as RF and XG-Boost [11,22] are used (along with several other methods like KNN, NB, DT, and SVM to compare the results with [22]) together with selected feature extraction (FE) algorithms. The ML methods results are assessed using the confusion matrix to calculate the performance (accuracy, precision, recall, and F1 Score).

KNN identifies classes based on a distance matrix, and the best K value classification [21] is found using Euclidean Distance (ED) matrix. The second ML method is NB, which uses binary features as vector attributes to identify words based on the probability of their occurrence. The method is robust and can handle noise and missing data. DT is a method, represented as a tree structure, that can be used for data classification and pattern prediction. The relationship between the attribute variable x and the target y is depicted using internal nodes as attribute tests, branches as test results, and outer nodes as labels. SVM are supervised learning models used for classification and regression. SVM requires training and testing phases to find the best hyperplane that acts as a separator of two data classes. It works on high-dimensional datasets and uses a few selected data points to form a model (support vector).

RF uses ensemble learning and can be used to solve regression and classification problems (also based on sentiment analysis [16]). RF can reduce the problems of overfitting and missing data, and it can handle datasets containing categorical variables. Extreme Gradient Boosting (XG-Boost) is a tree-based algorithm that can be applied to classification and regression problems [12]. This algorithm mimics the RF behavior and is combined with gradient descent/boosting. Gradient Boosting is a machine learning concept used to solve regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models. XG-Boost is a version of GBM (Gradient Boosting Machine) with some advantages, including accuracy, efficiency and scalability, which works well for applications such as regression, classification, and ranking.

The proposed approach also utilizes selected FE methods (see Figure 1), including CV, TF-IDF, HV, and Word2Vec, to convert text to its numerical representation that is then used by ML models. The CV method calculates the frequency of occurrence of the detected words, while TF-IDF is a numerical statistic method used for weighting text data. HV transforms a collection of text documents into a matrix of token occurrences. Word2Vec generates a vector representation for each word in a corpus, based on the context in which it appears.

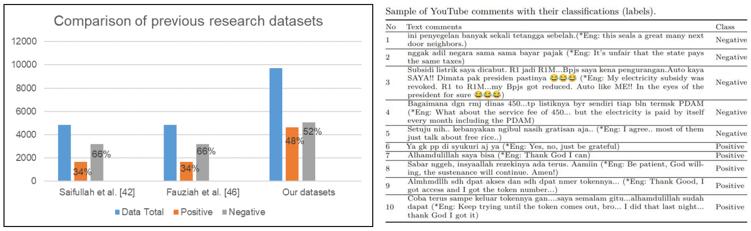
4 Results and Discussion

This section presents and discusses the experimental results of sentiment analysis for anxiety detection using ML approaches. We discuss some key issues like the used dataset and its labeling, the experimental evaluation of the proposed approach, and comparison with the previous research.

The dataset contains a total of 9,706 YouTube comments related to the Indonesian government's COVID-19 program, with 4,862 data from previous research [22,11] and 4,844 newly crawled comments. The comments are labeled as positive ("0") or negative, with negative comments indicating anxiety [5] but not necessarily the hate speech. The dataset was expanded to balance the number of positive and negative comments and avoid overfitting. Currently, the number of positive comments has almost equaled the number of negative comments (the number of positive comments is now about 90% of the number of negative comments, compared to about 50% in previous studies)—see Figure 2a.

Figure 2a compares Indonesian YouTube comments dataset with those used in previous studies. The dataset is labeled based on the application of sentiment analysis in the process of anxiety detection. The dataset cleaning process involves using "Literature" library and adding stop-words (757) and true-words (13770).

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_38



(a)

Sample of YouTube comments with their classifications (labels).

No	Text comments	Class
1	ini penyegelan banyak sekali tetangga sebelah.(*Eng: this seals a great many next door neighbors.)	Negative
2	nggak adil negara sama sama bayar pajak (*Eng: It's unfair that the state pays the same taxes)	Negative
3	Subsidi listrik saya dicabut. R1 jadi R1M...Bpjs saya kena pengurangan.Auto kaya SAYA!! Dimata pak presiden pastinya 😂😂😂 (*Eng: My electricity subsidy was revoked. R1 to R1M...my Bpjs got reduced. Auto like ME!! In the eyes of the president for sure 😂😂😂)	Negative
4	Bagaimana dgn rmj dinas 450...tp listiknya byr sendiri tiap bln termsk PDAM (*Eng: What about the service fee of 450... but the electricity is paid by itself every month including the PDAM)	Negative
5	Setuju nih.. kebanyakan ngibul nasih gratisan aja.. (*Eng: I agree.. most of them just talk about free rice..)	Negative
6	Ya gk pp di syukuri aj ya (*Eng: Yes, no, just be grateful)	Positive
7	Alhamdulillah saya bisa (*Eng: Thank God I can)	Positive
8	Sabar nggeh, insyaallah rezekinya ada terus. Aamiin (*Eng: Be patient, God willing, the sustenance will continue. Amen!)	Positive
9	Almhmdllh sdh dpat akses dan sdh dpat nmer tokennya... (*Eng: Thank Good, I got access and I got the token number...)	Positive
10	Coba terus sampe keluar tokennya gan....saya semalam gitu...alhamdulillah sudah dapat (*Eng: Keep trying until the token comes out, bro... I did that last night... thank God I got it)	Positive

(b)

Fig. 2. (a) Comparison with previous research data, and (b) sample datasets.

Figure 2b shows the sample YouTube comments in Indonesian, categorized as positive or negative.

The experimental results indicated that the dataset requires preprocessing, including tokenizing, filtering (slang words conversion, removing numbers, removing stop-words, removing figures, removing duplicates), stemming, tagging, and emoticon conversion. The dataset is split into the training set (80%) and testing set (20%) using a Python script. Additional rules are added to improve the data cleaning, such as assigning emotional trust, confidence, and anger based on specific criteria. The conversion of emoticons to text indicates the emotional expression. For example, in data in row 3 (presented in Figure 2b), the emoticon 😂 is converted to "*face_with_tears_of_joy.*". The result of preprocessing is clean data (without meaningless or useless text) that can be used as input to the next process.

In this research, we evaluated 24 modeling scenarios using 4 performance metrics of the confusion matrix. Furthermore, the new data was used to compute the confusion matrix, as shown in Figure 3. The best accuracy of 98.4% was achieved using RF-CV. The other ML methods, such as SVM, DT, and XG-Boost, can identify anxiety with the accuracy of over 92%. Word2Vec has the lowest accuracy when used with most ML methods, except when used with KNN, in which case it performs better than the rest of the FE methods used with KNN. It also obtained better results than the NB-HV method. It is because Word2Vec converts words into vectors and is trained between conditional sentences.

The RF algorithm when used with all FE methods is highly accurate and balanced in terms of precision, recall, and F1 Score. However, RF with Word2Vec has only 81.9% accuracy as compared to the other FE methods (above 96%). In addition, the RF method is superior in detecting anxiety based on sentiment analysis, with consistent performance close to or above 95%. The final results (Figure 3) showed that SVM, DT, RF, and XG-Boost methods used with CV, TF-IDF, and HV achieved the accuracy close to or above 90%. In addition, KNN and NB obtained lower accuracy compared to other ML methods. The Word2Vec obtained the lowest accuracy in each experiment (less than 82% for all ML models used). However, despite its poor performance, the Word2Vec method is superior to other FE methods when used with KNN algorithm. HV

DOI: 10.1007/978-3-031-36021-3_38

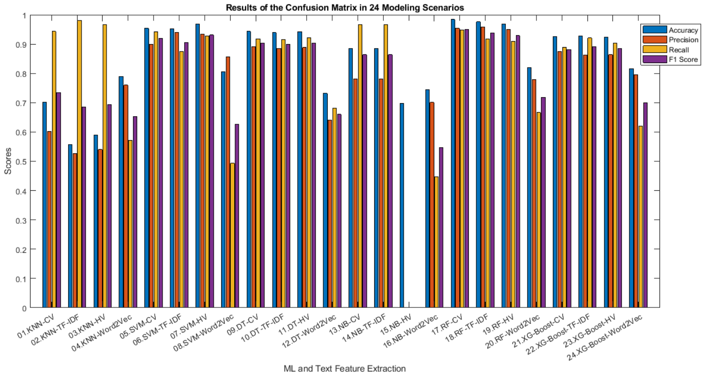


Fig. 3. The performance of ML models used with FE methods (24 scenarios).

method obtained higher accuracy (94%–96%) when used with the SVM, DT, and RF algorithms.

The research presented in this paper improves the previous results [22,11] on anxiety detection based on sentiment analysis, resulting in enhanced methods and outcomes, as shown in Table 1. In addition, we conducted experiments using previous research data and newly added crawled data, shown in Figure 2.

The proposed methods achieved better results than previous studies on anxiety detection based on sentiment analysis. The RF-CV method had the highest improvement, with a 13.4 percentage points increase in accuracy (from 85% to 98.4%). The other methods (SVM, DT, NB, and XG-Boost) outperformed previous studies with the accuracy above 85%. The addition of several processes in the preprocessing phase (including stop-words) and the extension of the dataset have improved the proposed method's performance.

5 Conclusions

The research presented in this paper improved the accuracy of methods proposed in [22] by using additional datasets, preprocessing, and text feature extraction to better detect psychological factors. The proposed machine learning based approach is a contribution to the research on detecting anxiety based on sentiment analysis. The best and recommended method is RF-CV, which has obtained the accuracy of 98.4% and consistent precision, recall, and F1 scores with values over 95%. SVM, DT, and XG-Boost methods had also good accuracy, but their performance still needs improvement. The future work will include the application of optimization algorithms and Deep Learning methods.

Acknowledgement. The research presented in this paper was partially supported by the funds of Polish Ministry of Education and Science assigned to AGH University of Science and Technology.

Table 1. Comparison of the obtained results (FE: Feature Extraction; Acc: Accuracy; Prec: Precision; Rec: Recall; F1: F1 Score).

No	Method	FE	Saifullah et al. [22]				Initial improvement				Our proposed model			
			Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
1	KNN	CV	0.601	0.468	0.927	-	0.606	0.471	0.924	0.624	0.701	0.602	0.943	0.735
2		TF-IDF	0.391	0.364	0.968	-	0.391	0.364	0.968	0.529	0.557	0.526	0.981	0.685
3		HV	-	-	-	-	0.464	0.395	0.971	0.561	0.589	0.540	0.966	0.693
4		Word2Vec	-	-	-	-	0.687	0.594	0.358	0.446	0.790	0.760	0.572	0.653
5	SVM	CV	0.815	0.797	0.640	-	0.815	0.801	0.634	0.708	0.954	0.899	0.941	0.920
6		TF-IDF	0.799	0.866	0.509	-	0.798	0.866	0.506	0.639	0.953	0.940	0.875	0.906
7		HV	-	-	-	-	0.802	0.854	0.529	0.654	0.968	0.934	0.928	0.931
8		Word2Vec	-	-	-	-	0.687	0.831	0.142	0.243	0.805	0.856	0.493	0.626
9	DT	CV	0.804	0.715	0.738	-	0.8	0.713	0.724	0.719	0.943	0.891	0.918	0.904
10		TF-IDF	0.806	0.720	0.738	-	0.806	0.716	0.747	0.731	0.939	0.885	0.915	0.900
11		HV	-	-	-	-	0.81	0.728	0.738	0.733	0.942	0.888	0.921	0.904
12		Word2Vec	-	-	-	-	0.665	0.526	0.526	0.526	0.732	0.640	0.681	0.660
13	NB	CV	0.823	0.839	0.619	-	0.824	0.839	0.622	0.715	0.885	0.781	0.966	0.864
14		TF-IDF	0.823	0.839	0.619	-	0.824	0.839	0.622	0.715	0.885	0.781	0.966	0.864
15		HV	-	-	-	-	0.646	0	0	0	0.698	0.000	0.000	0.000
16		Word2Vec	-	-	-	-	0.66	0.53	0.331	0.408	0.744	0.701	0.447	0.546
17	RF	CV	0.850	0.786	0.791	-	0.841	0.776	0.773	0.774	0.984	0.954	0.948	0.951
18		TF-IDF	0.826	0.785	0.701	-	0.827	0.778	0.715	0.745	0.976	0.959	0.918	0.938
19		HV	-	-	-	-	0.83	0.81	0.68	0.739	0.969	0.950	0.909	0.929
20		Word2Vec	-	-	-	-	0.73	0.683	0.439	0.535	0.819	0.779	0.666	0.718
21	XG-Boost	CV	0.732	0.895	0.273	-	0.732	0.895	0.273	0.419	0.925	0.874	0.889	0.881
22		TF-IDF	0.747	0.871	0.334	-	0.744	0.868	0.326	0.474	0.928	0.863	0.921	0.891
23		HV	-	-	-	-	0.738	0.83	0.326	0.468	0.924	0.864	0.904	0.884
24		Word2Vec	-	-	-	-	0.739	0.732	0.413	0.528	0.816	0.796	0.621	0.700

References

1. Ahmad, A.R., Murad, H.R.: 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), e19556 (2020). https://doi.org/10.2196/19556
2. Ahmed, A., Aziz, S., Khalifa, M., Shah, U., Hassan, A., Abd-Alrazaq, A., Househ, M.: Thematic analysis on user reviews for depression and anxiety chatbot apps: Machine learning approach. JMIR Formative Research **6**(3), e27654 (2022)
3. Almonayyes, A.: Tweets Classification Using Contextual Knowledge And Boosting. International Journal of Advances in Electronics and Computer Science **4**(4), 87–92 (2017)
4. Bhati, R.: Sentiment analysis a deep survey on methods and approaches. International Journal of Disaster Recovery and Business Continuity **11**(1), 503–511 (2020)
5. Cahyana, N.H., Saifullah, S., Fauziah, Y., Aribowo, A.S., Drezewski, R.: Semi-supervised text annotation for hate speech detection using k-nearest neighbors and term frequency-inverse document frequency. International Journal of Advanced Computer Science and Applications **13**(10) (2022). https://doi.org/10.14569/ijacsa.2022.0131020
6. Calderón-Monge, E.: Twitter to manage emotions in political marketing. Journal of Promotion Management **23**(3), 359–371 (2017)
7. Chin, D., Zappone, A., Zhao, J.: Analyzing Twitter Sentiment of the 2016 Presidential Candidates. Applied Informatics and Technology Innovation Conference (AITIC 2016) (2016)
8. Czornik, M., Malekshahi, A., Mahmoud, W., Wolpert, S., Birbaumer, N.: Psychophysiological treatment of chronic tinnitus: A review. Clinical Psychology & Psychotherapy **29**(4), 1236–1253 (2022). https://doi.org/10.1002/cpp.2708

9. Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., Tesconi, M.: Hate me, hate me not: Hate speech detection on Facebook. In: Proceedings of the First Italian Conference on Cybersecurity (ITASEC17). vol. 1816, pp. 86–95 (2017)
10. Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., Bhamidipati, N.: Hate speech detection with comment embeddings. In: Proceedings of the 24th International Conference on World Wide Web. ACM (2015)
11. Fauziah, Y., Saifullah, S., Aribowo, A.S.: Design Text Mining for Anxiety Detection using Machine Learning based-on Social Media Data during COVID-19 pandemic. Proceeding of LPPM UPN "Veteran" Yogyakarta Conference Series 2020–Engineering and Science Series **1**(1), 253–261 (2020)
12. Georganos, S., Grippa, T., Vanhuysse, S., Lennert, M., Shimoni, M., Wolff, E.: Very high resolution object-based land use–land cover urban classification using extreme gradient boosting. IEEE Geoscience and Remote Sensing Letters **15**(4), 607–611 (2018). https://doi.org/10.1109/lgrs.2018.2803259
13. Giannakis, M., Dubey, R., Yan, S., Spanaki, K., Papadopoulos, T.: Social media and sensemaking patterns in new product development: demystifying the customer sentiment. Annals of Operations Research **308**(1-2), 145–175 (2020)
14. Gitari, N.D., Zhang, Z., Damien, H., Long, J.: A lexicon-based approach for hate speech detection. International Journal of Multimedia and Ubiquitous Engineering **10**(4), 215–230 (2015). https://doi.org/10.14257/ijmue.2015.10.4.21
15. Kléma, J., Almonayyes, A.: Automatic Categorization of Fanatic Text Using random Forests. Kuwait Journal of Science and Engineering **33**(2), 1–18 (2006)
16. Kumar, S., Yadava, M., Roy, P.P.: Fusion of EEG response and sentiment analysis of products review to predict customer satisfaction. Information Fusion **52**, 41–52 (2019). https://doi.org/10.1016/j.inffus.2018.11.001
17. Muñoz, S., Iglesias, C.A.: A text classification approach to detect psychological stress combining a lexicon-based feature framework with distributional representations. Information Processing & Management **59**(5), 103011 (2022)
18. Ni, M.Y., Yang, L., Leung, C.M.C., Li, N., Yao, X.I., Wang, Y., Leung, G.M., Cowling, B.J., Liao, Q.: 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), e19009 (2020). https://doi.org/10.2196/19009
19. Nijhawan, T., Attigeri, G., Ananthakrishna, T.: Stress detection using natural language processing and machine learning over social interactions. Journal of Big Data **9**(1) (2022). https://doi.org/10.1186/s40537-022-00575-6
20. Ragini, J.R., Anand, P.R., Bhaskar, V.: Big data analytics for disaster response and recovery through sentiment analysis. International Journal of Information Management **42**, 13–24 (2018). https://doi.org/10.1016/j.ijinfomgt.2018.05.004
21. Rezwanul, M., Ali, A., Rahman, A.: Sentiment analysis on twitter data using KNN and SVM. International Journal of Advanced Computer Science and Applications **8**(6) (2017). https://doi.org/10.14569/ijacsa.2017.080603
22. Saifullah, S., Fauziyah, Y., Aribowo, A.S.: Comparison of machine learning for sentiment analysis in detecting anxiety based on social media data. Jurnal Informatika **15**(1), 45 (2021). https://doi.org/10.26555/jifo.v15i1.a20111
23. Schmidt, A., Wiegand, M.: A survey on hate speech detection using natural language processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/w17-1101