Depression Early Warning System Based on Social Media Using Multinomial Naïve Bayes Algorithm

Achmad Maududie [1,a)], Putri Armaini [1,b)], Priza Pandunata [1,c)]

[1] *Faculty of Computer Science, Jember University Indonesia*

[a)] *maududie@unej.ac.id*
[b)] *192410101094@mail.unej.ac.id*
[c)] *priza@unej.ac.id*

Abstract. Community mental health is a serious challenge that must be encountered by a country. The results of the I-NAMHS (Indonesia-National Adolescent Mental Health Survey) show that one in three teenagers in Indonesia has mental health problems, while one in twenty teenagers in Indonesia has had a mental disorder in the last 12 months. Based on the results of a study from the Mayo Clinic in 2019, it was revealed that there is a phenomenon regarding the impact of social media on teenagers aged 13–16 years, where using social media more than three times a day can have a negative influence on mental health. This research aims to develop an approach to detecting depressive symptoms in social media users through messages delivered via Twitter. The crawling data result is labeled according to the Patient Health Questionnaire-9 (PHQ-9) questionnaire with three categories of depression levels (defined by the American Psychiatric Association), namely mild, moderate, and severe. The classification algorithm used is Multinomial Naive Bayes with Frequency-Relevance Frequency (TF-RF) as the weight determination and Min-Max Scaler as the normalization data. The evaluation results show that the resulting model has an accuracy value of 87.6%. Using this model, we detected from 76483 tweets that 2385 accounts were categorized as having mild depression, 5069 accounts were in moderate depression, and 3636 accounts were in severe depression. Therefore, in general, this approach can be applied as a depression early warning system for social media platform users.

Keywords: Social media and mental health, depression detection algorithm multinomial naive bayes classification

INTRODUCTION

Prevalent symptoms of depression include melancholy, poor mood, exhaustion, loss of interest in life, lack of motivation, helplessness, and despair. Depression is a prevalent mental health condition and societal problem. A person is said to be in good mental health if they can accept who they are, deal with social situations, feel happy about themselves, and have a realistic self-image. Mental health is as important as physical health [1]. One in three Indonesian teenagers has mental health problems, and one in twenty has experienced a mental disorder in the last 12 months, according to the findings of the I-NAMHS (Indonesia-National Adolescent Mental Health Survey) [2].

Users of the microblogging site Twitter often exchange messages in real time. With a maximum 140 characters in length for brief messages, sometimes known as tweets [3], using the Twitter platform is an easy way to communicate, exchange tales, voice your thoughts, and express opinions [4]. Even so, people frequently feel at ease sharing their thoughts, sentiments, and emotions. A 2019 study from the Mayo Clinic found that teens who use social media three times a day may have mental health problems and poor wellbeing [5]. With the huge amount of data on personal expressions that it has, Twitter can be a data source for capturing a picture of the level of depression that occurs in society, especially for Twitter users. Therefore, regarding mental health problems in Indonesia and the availability of data on Twitter, this research attempts to establish an approach to measuring the degree of depression of Twitter users through their messages as a depression early warning system.

METHOD

The approach used in this research consists of two main parts: creating a model to classify depression levels consisting of three depression classes (mild, moderate, and severe) and developing a depression early warning system using the depression level classification model that has been created.

I. H. Agustin (ed.), *Proceedings of the 2nd International Conference on Neural Networks and Machine Learning 2023 (ICNNML 2023)*, Advances in Intelligent Systems Research 183,
https://doi.org/10.2991/978-94-6463-445-7_8

Building a Depression-Level Classification Model

Preparing keywords for data crawling. The depression level classification model was formed from training data consisting of four groups of tweets (the results of Twitter crawling) representing four classes of depression. To obtain the training data, keywords were compiled for the crawling process based on the items of the questionnaire used to evaluate the presence of depressive symptoms from the Patient Health Questionnaire-9 (PHQ-9). This questionnaire has the potential to serve as a dual-purpose instrument that can screen for the presence of depressive disorders and assess the severity of symptoms [4]. As a note, all keywords are in Indonesian.

TABLE 1. The patient health questionnaire-9 (PHQ-9)

No	The Patient Health Questionnaire-9 (PHQ-9)
1	little interest or pleasure in doing things?
2	feeling down, depressed, or hopeless?
3	trouble falling or staying asleep, or sleeping too much?
4	feeling tired or having little energy?
5	poor appetite or overeating?
6	feeling bad about yourself or that you're a failure or have let yourself or your family down?
7	trouble concentrating on things, such as reading the newspaper or watching television?
8	moving or speaking so slowly that other people could have noticed? Or the opposite—being so fidgety or restless that you have been moving around a lot more than usual?
9	thoughts that you would be better off dead or of hurting yourself in some way?

Preparing the dataset for model building. The dataset for model creation was obtained through crawling Twitter data using specific keywords. From this dataset, a portion was selected to build the model. In this research, the aim was to gather approximately 800 tweets for each class.

Preprocessing Data. In the data preprocessing stage, some steps are used to transform raw text into structured linguistic components and standard notations [4]. This stage consists of seven steps, as follows:

- Case Folding: the process of processing data by equalizing all characters (letters) that will be used [4]. In this step, we transform all characters to low-ercase. In this study, the author used the lower() function to standardize and align the data.
- Emoji Conversion: the process of processing emojis into words because of emojis. Emojis in each comment need to be analyzed as they have an impact on sentiment analysis results [7] Python's demojize package is utilized. At this point, the emojis in the dataset are converted into string words and given meanings. This step assumes that the emoji library has already been installed and uses the emoji.core module that was imported from demojize.
- Slang Word Conversion: the process of translating a slang word into a formal word. Slang words are often used on Twitter that do not follow formal grammar rules. In this step, to translate slang word by using Indonesian slang dictionary from https://github.com/nasalsabila/kamus-alay.
- Cleansing: the process of removing unimportant characters or symbols to reduce noise in the classification process [4]. In this step, to remove unimportant characters, we use regular expressions. In order to remove characters, a string pattern containing the characters to be eliminated is passed into thesub function, which makes use of the re package library.
- Tokenizing: is the process of transforming sentences into tokens. In this case, all sentences break into collections of words. This tokenizing stage utilizes the word_tokenize function from the nltk.tokenize library. The results of tokenizing are stored in the tweet_token table.
- Stop Word Removal: is the process of eliminating stop words that do not represent meaningful content to optimize the data in the training data for classification. Stop words are words that commonly occur and do not hold significant meaning in the data analysis stage [4].

- Stemming: is the process of removing suffixes or prefixes from words to get their base form [4]. In this step, we use the Indonesian stemmer library from Sastrawi.

Term Weighting. Term weighting is a text mining approach that assesses a term's significance inside a document. In this research, we use the term frequency-relevance frequency (TF-RF) algorithm, where the frequency of occurrence of terms in related categories is considered through document relationships [6]. The weight of the term could be defined by the following equation.

$$tf_{td}rf = tf_{td} X \log log\,(2 \frac{b}{(1, c)}) \quad (1)$$

$tf_{td}rf$: Document weighting in vector space modeling
tf_{td} : Number of occurrences of term t in the document
b : Number of documents that contain term *t*
c : Number of documents that do not contain term *t*

Normalization using MinMaxScaler. MinMaxScaler is a method used in sentiment analysis to transform the range of values of numeric features into a smaller range, typically between 0 and 1. This method helps in data normalization and ensures that all features have a similar scale, so no feature dominates the sentiment analysis process. By using MinMaxScaler, data can be transformed into a smaller range, making it easier for comparison and further analysis [8]

$$Xnorm = \frac{X-(x)}{(x)-(x)} \quad (2)$$

$Xnorm$: The normalized result of a value x
X : The I-th value from the set of data to be normalized
(x) : The minimum value from the set of data
(x) : The maximum value from the set of data

Multinomial Naïve Bayes. Multinomial naive Bayes is one of the text mining methods in the process of text classification using the probability values of a class. The process begins by inputting training data, which can be used for testing data later on, to calculate the probability of a class occurring in the training data.

$$P(c) = \frac{N_c}{N_{doc}} \quad (2)$$

c : Category or class
doc : Document
Nc : Number of category c in the training documents
$Ndoc$: Total number of training documents used.

Furthermore, the probability that a word belongs to a specific category or class can be calculated using the following equation.

$$P\left(w_i, c\right) = \frac{count\left(w_i, c\right)+1}{\sum w\, count\left(w_i, c\right)+ |V|} \quad (3)$$

w : The i-th word in all documents categorized as c
$count\left(w_i, c\right)$: The number of occurrences of a specific word in a category or class
$\sum w\, count\left(w_i, c\right)$: The total count of all words in the class
$|V|$: The total number of unique words in the classThe normalized result of a value x

Building Depression Early Warning System

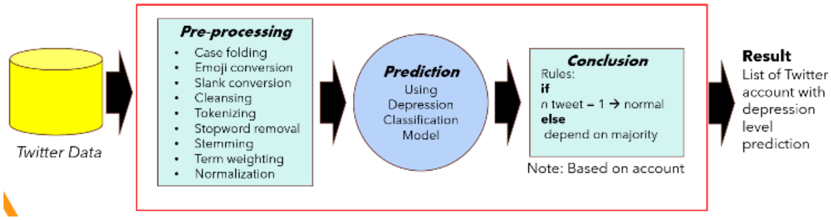

FIGURE 1. The purposed approach of depression early warning system

The process of creating a Depression Early Warning System begins with preparing the prediction dataset. This dataset will undergo several preprocessing steps, such as case folding, emoji conversion, slang conversion, cleansing, tokenizing, stopword removal, and stemming. Next, term weighting and normalization steps are performed to obtain normalized values for the data. Afterward, the previously built model is used to make predictions, allowing each tweet in the Twitter dataset to be predicted.

For the conclusion step, attention is given to each account. If an account has only one tweet, it is categorized as not experiencing depression. However, if an account has multiple tweets, the conclusion is based on the majority prediction. Thus, the account is categorized as experiencing depression with the severity level depending on the majority prediction.

Results and Discussion

Building Depression Early Warning System

Based on PHQ-9, the keywords for each item question could be shown in Table 2. A total of 12 keywords were used for the mild depression class, 13 keywords were used-for the moderate depression class, and 2 keywords were used for the severe depresion class. These guidelines are obtained from the PHQ-9 questionnaire, which divides each question and adjusts them based on the characteristics of depression and its respective levels.

TABLE 2. Labeling guidelines

No	PHQ-9 Questionnaire	Keywords	Class
1	little interest or pleasure in doing things?	malas, mager, malas gerak	mild depression
2	feeling down, depressed, or hopeless?	sedih, nangis, murung, putus asa, kecewa	mild depression
3	trouble falling or staying asleep, or sleeping too much?	insomnia, sulit tidur,	mild depression
4	feeling tired or having little energy?	lelah, lemah, lesu, cape, lunglai	moderate depression
5	poor appetite or overeating?	tidak nafsu	mild depression
6	feeling bad about yourself or that you're a failure or have let yourself or your family down?	Tidak percaya diri, malu, takut	moderate depression
7	trouble concentrating on things, such as reading the newspaper or watching television?	Ragu	moderate depression

No	PHQ-9 Questionnaire	Keywords	Class
8	moving or speaking so slowly that other people could have noticed? Or the opposite—being so fidgety or restless that you have been moving around a lot more than usual?	resah, gelisah, cemas, khawatir, ragu	moderate depression
9	thoughts that you would be better off dead or of hurting yourself in some way?	mati, selfharm	severe depression

Building Depression Classification

Multinomial naive Bayes is used to create depression classification models. The technique involves multiple stages, the first of which is preparation, which involves preprocessing, weighting, normalization, and modeling of the data.

Preprocessing Data. This study's data pretreatment step includes case folding, stopword removal, emoji preprocessing, unique word preprocessing, cleaning, tokenizing, and stemming.

TABLE 3. Case folding

Before	After
mau selfharm gaada silet mau pake cutter y mnimal kaga berkarat lah SOALNY GMW MATII cuma mau ngelampiasin doang	mau selfharm gaada silet mau pake cutter y mnimal kaga berkarat lah soalny gmw matii cuma mau ngelampiasin doang
Otak overheat tapi kalo dibawa tidur bawaannya gelisah dan ovt, pantes ya bnyk case org depresi gara2 skripshit 🤯	otak overheat tapi kalo dibawa tidur bawaannya gelisah dan ovt, pantes ya bnyk case org depresi gara2 skripshit 🤯
Alhamdulillah, sedikit cape aja tapi gpp https://t.co/iy3DKbBMzd	alhamdulillah, sedikit cape aja tapi gpp https://t.co/iy3dkbbmzd

the case folding process was successful. As seen in Table 4, the word "SOALNY GMW MATII" has been transformed into "soalny gmw matii" with all lowercase letters.

TABLE 4. Preprocessing emoji

Before	After
mau selfharm gaada silet mau pake cutter y mnimal kaga berkarat lah soalny gmw matii cuma mau ngelampiasin doang	mau selfharm gaada silet mau pake cutter y mnimal kaga berkarat lah soalny gmw matii cuma mau ngelampiasin doang
otak overheat tapi kalo dibawa tidur bawaannya gelisah dan ovt, pantes ya bnyk case org depresi gara2 skripshit 🤯	otak overheat tapi kalo dibawa tidur bawaannya gelisah dan ovt, pantes ya bnyk case org depresi gara2 skripshit
alhamdulillah, sedikit cape aja tapi gpp https://t.co/iy3dkbbmzd	alhamdulillah, sedikit cape aja tapi gpp https://t.co/iy3dkbbmzd

In Table 5, it is evident that the data in the second row successfully converted the emoji into words. The emoji has been transformed into ":exploding_head:".

TABLE 5. Preprocessing unique text data

Before	After
mau selfharm gaada silet mau pake cutter y mnimal kaga berkarat lah soalny gmw matii cuma mau ngelampiasin doang	mau selfharm enggak ada silet mau pakai cutter ya mnimal kagak berkarat lah soalny gmw mati cuma mau ngelampiasin doang
otak overheat tapi kalo dibawa tidur bawaannya gelisah dan ovt, pantes ya bnyk case org depresi gara2 skripshit	otak overheat tapi kalo dibawa tidur bawaannya gelisah dan ovt, pantes ya banyak case orang depresi gara gara skripshit
alhamdulillah, sedikit cape aja tapi gpp https://t.co/iy3dkbbmzd	Alhamdulillah, sedikit capek saja tapi enggak apa apa https://t.co/iy3dkbbmzd

the preprocessing step for slang data was successful, especially when using the available dictionary data. In Table 6, it is observed that some words have been successful transformed. For example, 'gaada' has been changed to 'enggak ada', 'pake' has beechanged to 'pakai', and 'y' has been changed to 'ya'.

TABLE 6. Cleansing

Before	After
mau selfharm enggak ada silet mau pakai cutter ya mnimal kagak berkarat lah soalny gmw mati cuma mau ngelampiasin doang	mau selfharm enggak ada silet mau pakai cutter ya mnimal kaga berkarat lah soalny gmw matii cuma mau ngelampiasin doang
otak overheat tapi kalo dibawa tidur bawaannya gelisah dan ovt, pantes ya banyak case orang depresi gara gara skripshit	otak overheat tapi kalo dibawa tidur bawaannya gelisah dan ovt pantes ya bnyk case org depresi gara skripshit
alhamdulillah, sedikit capek saja tapi enggak apa apa https://t.co/iy3DKbBMzd	alhamdulillah sedikit capek aja tapi tidak apa apa

The cleansing step was successful, for example, in removing punctuation marks and links.

TABLE 7. Tokenizing

Before	After
mau selfharm enggak ada silet mau pakai cutter ya mnimal kaga berkarat lah soalny gmw matii cuma mau ngelampiasin doang	‘mau’, ‘selfharm’, ‘enggak’, ‘ada’, ‘silet’, ‘mau’, ‘pake’, ‘cutter’, ‘ya’, ‘mnimal’, ‘kagak’, ‘berkarat’, ‘lah’, ‘soalny’, ‘gmw, ‘mati’, ‘cuma’, ‘mau’, ‘ngelampiasin’, ‘doang’
otak overheat tapi kalo dibawa tidur bawaannya gelisah dan ovt pantes ya bnyk case org depresi gara skripshit	‘otak’, ‘overheat’, ‘tapi’, ‘kalo’, ‘dibawa’, ‘tidur’, ‘bawaannya’, ‘gelisah’, ‘dan’, ‘ovt’, ‘pantes’, ‘ya’, ‘bnyk’, ‘case’, ‘orang’, ‘depresi’, ‘gara’, ‘gara’, ‘skripshit’
alhamdulillah sedikit capek aja tapi tidak apa apa	‘alhamdulillah’, ‘sedikit’, ‘capek’, ‘aja’, ‘tapi’, ‘tidak’, ‘apa’, ‘apa’

Tokenizing step was successful in converting sentences into word tokens.

TABLE 8. Stopword Removal

Before	After
'mau', 'selfharm', 'enggak', 'ada', 'silet', 'mau', 'pakai', 'cutter', 'ya', 'mnimal', 'kagak', 'berkarat', 'lah', 'soalny', 'gmw, 'mati', 'cuma', 'mau', 'ngelampiasin', 'doang'	'selfharm', 'silet', 'pakai', 'cutter', 'ya', 'mnimal', 'kagak', 'berkarat', 'soalny', 'gmw, 'mati', 'ngelampiasin', 'doang'
'otak', 'overheat', 'tapi', 'kalo', 'dibawa', 'tidur', 'bawaannya', 'gelisah', 'dan', 'ovt', 'pantes', 'ya', 'bnyk', 'case', 'orang', 'depresi', 'gara', 'gara', 'skripshit'	'otak', 'overheat', 'kalo', 'dibawa', 'tidur', 'bawaannya', 'gelisah', 'ovt', 'ya', 'case', 'orang', 'depresi', 'gara', 'gara', 'skripshit'
'alhamdulillah', 'sedikit', 'capek', 'aja', 'tapi', 'tidak', 'apa', 'apa'	'alhamdulillah', 'capek'

The stopword removal step was successful, as it removed certain words such as "mau," "enggak," "ada," "lah," "cuma," "tapi," "dan," "pantes," and "ya.".

TABLE 9. Stemming

Before	After
'selfharm', 'silet', 'pakai', 'cutter', 'ya', 'mnimal', 'kagak', 'berkarat', 'soalny', 'gmw, 'mati', 'ngelampiasin', 'doang'	'selfharm', 'silet', 'pakai', 'cutter', 'ya', 'mnimal', 'kagak', 'karat', 'soalny', 'gmw', 'mati', 'ngelampiasin', 'doang'
'otak', 'overheat', 'kalo', 'dibawa', 'tidur', 'bawaannya', 'gelisah', 'ovt', 'ya', 'case', 'orang', 'depresi', 'gara', 'gara', 'skripshit'	'otak', 'overheat', 'kalo', 'bawa', 'tidur', 'bawa, 'gelisah', 'ovt', 'pantes', 'ya', 'banyak', 'case', 'orang', 'depresi', 'gara', 'skripshit'
'alhamdulillah', 'capek'	'alhamdulillah', 'capek'

The stemming step was successful in converting words like 'berkarat' to its base form 'karat', 'dibawa' to 'bawa', and 'bawaannya' to 'bawa'.

Term Frequency - Relevance Frequency (TF-RF). Term Frequency – Relevance Frequency (TF-RF) weighting is a method where the frequency of term occurrence in related categories is considered through document relationships. Table 11 shows the value of TF which represents the occurrence of each word in each document while Table 12 shows the value of RF.

TABLE 10. TF value

Word	TF Value	
['selfharm', 'silet', 'pakai', 'cutter', …]	selfharm	0.071428
	silet	0.071428
	pakai	0.071428
	cutter	0.071428
['otak', 'overheat', …]	otak	0.066667
	overheat	0.066667
['alhamdulillah', 'capek']	alhamdulillah	0.5
	capek	0.5

TABLE 11. RF value

RF Value	
selfharm	0.336365

RF Value	
silet	0.301358
pakai	0.305419
cutter	0.301113
otak	0.30201
overheat	0.30111
alhamdulillah	0.30234
capek	0.318589

Then, the calculation of the TF-RF value will be performed by multiplying the TF value and the RF value.

TABLE 12. TF-RF value

TF-RF Value	
Selfharm	0.0240259
Silet	0.0215254
Pakai	0.0218155
Cutter	0.0215079
Otak	0.0201341
Overheat	0.0200741
alhamdulillah	0.1511702
Capek	0.1592946

MinMaxScaler. This research uses MinMaxScaler for normalization. The normalization stage utilizes MinMaxScaler with a range of [0,1]. This research utilizes the sklearn.preprocessing library by importing MinMaxScaler.

Multinomial Naïve bayes. During the model generation process, the data is classified into three categories: mild depression, moderate depression, and severe depression using Multinomial Naïve Bayes. The train_test_split function from the sklearn library is used in this study to split the data, with 90% going to training and 10% to testing. Next, multinomialNB() and the sklearn.naive_bayes module are used to generate the model. The variable "predicted" contains the modeling findings. Consequently, the model's accuracy comes out to be 87.6%.

Implementation Model into Depression Early Warning

The prediction stage uses a dataset of 76,483 to determine predictions based on the previously created model. In the depression prediction stage, it starts with performing preprocessing similar to the previous preprocessing stage for model creation. This stage can be time-consuming. Next, the weighting stage using TF-RF and normalization using MinMaxScaler is conducted. After obtaining the vector values from weighting, predictions are made using the created model. As a result, predictions are generated based on the available tweet data. The predicted results will classify the tweet data into classes 1, 2, and 3, representing mild depression, moderate depression, and severe depression.

```
                                                  tweet  prediksi
0            and he still pulled the capek out three times         2
1           capek capek war nyantol calo nya luar nalar ceee         2
2       live near capek verdean diaspora their descend...         2
3                                      capek malam ngakakin         3
4                         capek banget tonton nappeun eomma         3
...                                                   ...       ...
76479                               relax yuk biar selfharm         2
76480             mati lampu hujan tau mentioning of selfharm         2
76481                   menang tidur kadang suka selfharm sih         1
76482   nyokap tinggal sbmptn kayak kenal selfharm dll...         3
76483   mintak balas orang kecewa kecewa perhati jaga ...         3

[76484 rows x 2 columns]
```

FIGURE 2. Prediction depression

Based on figure 2, The predictions made by the model aim to determine the level of depression indicated by the sentence. In the first data, there is an indication of mild depression. In the second data, the sentence suggests moderate depression. And in the third data, the sentence indicates severe depression. This pattern continues for a total of 76,483 data points.

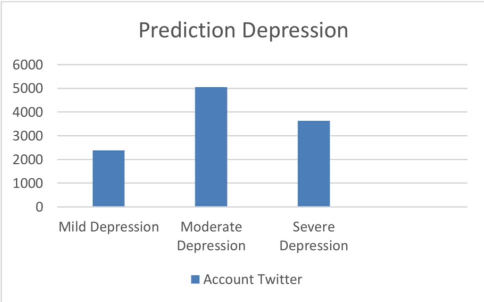

FIGURE 3. Bar chart prediction depression

Based on figure 3, using this model multinomial naïve bayes, we detected from 76483 tweets that 2385 accounts were categorized as having mild depression, 5069 accounts were in moderate depression, and 3636 accounts were in severe depression.

Conclusion

We have developed a Depression Early Warning System Based on social media with relatively high accuracy of model prediction, i.e., 87.6%. The system is able to detect Twitter accounts indicated to have symptoms of depression, either mild, moderate, or severe.

Acknowledgment

This research was supported by University of Jember. I thank all those who havehelped my research.

References

1. F. V. Salinas, (2022), "Contemporary Anesthetic Evaluation and Management for Electroconvulsive Therapy," *Adv Anesth*, vol. 40, no. 1, pp. 201–221, Dec. 2022, doi: 10.1016/j.aan.2022.07.005.
2. "I-NAMHS: *Indonesia-National Adolescent Mental Health Survey*."
3. D. Isari, A. Pontiggia, and F. Virili, (2016), "Working with tweets vs. working with chats: An experiment on collaborative problem solving," *Comput Human Behav*, vol. 58, pp. 130–140, May 2016, doi: 10.1016/j.chb.2015.12.052.
4. D. Sarkar, *Text Analytics with Python : A Practical Real-World Approach to Gaining Actionable. Berkeley*, CA: Apress, 2016. doi: 10.1007/978-1-4842- 2388-8.
5. R. M. Viner et al., (2019), "Roles of cyberbullying, sleep, and physical activity in mediating the effects of social media use on mental health and wellbeing among young people in England: a secondary analysis of longitudinal data," *Lancet Child Adolesc Health*, vol. 3, no. 10, pp. 685–696, Oct. 2019, doi: 10.1016/S2352-4642(19)30186-5.
6. G. Plemmons, (2019), "Depression and Suicide Screening," in *Adolescent Health Screening: an Update in the Age of Big Data, Elsevier*, pp. 135–149. doi: 10.1016/B978-0-323-66130-0.00011-9.
7. C. Liu et al., (2021), "Improving sentiment analysis accuracy with emoji embedding," *Journal of Safety Science and Resilience*, vol. 2, no. 4, pp. 246–252, Dec. 2021, doi: 10.1016/j.jnlssr.2021.10.003.

8. M. D. Devika, C. Sunitha, and A. Ganesh, (2016), "Sentiment Analysis: A Compara- tive Study on Different Approaches," *Procedia Comput Sci*, vol. 87, pp. 44– 49, doi: 10.1016/j.procs.2016.05.124.