JINAV: Journal of Information and Visualization *ISSN: 2746-1440 (online)*
Vol. 4 No. 1 (2023)
https://doi.org/10.35877/454RI.jinav1543

Analyzing *Cyberbullying* Negative Content on Twitter Social Media With The RoBERTa Method

Muh Akib A Yani* & Warih Maharani

Faculty of Informatics, Telkom University, Bandung, Indonesia

Abstract

Social media is an online-based communication tool that can make it easier for users to interact among fellow users without any area and time restrictions. Indonesia has the highest number of social media users in the world. The Twitter social media platform is a place where users can pour out their whole hearts in the form of tweets. Free and diverse interactions on Twitter have a considerable influence on the psychological condition of its users. Cyberbullying or online bullying is an act of humiliating or hurting other people's feelings intentionally and repeatedly on social media, messages, or in other ways. In this study, the RoBERTa classification method was used to detect cyberbullying tweets with a best accuracy score of 86.9% and an f1-score of 77.5%.

Keywords: Social Media; Twitter; Cyberbullying; RoBERTa.

1. Introduction

1.1. Background

Social media is an online-based means of communication that can make it easier for users to interact between fellow users without any regional and time restrictions. Indonesia has the highest number of social media users in the world. The Ministry of Communication and Informatics (Kemenkominfo) states that 95% of the 63 million internet users are social users (Fadli & Hidayatullah, 2021). One of the social media that is often used by users is Twitter, *Country Industry Head* Twitter Indonesia claims that Indonesia is the country with the highest active growth of daily users of Twitter social media (Fadli & Hidayatullah, 2021).

But not all users can use Twitter social media wisely. Not a few users use Twitter social media to take actions that can harm other users such as the spread of hoax news, fraud, to online harassment or *bullying (cyberbullying*) this problem arises along with the development of social media recently. *Cyberbullying* can be interpreted as the use of social media by an individual or group of users to harass other users which can result in other users feeling negative effects on the victim. One study by a national anti-bullying charity showed that two out of three children aged 13-22 surveyed had been victims of *cyberbullying* (Saravanaraj et al., n.d.).

Therefore, this study aims to detect *cyberbullying* on the Twitter *platform*. The benchmark in this study is based on the reciprocity and interaction of user tweets on Twitter. Previously, research conducted by Yinhan Liu regarding the expansion of the BERT architecture that had been developed and named RoBERTa, the research was carried out as a development of the BERT architecture which showed results in being significantly less trained compared to later models. Therefore, the method that will be used in this study is RoBERTa where RoBERTa is a retraining of BERT with an improved training methodology, allowing receiving more data, and computing power. Additionally, this method eliminates *Next Sentence Prediction* (NSP) on BERT and uses *Dynamic Masking*. With several advantages possessed by RoBERTa, making this method reliable with improved performance compared to the BERT method (Liu et al., 2019).

* Corresponding author.
E-mail address: muhammadakib@student.telkomuniversity.ac.id

1.2. Topics and Limitations

This study discusses the detection of Indonesian tweets containing *cyberbullying* with the keywords "fat", "eat", "government", "tadpole", "anj*ng", "tol*l" and some words that contain harsh words and use animal swear words through social media twitter with the BERT classification method that has been developed, namely RoBERTa.

1.3. Purpose

The purpose of this study was to determine the performance of the RoBERTa classification method by looking for the best accuracy value against Indonesian people's ciutan on twitter social media that contains *cyberbullying*.

1.4. Writing Organization

In this research report, related and related literature studies will be discussed as well as studies that are the reference for research in Chapter 2. Then the system built into this study can be seen in Chapter 3. The results of the research that has been carried out are listed in Chapter 4. As well as the conclusions of the study can be seen in Chapter 5.

2. Literature Review

Social media is an online-based communication platform that allows users to easily interact with each other without any restrictions on distance, time, and place. The positive impact of social media is that it makes it easier for users to interact with many people, expanding associations, distance and time is no longer a problem, it is easier to express themselves, information dissemination can take place quickly, costs less (Istiani & Islamy, 2020). Twitter is one of the microblogging developed by Twitter, Inc. It is called microblogging because this platform allows its users to send and read messages like blogs in general. The tweet is called a tweet, which is text that has a limit of 140 characters published on a user's profile page (Anggreini, 2016).

Cyberbullying or online bullying is an act of intentionally and repeatedly insulting or hurting the feelings of others on social media, messages, or in other ways. Cyberbullying activities on the social media platform Twitter are carried out by publishing tweets that contain harsh words or insulting words, to words that suggest insults to SARA (Fadli & Hidayatullah, 2021).

Research on *cyberbullying* detection has been carried out with various classification methods. A.Saravanaraj, J *et al*[2] conducted research on *cyberbullying* detection on Twitter social media using the *Naïve Bayes* and *Random Forest* classification methods. The stages of its research activities consist of data collection, pre-processing, classification, and evaluation. This study resulted in an accuracy value of 90%.

Nassharih Abdulloh *et al* (Abdulloh & Hidayatullah, 2019) conducted research on *cyberbullying* detection on Twitter social media tweets using 4 classification methods, namely *Multinomial Naïve Bayes*, *Linear SVM*, *Logistic Regression*, and *KNN* . The stages of its research activities consist of data collection, *preprocessing*, feature extraction, classification, and evaluation. This study resulted in accuracy values of 0.961 for the *Naïve Bayes Multinomial* algorithm, 0.994 for the *Logistic Regression* algorithm, 0.997 for the *Linear SVM* algorithm, and 0.918 for the *KNN* algorithm .

Ketsbaia & Chen conducted research on *cyberbullying* detection using two *datasets* sourced from Twitter social media by giving 3 labels namely "*hate speech*", "*offensive*", " *neither hate-speech or offensive*" and uses 3 classification methods namely *RoBERTa*, *XLNet* and *DistilBERT*. The stages of research activities consist of data collection, *preprocessing*, *tf-idf* features, *dataset* balancing techniques, namely *SMOTE* and *Random Under Sampling*, selection features, *Logistic Regression*, *Particle Swarm Optimisation*, *Genetic Algorithm*, classification method, and evaluation. And the results were obtained from the accuracy for the first *dataset*, namely 86.78% for the *DistilBBERT* algorithm, 87.78% for the *XLNet* algorithm, 86.43% for the *RoBERTa* algorithm. As for the second *dataset*, an accuracy value of 94% was obtained for the *DistilBBERT* algorithm, 94.47% for the *XLNet* algorithm, 93.89 % for the *RoBERTa* algorithm.

3. Methods

In this study, the system built was able to detect *cyberbullying* on twitter tweets. There are several stages to detect, namely, *crawling*, data *labelling*, data *preprocessing*, *data splitting* and is divided into two, namely dataset training and testing dataset where training dataset is used to train the algorithm to be used while *testing dataset* used to determine the performance of pre-trained algorithms, RoBERTa classification, and evaluation. The scheme in this study can be seen in figure 1.

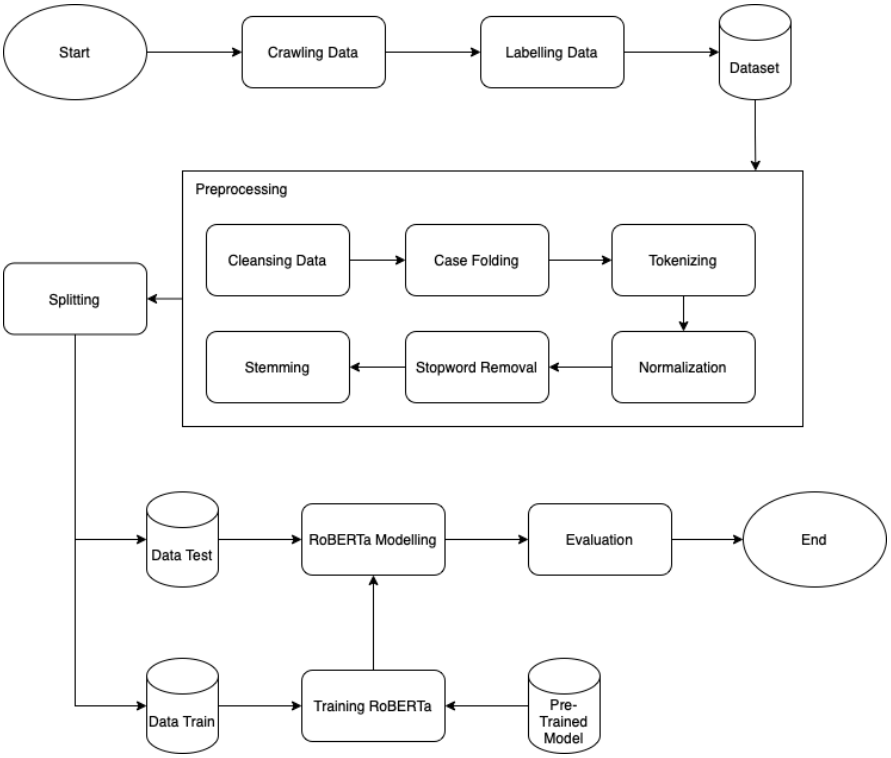

Figure 1. System Design Flowchart

3.1. Data Crawling

In this research, the dataset is data taken based on Indonesian people's tweets on twitter social media with the keywords "fat", "eat", "government", "tadpole", "anj*ng", "tol*l" and some words that contain harsh words and use animal swear words.

3.2. Data Labelling

After *crawling* the data , *the dataset* that has been collected will go through a manual data labeling process where the data is divided into 2 labels, namely positive and negative. Labeling is done by 3 people with the aim of reducing subjectivity in labeling. Labeling is done by paying attention to the words contained in the crawled tweets data, if there are harsh words in the *tweets* data or that contain *cyberbullying* is then given a negative label or 1, if the data does not contain harsh words or does not contain *cyberbullying* then it is given a positive label or 0.

Cyberbullying or harsh words have the meaning of swearing, slurs, spelling, or words that contain animal swear words.

Table 1. Data Labeling Example

Label	Tweets
Positive	kemarennya again said to be fat with nyokap family
Negative	people like fadli zonk , should be issued dr parliament, work g because, hobby bacot and nyinyir, just love to eat doang, tuh sampe fat, his brain shrinks

3.3. BERT

BERT is a classification method that is a language representation model designed to train bidirectional representation of unlabeled text by adjusting left and proper contexts in all layers. Two Way Encoder Representation from Transformers (BERT) optimizes Masked Language Model (MLM) and Next Sentence Prediction (NSP) in Pre-Trained processes. The Masked Model Language (MLM) is a process for predicting the words appearing from previous comments. Next Sentence Prediction (NSP) is a loss of binary classification which functions to bring up two words that follow each other in a text. BERT is the first finetuning-based representational model that outperforms many architectures. The BERT mode training analysis explores and calculates the important options for training the BERT model while maintaining the architectural mode. This starts from training the BERT model with the same configuration as the base BERT (training²L = 12, H= 768, A = 12, params 110M).

3.4. RoBERTa

Robustly optimized BERT approach (RoBERTa) is a replication version of the Pre-Trained approach from BERT, which has been optimized and can detect text that has no notation with a maximum amount of 160GB. RoBERTa removed Next Sentence Prediction (NSP) and added dynamic word hiding during the training period. These changes and features identify performance improvements compared to BERT in many NLP tasks, including text classification. Thus, RoBERTa uses more data to train, increases batch size, eliminates prediction of subsequent losses, and replaces static masking with dynamic masking in the pre-training stage to further improve performance. RoBERTa will give more optimal results than BERT because of these modifications. This has been proven by research using the architecture of BERT large L=24, H=1024, A=16, 355M parameters.

3.5. Pre-processing

After data collection is carried out, the next step is to change the data. Here are some steps in changing data:

1) Cleansing Data

*Cleansing Da*ta is the process of cleaning or removing unnecessary things such as punctuation, numbers, URLs, and words that are considered unimportant [5]. Examples of data that have passed the cleansing process can be seen in table 2.

Table 2. Data Cleansing

Tweets	***Cleansing***
Memwill be many but not fat : Kuli Indon https://t.co/4bLEdvTvEc	Eats a lot but is not fat coolie Indon

2) Case Folding

In this process words that have *an uppercase or uppercase* are changed to *lowercase* or lowercase. This process is done to avoid duplication that is distinguished from *case-sensitive*. Data that has passed through *case folding* will look like in tabel 3.

Table 3. Case Folding

Cleansing	***Case Folding***
Eats a lot but is not fat coolie Indon	eats a lot but not fat coolie indon

3) Tokenizing

Tokenization is the process of dividing or breaking a sentence that was previously separated by spaces into words that compose it. *This tokenization* needs to be done to facilitate classification. An example of tokenized data is found in table 4.

Table 4. Tokenizing

Case Folding	***Tokenizing***
eats a lot but not fat coolie indon	['eat', 'many', 'but', 'no', 'fat', 'coolie', 'indon']

4) Normalization

At the normalization stage, shortened words, incorrect words in writing, and informal words are changed to standard words and according to the writing of the KBBI. The dictionary used is a dictionary of kbbi words that has been created before. An example *of normalization* can be seen in table 5.

Table 5. Normalization

Tokenizing	***Normalization***
['eat', 'many', 'but', 'no', 'fat', 'coolie', 'indon']	['eat', 'many', 'but', 'no, 'fat', 'coolie', 'indon']

5) Stopword Removal

Stopword Removal is a stage of removing conjunctions that often appear. These words usually have a function but have no meaningful meaning and do not give weight to an opinion or sentence. An example of stopword removal can be seen in table 6.

Table 6. Stopword Removal

Normalization	***Stopword Removal***
['eat', 'many', 'but', 'no, 'fat', 'coolie', 'indon']	['eating', 'a lot', 'but', 'no, 'fat']

6) Stemming

Stemming is the stage of clearing affixes that includes prefixes, suffixes or a combination of the two. With stemming, words that have the same base word will be considered to have the same token. This helps in improving data processing performance. The results of the stemming data can be seen in table 7.

Table 7. Stemming

Stopword Removal	***Stemming***
['eating', 'a lot', 'but', 'no, 'fat']	['eat', 'many', 'but', 'no, 'fat']

3.6. RoBERTa Tokenize

After the dataset is divided into *train data* and *test data*, encoding is then carried out for each *dataset*. A *pre-trained* model on a RoBERTa model is required to perform an *encoding* on the *dataset*. In this study, *a pre-trained* RoBERTa model from ayameRushia was used which used *masked language modeling*. This model was previously trained using the Indonesian Wikipedia and the inputs of this model are a kat, sentence, and paragraph. In the *encoding* process, there are three inputs that will be processed in the model, namely *input_ids*, *attention_mask*, and *token_type*_ids. At this stage the *dataset* will be *tokenized* in the RoBERTa *modeling* process. This process will be adapted to the RoBERTa *pre-trained* model where the model will detect sentences from the *dataset*. The words present in the *dataset* will be combined into one when the *pre-trained* model can understand the sentence and will be separated when it cannot understand the sentence that is on the *dataset* (Toraman et al., 2022). You can see the example in table 8.

Table 8. Tokenize RoBERTa

Stages	**Result**
Dataset	come on diet again so you don't say add fat
Tokenize RoBERTa	['ay', 'oo', 'Ġdiet', 'Ġlagi', 'Ġbiar', 'Ġga', k', 'Ġdibilang', 'Ġn', 'ambah', 'Ġgend', 'ut']

3.7. Confusion Matrix

Furthermore, the last stage is to evaluate the stages that have been processed previously. At this stage it is necessary to be able to test the degree of accuracy of the previous method. To get the evaluation results, measurements are needed with the Confusion Matrix method which has 4 characteristics, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The performance of a matrix is measured based on *accuracy*, *precision*, *recall*, and *f1-scor*e which can be tested based on TP, TN, FP, and FN (Singh et al., 2021). The following is an example of calculating the evaluation value:

1) *Accuracy*
Accuracy presents the ratio to be classified by the formula :

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

2) *Precision*
Precision is the value obtained from the accuracy of a class with the total number of predictions for that class. The purpose of precision is to see the percentage of relevance of the classification results with the formula :

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

3) *Recall*
Recall is a value obtained from the accuracy of a class's predictions with the total number of facts for that class. Recall can be calculated by the formula:

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

4) *F1-Score*
F1-S core is an evaluation calculation performed by combining both precision and recall values. F1-S core can be calculated by the formula:

$$F1 = \frac{2.(Recall.Precision)}{Recall + Precision} \quad (4)$$

4. Result and Discussion

4.1. Data

The data used in this study amounted to 3034 Indonesian tweets with the keywords "fat", "eat", "government", "tadpole", "an j*ng", "tol*l" and several words containing harsh words and using animal swear words that had been *crawled* previously using netlytic websites. After that, *preprocessing* is carried out to filter the data to be tested and after obtaining the final *dataset* of 2135 tweets

Table 9. Data Distribution

Label	Amount of Data
Positive	1436
Negative	699

Labeling is done based on tweets. Label 1 for data that contains cyberbullying or negative and 0 for data that does not contain *cyberbullying* or is positive.

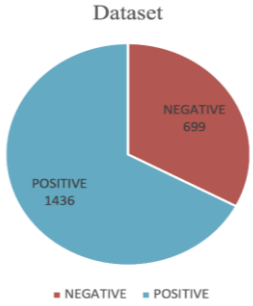

Figure 2. Data distribution

4.2. Scenarios and Test Results

In this study, 2 preprocessing comparison scenarios were carried out, namely *Preprocessing* which ends in the *case folding* or incomplete part and *preprocessing* to the final stage or *stemming*. After two preprocessing scenarios were carried out, it was continued until the classification stage by conducting 5 comparisons of train datasets and *test datasets* and the following results were obtained:

Table 10. Comparison of preprocessing and full preprocessing on the data train

Preprocessing (Data Cleansing, Case Folding)		***Full Preprocessing***	
Data Train	***Accuracy***	***Data Train***	***Accuracy***
90%	**86.92%**	90%	**82.24%**
80%	83.23%	80%	81.97%
70%	84.09%	70%	80.81%
60%	83.14%	60%	80.44%
50%	81.27%	50%	77.06%

After 2 *preprocessing* comparison scenarios obtained the highest accuracy value on the data that has been trained, namely on the *data train* of 90%, after getting the highest accuracy value, then testing was carried out with 2 scenarios, namely:

1) Performance testing using *preprocessing* (*data cleansing, case folding)* with 90% *data train*
2) Performance testing using *full preprocessing* with *a data train* of 90%

4.2.1. First Test Scenario

The first scenario was carried out with incomplete *preprocessing* or only *data cleansing* and *case folding* using a 90% *data train* by testing *batch sizes* of 8, 16, and 32 by conducting 5 experiments. The results can be seen from tables 11, 12, and13.

Table 11. Preprocessing batch size 8

EPOCH	***Accuracy***
1	82.7%
2	84.1%
3	84.5%
4	**86.9%**
5	83.6%

Table 12. Preprocessing batch size 16

EPOCH	***Accuracy***
1	85.1%
2	85.5%
3	**86.9%**
4	80.8%
5	85.1%

Table 13. Preprocessing batch size 32

EPOCH	***Accuracy***
1	80.8%
2	83.1%
3	**85.1%**
4	83.1%
5	83.6%

4.2.2. Second Test Scenario

The first scenario was performed with *full preprocessing* using a 90% *data train* by testing *batch sizes* of 8, 16, and 32 by conducting 5 attempts. The results can be seen from tables 14, 15, and 16.

Table 14. Full preprocessing batch size 8

EPOCH	***Accuracy***
1	78.0%
2	70.5%
3	**85.1%**
4	77.5%
5	82.7%

Table 15. Full preprocessing batch size 16

EPOCH	***Accuracy***
1	81.7%
2	83.6%
3	83.1%
4	84.5%
5	**85.9%**

Table 16. Full preprocessing batch size 32

EPOCH	***Accuracy***
1	80.8%
2	83.2%
3	**85.1%**
4	83.2%
5	83.6%

4.3. Analysis of Test Result

Table 17. Accuracy comparison

	Accuracy	***F1-Score***
***Preprocessing batch size* 8**	86.9%	76.3%
***Preprocessing batch size* 16**	86.9%	77.5%
***Preprocessing batch size* 32**	85.1%	71.6%
***Full preprocessing batch size* 8**	85.1%	75.3%
***Full preprocessing batch size* 16**	85.9%	75.7%
***Full preprocessing batch seze* 32**	85.1%	70.1%

There is a comparison of performance scores in table 17, it can be found that the *batch size* of 16 in tests using preprocessing and *full preprocessing* gets an accuracy value and *an f1-score* of 86.9% accuracy value and 77.5% *f* 1-score for preprocessing while for *full preprocessing* get a score of 85.9% accuracy and 75.7% *f1-score*. It can be ascertained that the batch size of 16 has a higher score than *the batch size* of 8 and 32.

Based on the test results of the two scenarios that have been carried out, it was found that the *datasets* were processed through the *preprocessing* stage without going through the *normalization*, *stopwrod removal* , and *stemming* stages get a higher accuracy value than *datasets* that are processed in *full preprocessing*, because the *dataset* is tested without word deletion which is considered non-standard and without through the process of clearing affix words can increase the detection rate on the RoBERTa classification method. The ratio of the comparison of the trained dataset to the untrained dataset affects the accuracy value, the more the *dataset* is trained the higher the accuracy obtained . And the batch size in the test affects the accuracy value obtained by decreasing the size of the *batch size*, the possibility that the level of accuracy will be as high as it is obtained because The algorithm converges faster but will generate *noise* on larger computations.

5. Conclusion

This research was conducted to detect Indonesian people's opinions regarding content on Twitter social media with the keywords "fat", "eat", "government", "tadpole", "dog", "tolol" and some words that contain harsh words and use animal swear words. In this study, two comparisons of preprocessing (data cleansing and case folding) and *full preprocessing* (*data cleansing*, *case folding*, *tokenizing*, *normalization*, *stopword removal* , and *stemming*) based on the two tests, the accuracy value of preprocessing was higher than using *full preprocessing*. In the RoBERTa model greatly affects the sentences in the *dataset* so that if there are words deleted in a sentence it will affect the detection of the RoBERTa model which causes the value of its accuracy is reduced.

In the next study, testing can be carried out using several different classification methods by comparing with the RoBERTa classification method in order to obtain different accuracy values from each method.

References

Abdulloh, N., & Hidayatullah, A. F. (2019). Deteksi Cyberbullying pada Cuitan Media Sosial Twitter. *Automata*, *Vol 1*(1), 1–5.

Anggreini, N. M. (2016). Pemanfaatan Media Sosial Twitter di Kalangan Pelajar SMK Negeri 5 Samarinda. *EJournal Sosiatri-Sosiologi*, *4*(2), 239–251.

Fadli, H., & Hidayatullah, A. (2021). Identifikasi Cyberbullying pada Media Sosial Twitter Menggunakan Metode LSTM dan BiLSTM. *Universitas Islam Indonesia (UII)*, *2*(No. 1), 1–6. https://journal.uii.ac.id/AUTOMATA/article/view/17364

Istiani, N., & Islamy, A. (2020). Pengaruh Media Sosial Terhadap Perubahan Sosial Masyarakat di Indonesia PENGARUH. *Asy Syar'Iyyah: Jurnal Ilmu Syari'Ah Dan Perbankan Islam*, *5*(2), 202–225.

Ketsbaia, L., & Chen, X. (n.d.). *Evaluation of Cyberbullying using Optimized Multi-Stage ML Framework and NLP*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. http://arxiv.org/abs/1907.11692

Saravanaraj, A., Sheeba, J. I., & Devaneyan, S. P. (n.d.). Automatic Detection of Cyberbullying From Twitter. *IRACST-International Journal of Computer Science and Information Technology & Security (IJCSITS)*, *6*(6), 2249–9555. https://www.researchgate.net/publication/333320174

Singh, P., Singh, N., Singh, K. K., & Singh, A. (2021). Diagnosing of disease using machine learning. *Machine Learning and the Internet of Medical Things in Healthcare*, 89–111. https://doi.org/10.1016/B978-0-12-821229-5.00003-3

Toraman, C., Yilmaz, E. H., Şahinuç, F., & Ozcelik, O. (2022). *Impact of Tokenization on Language Models: An Analysis for Turkish*. http://arxiv.org/abs/2204.08832

Pericherla, S., & Ilavarasan, E. (2021). Cyberbullying detection on multi-modal data using pre-trained deep learning architectures. https://revistas.ucc.edu.co/index.php/in/article/view/4001

Wang, X., Chen, S., Li, T., Li, W., Zhou, Y., & Zheng, J. (2020). Depression Risk Prediction for Chinese Microblogs via Deep-Learning Methods : Content Analysis.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding.