Feature Extraction for Sentiment Analysis in Indonesian Twitter

Eka Dyar Wahyuni*, Amalia Anjani Arifiyanti, Mohamad Irwan Afandi

Information System Department, Faculty of Computer Science, Universitas Pembangunan Nasional “Veteran” Jawa Timur, Indonesia

*Corresponding author: E-mail: ekawahyuni.si@upnjatim.ac.id

ABSTRACT

Twitter’s sentiment analysis is one of the most interesting fields of research lately. It intertwines the natural language processing techniques with data mining. Up to this point, many algorithms have been proposed to better understand sentiment from text. The proposed method can be focused on the preprocessing step, dataset splitting method (training and testing), dataset balancing method (when the data is unbalanced), to the improvement of the existing algorithm. But, the main focus of this paper is on feature extraction from tweets using TF-IDF. The features obtained from this process are expected to improve the accuracy of the classification process. The dataset used in this research is in Indonesian, which has a very different form when compared to English. This dataset consists of 1068 manually labeled tweets related to the "school from home" policy caused by the COVID-19 outbreak, taken from March to July. All steps required to process this data will be implemented using python. To validate its utility, the performance of the proposed method is compared with each other. Finally, the results are summarized by reflecting on the impact of the inclusion of the proposed features for each classification algorithm for sentiment detection

Keywords: Feature extraction, TF-IDF, Indonesian Tweet

Introduction

The coronavirus outbreak in Wuhan at the end of 2019 has spread to other countries, including Indonesia, resulting in the infection scale turning into a pandemic. Because until now, the vaccine for this virus is still being developed, the government has implemented a lockdown, quarantine, and social distancing policy as one of the countermeasures to reduce the spread of this virus. Indonesia also started implementing this policy in March 2020 (kmt, 2020). This policy's impact is the temporary closure of various public facilities, which have a significant risk of being used by large numbers of people, such as schools, offices, places of worship, public transportation facilities, and even malls (kmt, 2020).

Even though schools and offices are temporarily closed, the activities in them continue. Students still learning from their home; therefore, the term "school from home" has risen. Because all activities are carried out from home, to maintain relationships with family, colleagues, and friends, and still be able to do their jobs, the internet has become a major necessity. One application that has experienced a surge in usage during this pandemic is social media (Cahya, 2020).

In Indonesia, some of the most commonly used media are WhatsApp, Instagram, Twitter, and Facebook. This research will process twitter data related to SFH in Indonesia and analyze the sentiment. Current sentiment analysis plays a somewhat curvy role. By using sentiment analysis, an organization can get an opinion on a keyword. This method is considered faster and can get a wider respondent than questionnaires and surveys to find out user opinions (Mas'udah et al., 2020).

The many algorithms that can be used in sentiment analysis, this study will compare the SVM, Naive Bayes, and Decision Tree algorithms to obtain an algorithm with good performance through...
its ROC curve. The final objective of this research is to get an idea of how many positive and negative sentiments are so that the related parties can immediately take the right policy

**Material and Methods**

The first step begins with data collection, which is getting the tweets to be processed by using the twisted library from python. There are many ways to get tweets by searching for tweets that contain a specific hashtag or searching for tweets from a particular account. This research will look for tweets that contain certain hashtags. Hashtags used are "belajardarirumah" or "BelajarDariRumah" or "belajardirumah" or "BelajarDiRumah". In English, those similar terms mean "LearningFromHome" or "Studying / ScholFromHome".

We will get many columns of tweets, including usernames, tweets, likes, retweets, and others. This research will use a supervised method, whereas of the many columns, there is not a single column that can be used to categorize which class a tweet belongs to, so a labeling process is required. In general, this labeling can be done manually or automatically, and in this study, labeling was done manually. The data will be grouped into two classes, which are, positive and negative sentiments.

The next step is EDA which is usually carried out to understand the data better, look for anomalous data, see important variables and also get initial insights. EDA is usually carried out either statistically or using graphs (data visualization) (Cox, 2017; Guzman et al., 2017; Halibas et al., 2019; Wahyuni et al., 2019). EDA carried out in this study will also use statistics and graphics. After we understand the data we have, we can start preprocessing the data. The preprocessing includes Case Folding, Cleansing, processing of abbreviations, tokenization, stopword removal, stemming, feature extraction using TF-IDF which will be continued to POS TAGs.

![Figure 1. Research methodology](image)

After the preprocessing is complete, the data is ready to be entered into the model building process. The algorithm will carry out a learning process for data to produce knowledge. This research will use the SVM algorithm, Naive Bayes, and Decision Tree. For this process, the data will be split into training data and test data. Data splitting will be done using the holdout method with the composition of the training and test data of 70:30. After the model has been built, a performance assessment will be carried out for each model. It will be evaluated which model has the best performance. The evaluation will be carried out using the values from the ROC curve.
Results and Discussion

EDA and data preprocessing

After the data collection and labeling process is complete, we get 1058 tweets related to the hashtag learning from home "or" BelajarDariRumah "or" learning from home "or" BelajarDiRumah "caused by the Covid pandemic. A lot of data needs to be removed from the dataset because when the twint script is run, the earliest date of data that can be retrieved is in 2010. Even though in 2010 there has been no Covid pandemic, because it is not related to a pandemic, this kind of data needs to be removed from the dataset to be processed. When is the tweet traffic hitting its peak? To answer this question, we have created a function that counts the occurrences of tweets each month. The results of this function are then displayed in raphical form, as shown in figure 2.

![Tweet Distribution](image1.png)

Figure 1. Tweet distribution each month

What is going on at the highest peak in March (21st) and in April (13th)? It turns out that on March 21st, in many Indonesian cities, the local government announced an extension of the school from home policy, which had previously begun on 16-21. Because the condition was not yet favorable, the strategy has been extended again. While in April 13th Nadiem Anwar Makarim, the Minister of Education and Culture (Mendikbud), partnered with one of Indonesia's television stations, TVRI, to expand the teaching-learning process students at home due to the Corona pandemic (Covid-19), so that people who have limited access to the internet can continue to study at home, through their television.

![Token Distribution in March 21st](image2.png)

Figure 3. Token distribution in March 21st
What are they talking about in peak time? To answer what is being commented on, we can use a token. A token represents a word in a sentence. The tokens that often appear on March 21st (in Fig. 3) are "belajardirumah", "com", "Twitter", "pic" and "tugas". The first token most likely indicates a hashtag for this policy. While "com", "Twitter", "pic", more likely, this token does not impact the tweet content that we are going to process, so it is likely to get removed using the stopword removal process. 5th token has meaning, is not a stopword, and if this is deleted, our tweet will change its information content. After the stopword removal process is complete, we get a reasonably clean data set, and it no longer contains meaningless words. Token distribution has changed when compared to the graph image on the left.

The most frequently mentioned users in March are adhiemassardi and darmaningtyas (shown in Fig. 4). Both are individuals, Adi Massardi and Darmaningtyas are an activist, who was very vocal in criticizing the government, especially concerning the implementation of SFH policies. While in April, the most mentioned user was tvrinasional and itjen kemdikbud. Both are not individuals but are representative of an institution. In April, from the previous paragraph, we know that the Ministry of Education and Culture began collaborating with TVRI. So that if these two users get the most frequent mention, it is normal. Not contrary to previous findings.

Modelling

Before modeling can be done, the distribution of data for each class must be checked first, whether it is balanced or not, because this affects the performance of the model. From our dataset, 0 is a label that indicates negative sentiment, while 1 indicates positive sentiment. From the following figure, the distribution of data between classes is not balanced (the amount of data distribution between classes is not the same). To overcome this unbalanced distribution of data between classes, various techniques can be used, one of which is SMOTE, which will be applied in this research. After implementing SMOTE, the distribution of data between classes becomes more balanced (the amount of distribution is the same).
The next step is to train the model and test the model. For this stage, the dataset needs to be divided into 2 using the holdout method. The comparison of the dataset for training and testing is 70:30. The algorithms used to build the model are SVM, Naive Bayes (Multinomial), and Decision Tree. This study will compare how the model performs without SMOTE and with SMOTE. The distribution of the number of training and test data for each is as follows.

After the model is made, it is then tested with test data, and the performance of the model is seen from the AUC value (areas under the curve), it measures the areas under the ROC curve. The closer the AUC is to 1, the better performance the classifier has. AUC graph of each model made (3 models - SVM, Multinomial Naive Bayes, Decision Tree) is as follows.

For models developed with SVM, better performance is obtained when using data handled first with SMOTE, an increase from data without SMOTE by 18 points.
The same thing also happens for the model developed with Multinomial Naive Bayes. Better performance is obtained when using data handled first with SMOTE, the increase in data without SMOTE is greater than that of SVM, which is 27 points.

For models developed with a Decision Tree, better performance is obtained when using data handled first with SMOTE, the increment from data without SMOTE is not much different, which is only 5 points. From the performance assessment of 3 models with SMOTE and without SMOTE, the best AUC value is obtained by a model developed using SVM with SMOTE, the resulting confusion matrix is as follows.

The picture above shows that the class with the right prediction is mostly dominated by negative sentiment, namely 209 data, followed by positive sentiment with 163 data.
Conclusion

The results of the study can be concluded that:

1. We can get a better picture of the situation we are facing, regarding the data that we had, by using EDA.
2. EDA can also give us an overview of some preprocessing steps that might be required to make our tweet data cleaner.
3. The highest model performance value obtained in carrying out the classification process (AUC value) is 0.88 using the SVM algorithm with data sharing using the holdout method and performing the data balancing stage using the SMOTE method.

Acknowledgment

The author would like to thank UPN "Veteran" Jawa Timur who has funded this research and also LPPM UPN "Veteran" Jawa Timur who has been willing to be the research subject.

References

Cahya, I. (2020). Nasib raksasa media sosial kala pandemi, pengguna naik namun iklan turun | merdeka.com. Available at: https://www.merdeka.com/teknologi/nasib-raksasa-media-sosial-kala-pandemi-pengguna-naik-namun-iklan-turun.html (Accessed: 7 August 2020).

Cox, V. (2017). Exploratory data analysis what data do i have?, 47–74. doi: 10.1007/978-1-4842-2256-0_3.

Guzman, E., Alkadhi, R., & Seyff, N. (2017). An exploratory study of Twitter messages about software applications. Requirements Engineering, 22(3), pp. 387–412. doi: 10.1007/s00766-017-0274-x.

Halibas, A., Matthew, A. C., & Reazol, L. B. (2019). Determining the intervening effects of exploratory data analysis and feature engineering in telecoms customer churn modelling in 2019 4th MEC International Conference on Big Data and Smart City, ICBDSC 2019. Institute of Electrical and Electronics Engineers Inc., 1–7. doi: 10.1109/ICBDSC.2019.8645578.

Kmt. (2020). Jokowi calls for ‘social distancing’ to stem virus spread - National - The Jakarta Post. Available at: https://www.thejakartapost.com/news/2020/03/15/jokowi-calls-for-social-distancing-to-stem-virus-spread.html (Accessed: August 7th 2020).

Mas'udah, E., Wahyuni, E. D., & Arifiyanti, A. A. (2020). Analisis sentimen: Pemindahan ibu kota Indonesia pada twitter. Jurnal Informatika dan Sistem Informasi (JIFoSI). Doi:10.1109/ICBDSC.2019.8645578

Wahyuni, E. D., Arifiyanti, A. A., & Kustyani, M. (2019). Exploratory data analysis dalam konteks klasifikasi data mining. Prosiding Nasional Rekayasa Teknologi Industri dan Informasi XIV Tahun 2019 (ReTII), 2019(November), 263–269.