Springer Link

Conference proceedings | © 2021

Proceedings of International Conference on Smart Computing and Cyber Security

Strategic Foresight, Security Challenges and Innovation (SMARTCYBER 2020)

Editors: (view affiliations) Prasant Kumar Pattnaik, Mangal Sain, Ahmed A. Al-Absi, Pardeep Kumar

Presents research works in the field of smart computing and cyber security

Discusses results of SMARTCYBER 2020 held in the Department of Smart Computing, Kyungdong University, Global Campus, South Korea

Serves as a reference for researchers and practitioners in academia and industry

Part of the book series: Lecture Notes in Networks and Systems (LNNS, volume 149)

Conference series link(s): SMARTCYBER: International conference on smart computing and cyber security : strategic foresight, security challenges and innovation

Conference proceedings info: SMARTCYBER 2020.

Table of contents (35 papers)

Search within book

← Previous Page 1 of 2 Next→

Editors and Affiliations

School of Computer Engineering, Kalinga Institute of Industrial Technology, KIIT Deemed to be University, Bhubaneswar, India
Prasant Kumar Pattnaik

Division of Information and Communication Engineering, Dongseo University, Busan, Korea (Republic of)
Mangal Sain

Department of Smart Computing, Kyungdong University Global Campus, Gangwondo, Korea (Republic of)
Ahmed A. Al-Absi

Department of Computer Science, Swansea University, Bay Campus, Swansea, UK
Pardeep Kumar

Back to top

About the editors

Prasant Kumar Pattnaik, Ph.D. (Computer Science), Fellow IETE, Senior Member IEEE, is a Professor at the School of Computer Engineering, KIIT Deemed University, Bhubaneswar. He has more than a decade of teaching and research experience and awarded half dozen of Ph.D. Dr. Pattnaik has published numbers of research papers in peer-reviewed international journals and conferences and filed many patents. He also edited book volumes in Springer and IGI Global Publication. His areas of interest include mobile computing, cloud computing, cyber security, intelligent systems, and brain–computer interface. He is one of the Associate Editors of Journal of Intelligent & Fuzzy Systems, IOS Press, and Intelligent Systems Book Series Editor of CRC Press, Taylor Francis Group.

Mangal Sain received the Master of Application degree from India in 2003 and the Ph.D. degree in Computer Science from Dongseo University, Busan, South Korea, in 2011. Since 2011, he has been an Assistant Professor with the Department of Information and Communication Engineering, Dongseo University, Busan, South Korea. He has published over 40 international publications. His current research interests include wireless sensor network, middleware, cloud computing, embedded system, and the Internet of Things. He is a member of TIIS and has participated as a TPC member in several international conferences. Pardeep Kumar received the B.E. degree in Computer Science from Maharishi Dayanand University, Haryana (India), in 2002, the M.Tech. degree in Computer Science from Chaudhary Devi Lal University, Haryana (India), in 2006, and the Ph.D. degree in Ubiquitous Computing from Dongseo University, Busan (South Korea) in 2012. He is currently a Lecturer/Assistant Professor with the Department of Computer Science, Swansea University, Swansea, UK. From 2012 to 2018, he had held postdoc positions at the Department of Computer Science, Oxford University, Oxford UK (08/2016–09/2018), at the Department of Computer Science, The Arctic University of Norway, Tromso, Norway (08/2015–08/2016), and at Centre for Wireless Communications and the Department of Communications Engineering, University of Oulu, Finland (04/2012 to 08/2015).
Ahmed A. Al-Absi, Ph.D (Computer Science), is an Associate Professor at the Smart Computing Department, Kyungdong University Global Campus, South Korea. He is currently Dean of International Faculty and Director of Global Academic Collaboration Centers at Kyungdong University Global. He has more than ten years of experience in teaching and university lecturing in the areas of database design and computer algorithms. Dr. Al-Absi has published numbers of research papers in peer-reviewed international journals and conferences. His research areas are Big Data, Large Scale Data Process Systems, Cloud Computing, IoT, VANET, Deep Learning, Parallel Computing, Security, and Bioinformatics. His professional experience includes being a speaker at a number of renowned research conferences and technical meetings such as IEEE, Korea ICT leaders forum, and reviewer for refereed journals and conferences on data-intensive computing as well as an examiner for postgraduate scholars in his research areas.

Back to top

Towards A Sentiment Analyzer for Low-Resource Languages

Dian Indriani[1], Arbi Haza Nasution[1], Winda Monika[2], and Salhazan Nasution[3]

[1] Informatics Engineering, Universitas Islam Riau, Indonesia
[2] Library Science, Universitas Lancang Kuning, Indonesia
[3] Informatics Engineering, University of Riau, Indonesia
diianindrianii23@student.uir.ac.id, arbi@eng.uir.ac.id,
windabi.wm@gmail.com, salhazan@lecturer.unri.ac.id

Abstract. Twitter is one of the top influenced social media which has a million number of active users. It is commonly used for microblogging that allows users to share messages, ideas, thoughts and many more. Thus, millions interaction such as short messages or tweets are flowing around among the twitter users discussing various topics that has been happening world-wide. This research aims to analyse a sentiment of the users towards a particular trending topic that has been actively and massively discussed at that time. We chose a hashtag *#kpujangancurang* that was the trending topic during the Indonesia presidential election in 2019. We use the hashtag to obtain a set of data from Twitter to analyse and investigate further the positive or the negative sentiment of the users from their tweets. This research utilizes rapid miner tool to generate the twitter data and comparing Naive Bayes, K-Nearest Neighbor, Decision Tree, and Multi-Layer Perceptron classification methods to classify the sentiment of the twitter data. There are overall 200 labeled data in this experiment. Overall, Naive Bayes and Multi-Layer Perceptron classification outperformed the other two methods on 11 experiments with different size of training-testing data split. The two classifiers are potential to be used in creating sentiment analyzer for low-resource languages with small corpus.

Keywords: Twitter · Sentiment Analysis · Low-resource languages · Naive Bayes · K-Nearest Neighbor · Decision Tree · Multi-Layer Perceptron

1 Introduction

A rich sentiment analysis corpus is crucial in creating a good sentiment analyzer. Unfortunately, low-resource languages like Indonesian lack such resources. Some prior studies focused on enriching low-resource languages [6,7,8,9,10,11,12,13]. The rapid growth of online textual data creates an urgent need for powerful text mining techniques [1]. Sentiment analysis or opinion mining is a part of text mining. Sentimen analysis basically is a computational research that analyses the textual expression from opinion, sentiment and emotion of the social media users

[4]. It extracts attributes and components of the documented object. Through the sentiment analysis of the text, information such as the public's emotional status, views on some social phenomena, and preferences for a product can be obtained [20]. Hence, the perspective of the users either positive or negative could be revealed.

During the Indonesia 2019 presidential election, the competition was quite fierce where there were only two candidates fighting in the battle. Most of supporters from these two candidates were actively campaigning their candidates on social media and twitter was the highly used social media chosen by them. Due to the huge enthusiasm of those two supporters, most of the time fierce debate among them could not be avoided. One of the trending topic emerged was during the recapitulation of the votes. Twitter users reacted to the several findings showed that the calculation of the votes led to deception. Foremost, supporters from one party, from Prabowo Subianto volunteers found that many evidence of the wrong data were inputed to the system. Thus, the real count results was irrelevant with the information displayed on the system. This finding made the situation in Indonesia heating up. Supporters from Prabowo Subianto was upset and condemned the General Election Commision as the legal institution to take full responsibility of this matter. To express their disappointment, most of the twitter users created hashtag *#kpujangancurang* or "The General Election Commision should not be unfair". However, this issue was objected by the opponent supporters. They argued that this issue was merely caused by human error. The same hashtag actually was being used by the both parties, so that no one knows the exact sentiment of the tweets. Therefore, sentiment analyzer that could analyse the sentiment of the tweets is crucial

In sentiment analysis, the available corpus in Indonesian language is scarce. The existing machine learning tool such as rapidminer has two sentiment analyzer which are Aylien and Rosette, do not cover Indonesian language. We run an experiment by using the *#kpujangancurang* hastag to obtain corpus using rapidminer to extract the tweets and then analyse the sentiment of users by using four machine learning methods which are Naive Bayes, K-Nearest Neighbor, Decision Tree, and Multi-Layer Perceptron classification. The objective of this research is to find out which classifier is more suitable to be used in creating sentiment analyzer for low-resource languages with small corpus.

2 Literature Study

Several researches have been done on sentiment analysis. A study attempted to analyze the online sentiment changes of social media users using both the textual and visual content by analysing sentiment of twitter text and image [19]. Another related study performed linguistic analysis of the collected corpus and explain discovered phenomena to build a sentiment classifier, that is able to determine positive, negative and neutral sentiments for a document [14].

Furthermore, several studies have been done using machine learning method on sentiment analysis, for instance a study showed that a similar research on a twitter sentiment analysis by applying Naive Bayes classifier method to investigate the sentiment analysis of the twitter users on the traffic jam in Bandung [18]. Another study focused on data classification using k-NN (k-Nearest Neighbors) and Naive Bayes where the Corpus was downloaded from TREC Legal Track with a total of more than three thousand text documents and over twenty types of classifications [17]. A study utilized maximum entropy part of speech tagging and support vector machine to analyse the public sentiment. The study used dataset in Indonesian language and implemented machine learning approached due to its efficiency for integrating a large scale feature into a model. This kind of approach has been successfully implemented in various tasks such as natural language processing [16]. A study proposed a semi-automatic, complementary approach in which rule-based classification, supervised learning and machine learning are combined into a new method to achieve a good level of effectiveness [15]. Another study about opinion mining for hotel rating through reviews using decision tree classifier shows the advantage of using the algorithm is that the rule set can be easily generated and by analyzing each level of the tree, a particular service quality can be improved [3]. Deep learning methods also have been widely used in sentiment analysis tasks [5,21]. However, these studies show different accuracy from each machine learning method used depending on the size of the corpus.

RapidMiner is an open source software[4]. RapidMiner is one of the solutions for doing analysis on data mining, text mining and prediction analysis. RapidMiner uses various descriptive technique and prediction in giving a new insight to the users so that allows and helps users to make a better decision. RapidMiner is a standalone software and enable to be integrated with its own products. RapidMiner provides GUI (Graphic User Interface) for designing an analytical pipeline. GUI will generate XML (Extensible Markup Language) that defines the analytical process of the users need to be applied on the data. This file is later on read by rapid miner to be automatically analyzed. We use rapid miner due to several reasons: it eases in getting the dataset, it can be connected to twitter, it enables to search the topic as query so that the intended topic will emerge and can be saved in excel file, furthermore it allows extracting plentiful data. A study examined an anomaly detection extension for RapidMiner in order to assist non-experts with applying eight different k-nearest-neighbor and clustering based algorithms on their data [2]. However, in this study, we only use RapidMiner to extract data from Twitter.

3 Research Methodology

In this study, we use a dataset that was gotten from the tweets' document. We utilized rapid miner to obtain the tweets from the hashtag *#kpujangancurang*. To investigate further about the hashtag *#kpujangancurang*, we compare Naive Bayes, K-Nearest Neighbor, Decision Tree, and Multi-Layer Perceptron classification methods to classify the sentiment of the twitter data. There are two

[4] https://rapidminer.com

steps of the document classification: the first one is training the document that has been categorized. And the second one is training the uncategorized document. The four methods classify the distribution of the positive and negative sentiments. There are overall 200 labeled data in this experiment. To evaluate the performance of the sentiment analyzer, we use accuracy as the evaluation measure.

3.1 System Workflow

Overview Sentiment analysis overview is described in details which is depicted in the Fig. 1 below.

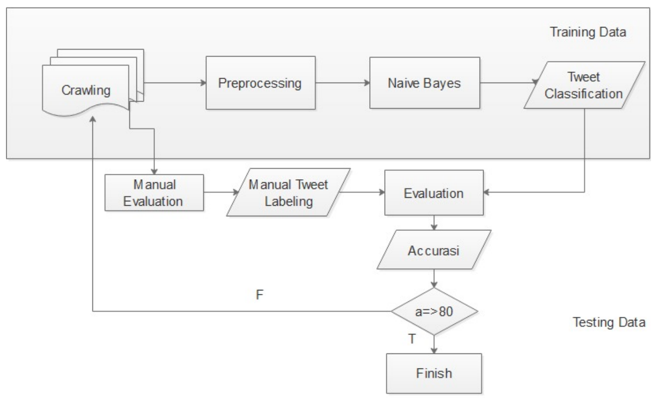

Fig. 1. Example of sentiment analysis workflow using Naive Bayes method.

- Data Crawling: It is a process of aggregating data from twitter using rapid miner as a tool. The aggregated data from hashtag *#kpujangancurang* is used as training dataset and testing dataset.
- Preprocessing: It is a process of cleaning the data by deleting common words by referring to stopwords.
- Classification: Naive Bayes method is applied to classify the sentiment into positive and negative sentiments. The rest of methods will be used in the same manner.
- Evaluation: The classification result from classifiers is evaluated with the manual labeling classification. The accuracy of the classification determine whether a new training dataset need to be added or not to reach the accuracy threshold of 80%.

Dataset How do we get the dataset is depicted in Fig. 2 below. The dataset that we analyse is in Indonesian language. Firstly, the tweet was queried by using the hashtag *#kpujangancurang*.

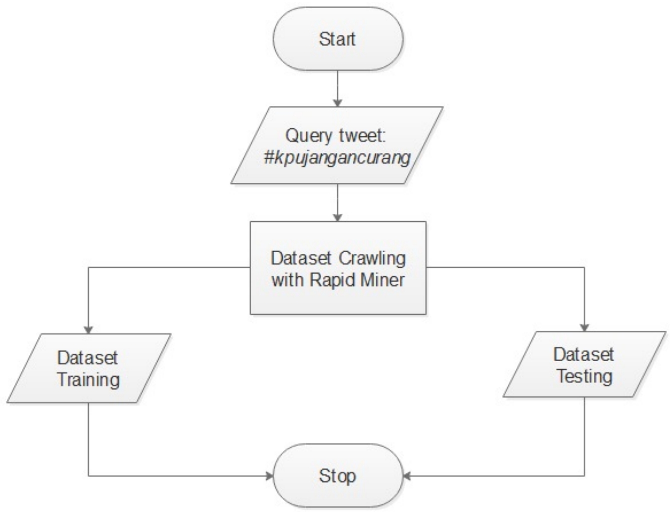

Fig. 2. Dataset Flow.

Then, the queried data is crawled by using rapidminer. The result from the query is divided into two part: training data and testing data. Testing data is classified by using the classifiers and then the result was marked with negative and positive sentiment label. Whereas, the training data is classified manually and the result was marked the same way as testing data is treated. Training data will be used during the evaluation to determine the accuracy of the result. Table 1 shows example of evaluation of the predicted sentiment by the classifiers.

Table 1. Example of Evaluation of The Predicted Sentiment

Testing Data	**Predicted Sentiment**	**Manually Labeled Sentiment**	**Accuracy**
kalau terus melanggar, hukuman-nya segera diterapkan	Positive	Positive	Accurate
kalau bersih kenapa takut audit forensic	Negative	Negative	Accurate
harus banyak belajar ke @BKNgoid dalam hal penyelenggaraan akbar	Positive	Positive	Accurate
Kebenaran meninggikan derajat bangsa tetapi dosa adalah noda bangsa	Negative	Positive	Inaccurate

Preprocessing Preprocessing process is an important step for the next step which disposes the non-useful attribute that can be noise for the classification process. Data that is imported in this process is a raw data, thus the result of this process is a high-quality document expected to ease the classification process. Preprocessing process is depicted in Fig. 3.

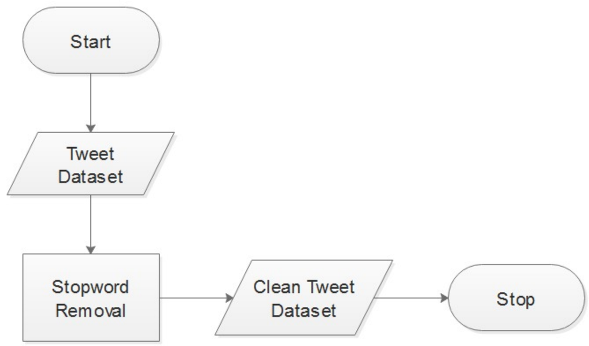

Fig. 3. Preprocessing Flow.

Table 2. Preprocessing Process Example

Stage	Before	After
1	Benar juga, kpu yang membuat rakyat resah. Aduh kejamnya kecurangan.	Benar juga kpu yang membuat rakyat resah Aduh kejamnya kecurangan
2	Benar juga kpu yang membuat rakyat resah Aduh kejamnya kecurangan	benar juga kpu yang membuat rakyat resah aduh kejamnya kecurangan
3	benar juga kpu yang membuat rakyat resah aduh kejamnya kecurangan	-benar- -juga- -kpu- -yang- -membuat- -rakyat- -resah- -aduh- -kejamnya- -kecurangan-
4	-benar- -juga- -kpu- -yang- -membuat- -rakyat- -resah- -aduh- -kejamnya- -kecurangan-	-benar- -kpu- -membuat- -rakyat- -resah- -kejamnya- -kecurangan-

This step is started with punctuation removal, case folding, tokenizing, and finally stopword removal which is intended to remove words that are not relevant with the topic. If in the tweet document exists irrelevant words, then these words will be removed. An example of each stage of the preprocessing process is listed in Table 2. The detailed preprocessing stage is as follow:

- Removing punctuation. This stage is the initial process in order to get pure text containing words only so that further processing becomes easier.
- Case Folding. This stage is the process of changing uppercase letters to lowercase letters.
- Tokenizing. In this stage, each word will be separated based on the specified space.
- Filtering. This stage is the removal of unimportant words based on Indonesian stopwords.

Term Frequency - Inverse Document Frequency (TF-IDF) After doing the preprocessing, the next step is to weight the words using the tf-idf calculation. Tf-idf is a way of giving the weight of a word (term) to words. For single words, each sentence is considered as a document. The following is an example of tf-idf calculation. The example of documents that want to be weighted is shown in Table 3 and the sample tf-idf result of Document A is shown in Table 4.

Table 3. Example of Documents

Tweet Document	Text
Document A	Jangan ancam rakyat, rakyat indonesia pintar
Document B	Rakyat tidak pernah gagal bernegara, pemerintah yang gagal bernegara
Document C	Suara rakyat dicuri, bagaimana uang rakyat

Table 4. TF-IDF Score of Document A

Word	TF	IDF	Weight
ancam	1	0.477	0.477
bernegara	0	0.176	0
gagal	0	0.176	0
jangan	1	0.477	0.477
rakyat	0.4	-0.2218	-0.0887
indonesia	1	0.477	0.477
pintar	1	0.477	0.477
tidak	0	0.477	0
pernah	0	0.477	0
pemerintah	0	0.477	0
dicuri	0	0.477	0
bagaimana	0	0.477	0
uang	0	0.477	0

Classifier The last step is classifying the weighted data with Naive Bayes, K-Nearest Neighbor, Decision Tree, and Multi-Layer Perceptron classification methods. To evaluate which classifiers are best for scarce corpus, we experimented by changing the size of the training-testing data split from 0.25-0.75 to 0.75-0.25. The evaluation is done by measuring the accuracy of the classifiers for each scenario as shown in Fig. 4.

4 Result

We obtained 200 twitter data using rapidminer. From the 200 twitter data, we conducted 11 experiments with different size of training-testing data split. Every classifiers shows a trend of increased accuracy on larger size of training data. However, Naive Bayes and Multi-Layer Perceptron classifier outperformed the other two methods in overall experiment as shown in Fig. 4. Decision Tree classifier shows a very low performance on small data, while K-Nearest Neighbor classifier shows accuracy below 0.76 on all combination size of training-testing data split. Both Naive Bayes and Multi-Layer Perceptron classifier have the highest accuracy on all combination size of training-testing data split and show consistent increased of accuracy as the training data size is increased.

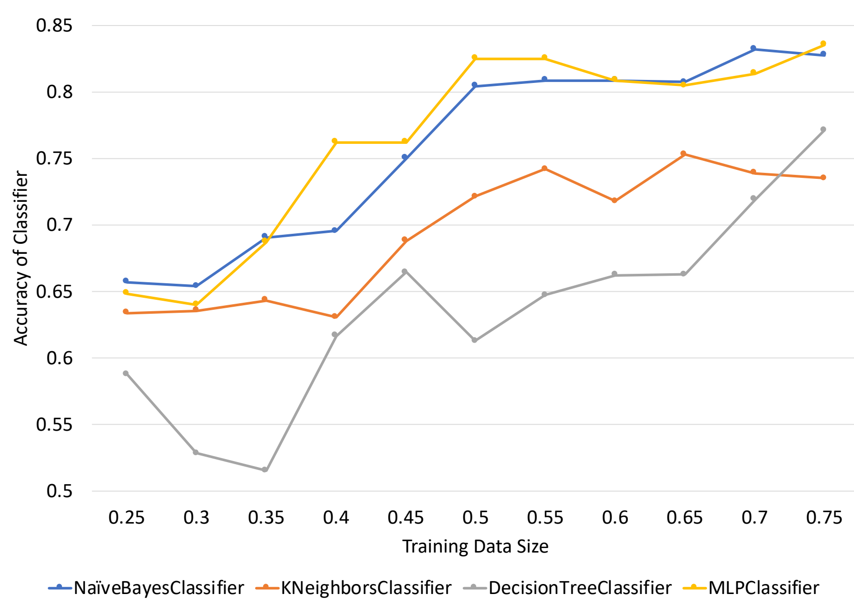

Fig. 4. Accuracy Comparison of Classifiers.

5 Conclusion

We have build a sentiment analyzer to identify users' sentiment from Twitter hashtag *#kpujangancurang* toward the General Election Commission We use the hashtag to obtain a set of data from Twitter to analyse and investigate further the positive and the negative sentiment of the users from their tweets. This research utilizes rapid miner tool to generate the twitter data and comparing Naive Bayes, K-Nearest Neighbor, Decision Tree, and Multi-Layer Perceptron classification methods to classify the sentiment of the twitter data. There are overall 200 labeled data in this experiment. Overall, Naive Bayes and Multi-Layer Perceptron classifier outperformed the other two methods on 11 experiments with different size of training-testing data split. The two classifiers are potential to be used in creating sentiment analyzer for low-resource languages with small corpus. In our future work, we will compare the accuracy of both Naive Bayes and Multi-Layer Perceptron classifier on bigger size of corpus.

Acknowledgment

This research is funded by Universitas Islam Riau.

References

1. Aggarwal, C.C., Zhai, C.: Mining text data. Springer Science & Business Media (2012)
2. Amer, M., Goldstein, M.: Nearest-Neighbor and Clustering based Anomaly Detection Algorithms for RapidMiner. Proceedings of the 3rd RapidMiner Community Meeting and Conferernce (RCOMM 2012) pp. 1–12 (2012)
3. Gupta, S., Jain, S., Gupta, S., Chauhan, A., et al.: Opinion mining for hotel rating through reviews using decision tree classification method. International Journal of Advanced Research in Computer Science **9**(2), 180 (2018)
4. Liu, B.: Sentiment analysis and opinion mining. Synthesis lectures on human language technologies **5**(1), 1–167 (2012)
5. Mukherjee, S., Adhikari, A., Roy, M.: Malignant melanoma detection using multi layer perceptron with optimized network parameter selection by pso. In: Contemporary Advances in Innovative and Applicable Information Technology, pp. 101–109. Springer (2019)
6. Nasution, A.H.: Pivot-based hybrid machine translation to support multilingual communication for closely related languages. World Transactions on Engineering and Technology Education **16**(2), 12–17 (2018)
7. Nasution, A.H., Murakami, Y., Ishida, T.: Constraint-based bilingual lexicon induction for closely related languages. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). pp. 3291–3298. Paris, France (May 2016)
8. Nasution, A.H., Murakami, Y., Ishida, T.: A generalized constraint approach to bilingual dictionary induction for low-resource language families. ACM Trans. Asian Low-Resour. Lang. Inf. Process. **17**(2), 9:1–9:29 (Nov 2017). https://doi.org/10.1145/3138815, `http://doi.acm.org/10.1145/3138815`

9. Nasution, A.H., Murakami, Y., Ishida, T.: Plan optimization for creating bilingual dictionaries of low-resource languages. In: 2017 International Conference on Culture and Computing (Culture and Computing). pp. 35–41 (Sept 2017). https://doi.org/10.1109/Culture.and.Computing.2017.21
10. Nasution, A.H., Murakami, Y., Ishida, T.: Similarity cluster of indonesian ethnic languages. In: Proceedings of the First International Conference on Science Engineering and Technology (ICoSET 2017). pp. 12–27. Pekanbaru, Indonesia (November 2017)
11. Nasution, A.H., Murakami, Y., Ishida, T.: Designing a collaborative process to create bilingual dictionaries of indonesian ethnic languages. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). pp. 3397–3404. European Language Resources Association (ELRA), Paris, France (may 2018)
12. Nasution, A.H., Murakami, Y., Ishida, T.: Generating similarity cluster of indonesian languages with semi-supervised clustering. International Journal of Electrical and Computer Engineering (IJECE) **9**(1), 1–8 (2019)
13. Nasution, A.H., Syafitri, N., Setiawan, P.R., Suryani, D.: Pivot-based hybrid machine translation to support multilingual communication. In: 2017 International Conference on Culture and Computing (Culture and Computing). pp. 147–148 (Sept 2017). https://doi.org/10.1109/Culture.and.Computing.2017.22
14. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: LREc. vol. 10, pp. 1320–1326 (2010)
15. Prabowo, R., Thelwall, M.: Sentiment analysis: A combined approach. Journal of Informetrics **3**(2), 143–157 (2009)
16. Putranti, N.D., Winarko, E.: Analisis sentimen twitter untuk teks berbahasa indonesia dengan maximum entropy dan support vector machine. IJCCS (Indonesian Journal of Computing and Cybernetics Systems) **8**(1), 91–100 (2014)
17. Rasjid, Z.E., Setiawan, R.: Performance Comparison and Optimization of Text Document Classification using k-NN and Naïve Bayes Classification Techniques. Procedia Computer Science **116**, 107–112 (2017). https://doi.org/10.1016/j.procs.2017.10.017, `https://doi.org/10.1016/j.procs.2017.10.017`
18. Rodiyansyah, S.F., Winarko, E.: Klasifikasi posting twitter kemacetan lalu lintas kota bandung menggunakan naive bayesian classification. IJCCS (Indonesian Journal of Computing and Cybernetics Systems) **6**(1) (2012)
19. You, Q.: Sentiment and emotion analysis for social multimedia: Methodologies and applications. Proceedings of the 2016 ACM Multimedia Conference (MM'16) pp. 1445–1449 (2016). https://doi.org/10.1145/2964284.2971475, `http://dl.acm.org/citation.cfm?doid=2964284.2971475`
20. Yuan, H., Wang, Y., Feng, X., Sun, S.: Sentiment Analysis Based on Weighted Word2vec and Att-LSTM pp. 420–424 (2019). https://doi.org/10.1145/3297156.3297228
21. Zhang, L., Wang, S., Liu, B.: Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **8**(4), e1253 (2018)