Towards a Sentiment Analyser for Low-resource Languages

by Arbi Haza Nasution

Towards a Sentiment Analyser for Low-resource Languages

Dian Indriani, Arbi Haza Nasution, Winda Monika, and Salhazan Nasution

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 of interactions such as short messages or tweets are flowing around among the twitter users discussing various topics that have been happening worldwide. 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 RapidMiner 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 analyser for low-resource languages with small corpus.

Keywords Twitter · Sentiment analysis · Low-resource languages · Naive Bayes · K-nearest neighbor · Decision tree · Multi-layer perceptron

D. Indriani · A. H. Nasution (✉)
Informatics Engineering, Universitas Islam Riau, Riau, Indonesia
e-mail: arbi@eng.uir.ac.id

D. Indriani
e-mail: diianindriani23@student.uir.ac.id

W. Monika
Library Science, Universitas Lancang Kuning, Riau, Indonesia
e-mail: windamonika@unilak.ac.id

S. Nasution
Informatics Engineering, Universitas Riau, Riau, Indonesia
e-mail: salhazan@lecturer.unri.ac.id

P. K. Pattnaik et al. (eds.), *Proceedings of International Conference on Smart Computing and Cyber Security*, Lecture Notes in Networks and Systems 149,
https://doi.org/10.1007/978-981-15-7990-5_10

1 Introduction

A rich sentiment analysis corpus is crucial in creating a good sentiment analyser. Unfortunately, low-resource languages like Indonesian lack such resources. Some prior studies focused on enriching low-resource languages [6–13]. 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. Sentiment 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 huge enthusiasm of those two supporters, most of the time fierce debate among them could not be avoided. One of the trending topics 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 evidences of the wrong data were inputted to the system. Thus, the real count results were 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 Commission 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 Commission 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, a sentiment analyser 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 analysers which are Aylien and Rosette and do not cover Indonesian language. We run an experiment by using the #kpujangancurang hashtag 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 analyser for low-resource languages with small corpus.

2 Literature Study

Several researches have been done on sentiment analysis. A study attempted to analyse the online sentiment changes of social media users using both the textual and visual contents 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 and Naive Bayes where the corpus was downloaded from TREC Legal Track with a total of more than 3000 text documents and over 20 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 shows opinion mining for hotel rating through reviews using decision tree classifier. The advantage of using the algorithm is that the rule set can be easily generated and by analysing each level of the tree, so 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. RapidMiner is one of the solutions for doing analysis on data mining, text mining and prediction analysis. RapidMiner uses various descriptive techniques and predictions 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 enables to be integrated with its own products. RapidMiner provides graphic user interface (GUI) for designing an analytical pipeline. GUI will generate Extensible Markup Language (XML) that defines the analytical process of the users need to be applied on the data. This file is later on read by RapidMiner to be automatically analysed. We use RapidMiner 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 obtained from the tweets' document. We utilized RapidMiner 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 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 analyser, we use accuracy as the evaluation measure.

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

- Data Crawling: It is a process of aggregating data from Twitter using RapidMiner 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 determines whether a

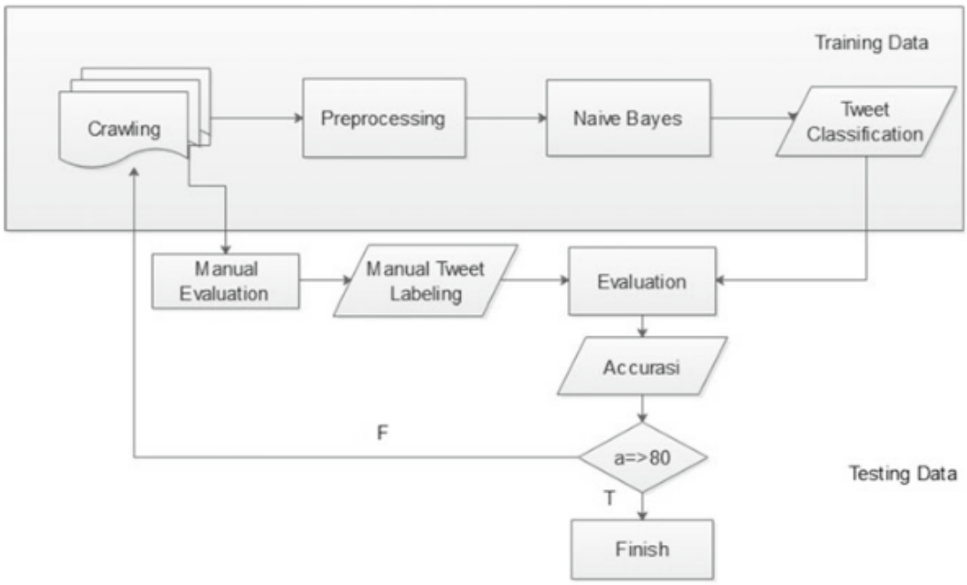

Fig. 1 Example of sentiment analysis workflow using Naive Bayes method

new training dataset needs to be added or not to reach the accuracy threshold of 80%.

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

Then, the queried data is crawled by using RapidMiner. The result from the query is divided into two parts: 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.

Preprocessing. This 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.

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 follows:

– Removing punctuation. This stage is the initial process in order to get pure text containing words only so that further processing becomes easier.

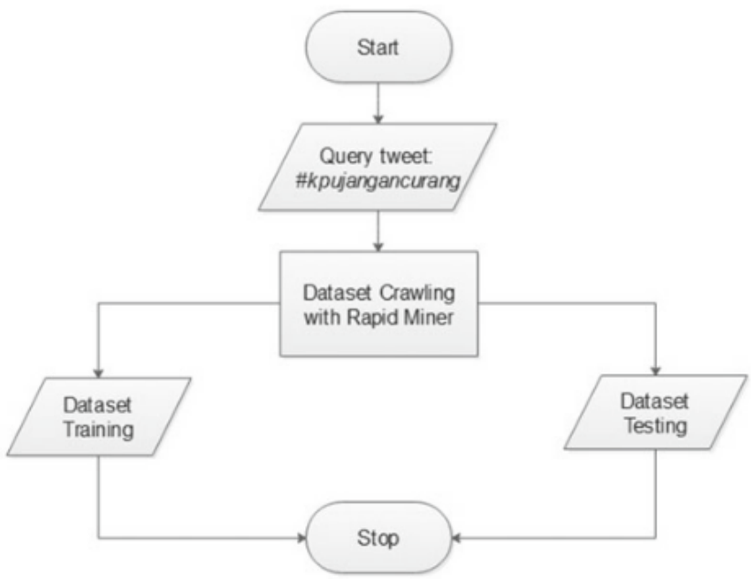

Fig. 2 Dataset flow

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

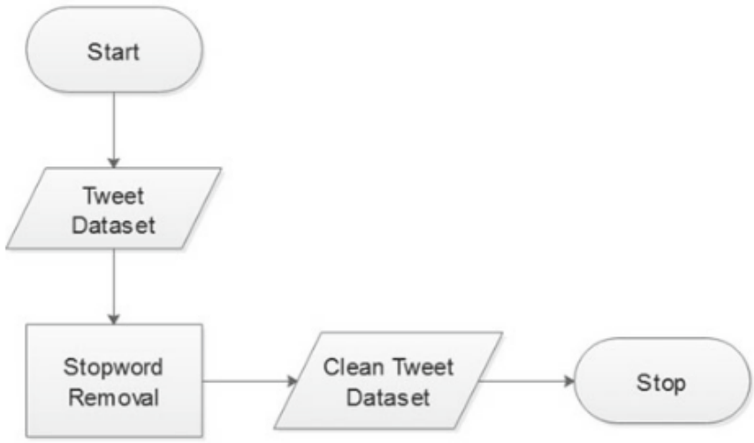

Fig. 3 Preprocessing flow

Table 2 Preprocessing process example

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

- 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.

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.

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

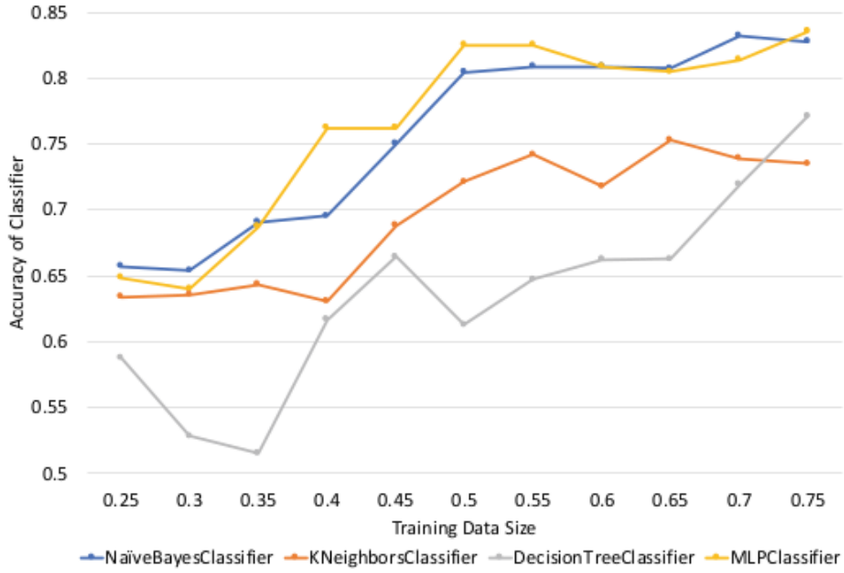

Fig. 4 Accuracy comparison of classifiers

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 classifier 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.

5 Conclusion

We have built a sentiment analyser 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 RapidMiner 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 for creating sentiment analyser 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.

Acknowledgements This research is funded by Universitas Islam Riau.

References

1. C.C. Aggarwal, C. Zhai, *Mining Text Data*. Springer Science & Business Media (2012)
2. M. Amer, M. Goldstein, Nearest-neighbor and clustering based anomaly detection algorithms for RapidMiner, in *Proceedings of the 3rd RapidMiner Community Meeting and Conference (RCOMM 2012)* (2012), pp. 1–12
3. S. Gupta, S. Jain, S. Gupta, A. Chauhan et al., Opinion mining for hotel rating through reviews using decision tree classification method. Int. J. Adv. Res. Comput. Sci. **9**(2), 180 (2018)
4. B. Liu, Sentiment analysis and opinion mining. Synthesis Lect. Human Language Technol. **5**(1), 1–167 (2012)
5. S. Mukherjee, A. Adhikari, M. Roy, Malignant melanoma detection using multi layer perceptron with optimized network parameter selection by pso, in *Contemporary Advances in Innovative and Applicable Information Technology* (Springer, Heidelberg, 2019), pp. 101–109
6. A.H. Nasution, Pivot-based hybrid machine translation to support multilingual communication for closely related languages. World Trans. Eng. Technol. Educ. **16**(2), 12–17 (2018)
7. A.H. Nasution, Y. Murakami, T. Ishida, Constraint-based bilingual lexicon induction for closely related languages, in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France (May 2016), pp. 3291–3298
8. A.H. Nasution, Y. Murakami, T. Ishida, 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
9. A.H. Nasution, Y. Murakami, T. Ishida, Plan optimization for creating bilingual dictionaries of low-resource languages, in *2017 International Conference on Culture and Computing (Culture and Computing)* (Sept 2017), pp. 35–41. https://doi.org/10.1109/Culture.and.Computing.2017.21
10. A.H. Nasution, Y. Murakami, T. Ishida, Similarity cluster of Indonesian ethnic languages, in *Proceedings of the First International Conference on Science Engineering and Technology (ICoSET 2017)*, Pekanbaru, Indonesia (November 2017), pp. 12–27
11. A.H. Nasution, Y. Murakami, T. Ishida, 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)*. European Language Resources Association (ELRA), Paris, France (May 2018), pp. 3397–3404
12. A.H. Nasution, Y. Murakami, T. Ishida, Generating similarity cluster of Indonesian languages with semi-supervised clustering. Int. J. Electrical Comput. Eng. (IJECE) **9**(1), 1–8 (2019)
13. A.H. Nasution, N. Syafitri, P.R. Setiawan, D. Suryani, Pivot-based hybrid machine translation to support multilingual communication, in *2017 International Conference on Culture and Computing (Culture and Computing)* (Sept 2017), pp. 147–148. https://doi.org/10.1109/Culture.and.Computing.2017.22
14. A. Pak, P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining. LREc **10**, 1320–1326 (2010)
15. R. Prabowo, M. Thelwall, Sentiment analysis: a combined approach. J. Informetrics **3**(2), 143–157 (2009)

16. N.D. Putranti, E. Winarko, Analisis sentimen twitter untuk teks berbahasa Indonesia dengan maximum entropy dan support vector machine. IJCCS (Indonesian J. Comput. Cybern. Syst.) **8**(1), 91–100 (2014)
17. R. Setiawan, Performance comparison and op-text document classification using k-NN and classification techniques. Proc. Comput. Sci. **116**, 107–112 (2017). https://doi.org/10.1016/j.procs.2017.10.017
18. S.F. Rodiyansyah, E. Winarko, Klasifikasi posting twitter kemacetan lalu lin-tas kota bandung menggunakan naive bayesian classification. IJCCS (Indonesian J. Comput. Cybern. Syst.) **6**(1) (2012)
19. Q. You, Sentiment and emotion analysis for social multimedia: methodologies and applications, in *Proceedings of the 2016 ACM Multimedia Conference (MM'16)* (2016), pp. 1445–1449. https://doi.org/10.1145/2964284.2971475, https://dl.acm.org/citation.cfm?doid=2964284.2971475
20. H. Yuan, Y. Wang, X. Feng, S. Sun, *Sentiment Analysis Based on Weighted Word2vec and Att-LSTM*, pp. 420–424 (2019). https://doi.org/10.1145/3297156.3297228
21. L. Zhang, S. Wang, B. Liu, Deep learning for sentiment analysis: a survey. Wiley Interdisc. Rev. Data Mining Knowl. Discovery **8**(4), e1253 (2018)

Towards a Sentiment Analyser for Low-resource Languages

ORIGINALITY REPORT

10%	7%	7%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Yudhi Arta, Evizal Abdul Kadir, Ari Hanggara, Des Suryani, Nesi Syafitri. "Chapter 15 Implementation of Motorcycle Monitoring Using Bluetooth with an Android-Based Microcontroller Using Arduino", Springer Science and Business Media LLC, 2021 Publication	2%
2	www.science.gov Internet Source	1%
3	www.ijarcs.info Internet Source	1%
4	Huda Jamal Abdelhameed, Susana Munoz-Hern'andez. "Emotion and opinion retrieval from social media in Arabic language: Survey", 2017 Joint International Conference on Information and Communication Technologies for Education and Training and International Conference on Computing in Arabic (ICCA-TICET), 2017 Publication	1%
5	madm.dfki.de Internet Source	1%
6	Zulfany Erlisa Rasjid, Reina Setiawan. "Performance Comparison and Optimization of Text Document Classification using k-NN and Naïve Bayes Classification Techniques", Procedia Computer Science, 2017 Publication	1%
	www-kiv.zcu.cz	

7	Internet Source	1 %
8	Prabowo, R.. "Sentiment analysis: A combined approach", Journal of Informetrics, 200904 Publication	1 %
9	123doc.org Internet Source	1 %
10	Lecture Notes in Computer Science, 2015. Publication	1 %
11	dl.acm.org Internet Source	1 %

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		