ISSN 2614-0047

Uncovering negative sentiments: a study of indonesian twitter users' health opinions on coffee consumption

Laksono Budiarto [a,1,*], Nissa Mawada Rokhman [a,2], Wako Uriu [b,3]

[a] Universitas Negeri Malang, Jl. Semarang No. 5 Malang 65145, Jawa Timur, Indonesia
[b] Chikushi Jogakuen University,2-chōme-12-1 Ishizaka, Dazaifu, Fukuoka 818-0118, Japan
[1] laksono.budiarto@um.ac.id; [2] nissa.mawarda@um.ac.id; [3] ue2017119@chikushi-u.ac.jp
* corresponding author

ARTICLE INFO

Article history
Received January 24, 2023
Revised February 6, 2023
Accepted February 26, 2023

Keywords
Sentiment analysis
Twitter
Coffee effect
Negative opinion
RapidMiner

ABSTRACT

The increase in coffee consumption among the public is due to several reasons, including health and lifestyle. Awareness of coffee consumption's positive and negative effects has also increased. This research is a sentiment analysis that aims to investigate Twitter users' opinions about the impact of coffee consumption on their health. The method involves data collection using the RapidMiner application, which utilizes the Twitter Application Programming Interface (API) function connected to a prepared Twitter account. The obtained data underwent data cleaning, saved as an Excel file type, training and testing, and model evaluation. Then, the data were classified into Negative, Neutral, and Positive Opinions. The results showed that less than 10% of opinions were positive, 19% were neutral, and 73% were negative. The opinions obtained are useful information for stakeholders in the coffee industry. They can also be used to determine better steps in educating the public about coffee.

1. Introduction

In previous research, it has been found that coffee consumption is generally safe at normal intake levels, based on a summary indicating low to no increased health risk for consumption of three to four cups per day, and may even be more beneficial than harmful to health. Importantly, outside of pregnancy, available evidence suggests that coffee can be tested as an intervention without significant risk of harm [1]. Caffeine, trigonelline, chlorogenic acid, amino acids, carbohydrates, lipids, organic acids, minerals, and volatile aroma compounds are just a few numbers of many chemical components contained in coffee that have both positive and negative impacts on the health of coffee drinkers [2], [3].

Recent research on coffee consumption and consumer purchasing patterns may help us better understand food choices and lifestyles. It is also possible to further improve and establish consumption recommendations and food purchasing behavior by integrating knowledge of food nutrient quality [4]. Evidence shows that consuming coffee may reduce the risk of noncommunicable diseases (NCDs) [1]. Understanding modifiable risk factors through unhealthy dietary patterns could help the World Health Organization (WHO) achieve its goal of reducing the relative risk of premature deaths from NCDs by 25% by025 [5].

Although there have been numerous studies on the general impact of coffee on health, there is still limited discussions on public opinion regarding to the effects of coffee on health. Opinions on the health effects of coffee vary widely among coffee drinkers, depending on their physical health and knowledge of the beverage itself. The same cup of coffee can have different opinions on each individual, not only in terms of taste but also its effects on their body, which sometimes contradicts the facts. According to Liu [6], there are two main types of textual information: facts and opinions.

Current processing techniques (such as search engines) work with facts (assuming the information is true), which can be expressed through topic keywords. However, search engines are not used to search for opinions because opinions are difficult to express with just a few keywords [7]. To mine opinions, it is more appropriate to use a dataset consisting of messages collected from Twitter, which contains a large number of short messages created by users of this microblogging platform [8]. Twitter is a well-known social media platform with 18.45 million users as of January 2022 [9]. Twitter is a popular social media platform widely used to express opinions and thoughts in short messages [10]. The limited character count on Twitter makes it an ideal platform for opinion mining. By analyzing tweets and opinions about coffee, it is possible to determine the sentiment or polarity of these opinions, whether they are positive, negative, or neutral. This information is valuable for understanding how users perceive and feel about coffee. By conducting sentiment analysis, a more accurate picture can be obtained regarding public opinion about coffee, particularly its impact on health. If the results show a higher negative score, education about coffee should be improved by emphasizing more research facts about the effects of coffee on the body. This research was conducted on Twitter posts, both recent and popular posts. The data was collected in October 2022, on Indonesian-language posts.

2. Method

Sentiment analysis, sometimes called opinion mining, is a method or process that classifies text resulting in Natural Language Processing (NLP) [11]. This is a common method of defining and grouping opinions about goods, services, or concepts, involving data mining applications, artificial intelligence (AI), and machine learning (ML) to mine text with a sentiment or subjective meaning. Opinion analysis is one type of study used in social media because its content can become a trending topic and significantly impact social life [12]. Text Mining is one of the main subfields of data mining. Its goal is to uncover previously undiscovered but potentially useful information from semi-structured or unstructured text data [13].

The mining process uses the RapidMiner application, one of the most popular, comprehensive and adaptable data mining tools that can be accessed, with over 400 data mining modules or operators. RapidMiner is an open-source data mining tool that was one of the top three data mining tools overall in 2007 and 2008, according to a survey conducted by the well-known data mining website KDnuggets.com among several hundred data mining specialists. Data loading, preprocessing, visualization, interactive mining process design and review, automatic modeling, automatic parameter and process optimization, automatic feature creation and feature selection, evaluation, and implementation are all supported by RapidMiner. RapidMiner is a powerful platform for mining and analyzing data [14], [15]. With the RapidMiner application, the analysis results will be processed through several modeling methods, one of which is the Naïve Bayes method [16]. The Bayes' theorem forms the basis of the Naive Bayes classification. Classification is a set of algorithms for classification. Naive Bayes classification is based on the idea that the features being classified are all independent of each other [17], [18]. As shown in Fig. 1, the steps of the process can be explained as follows, according to the RapidMiner Studio manual book [19].

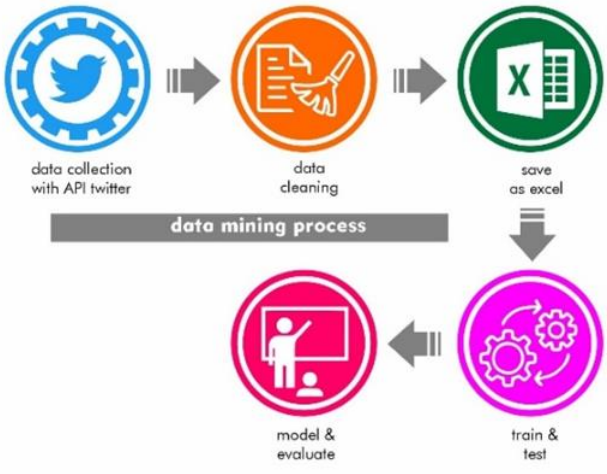

Fig. 1. The Stages of the Sentiment Analysis Process

2.1. Data collection with API Twitter

- Testing the prepared query on the Twitter search section. The obtained results can at least provide a definite picture of the relevance of the resulting message posts to the research objective.
- Selecting Twitter as the connection type to input the Twitter API access code in the token input field.
- Coming into the Twitter search operator, which is located under the Data Access, Application, and Twitter option.
- Select the connection type created in the connection entry option, and enter the prepared query in the query entry field.

2.2. Data Cleaning

- The cleaning process involves using the Select Attributes operator under the Blending, Attributes, and Selection options. Select “single” for the attributes file type option and “Text” for the attributes option.
- Enter the “Remove Duplicates” operator, which is located in the Cleansing, Duplicates option. Select “Single” in the attributes file type options, and “Text” in the attributes option.

2.3. Save as Excell

- The mining result data is saved in an Excel format file using the Write Excel operator located in the Data Access, Files, Write option.
- Specify the filename to save the results by filling in the filename in the Excel file input field.

2.4. Train and Test

- Adding a label column to the Excel file to enter Negative, Neutral, and Positive inputs.
- RapidMiner handles this process by using the XValidation operator. This operator automatically divides the data into various subsets needed for cross-validation. Several sample experiments can be found, including experiments that use XValidation for performance measurement, which are available under the Sample, Repository option.

2.5. Model and Evaluate

- By default, AutoModel uses multi-hold-out-set validation instead of cross-validation to validate the model.
- RapidMiner creates the resulting process after running AutoModel to see how the model’s performance is estimated.

The data obtained through mining consists of the date, time, sender, message content, and other determined attributes. The reading process is done by utilizing the Application Programming Interface (API) service provided by Twitter. Connection entry is done by entering the given API token and then entering the “Search Twitter” operator with the query input “minum kopi” sakit and “minum kopi” aman, result type “recent or popular”. A total of 328 data were obtained. The author did not use the query “minum kopi” sehat as an antonym of sakit because even though the results obtained were more (357 data), the negative value was higher than the positive value. It is because the word “sehat” (healthy) is more often associated with “tidak sehat” (unhealthy) or “kurang sehat” (rather sick).

The process is shown in Fig. 2. It starts with searching data based on the query entered in the search Twitter operator, followed by combining data using the union operator. The next step is to select the attribute to be processed in the text column containing user tweets. The following step is to remove duplicate data that may occur because users only retweet. The resulting data is then saved in Excel format for further processing.

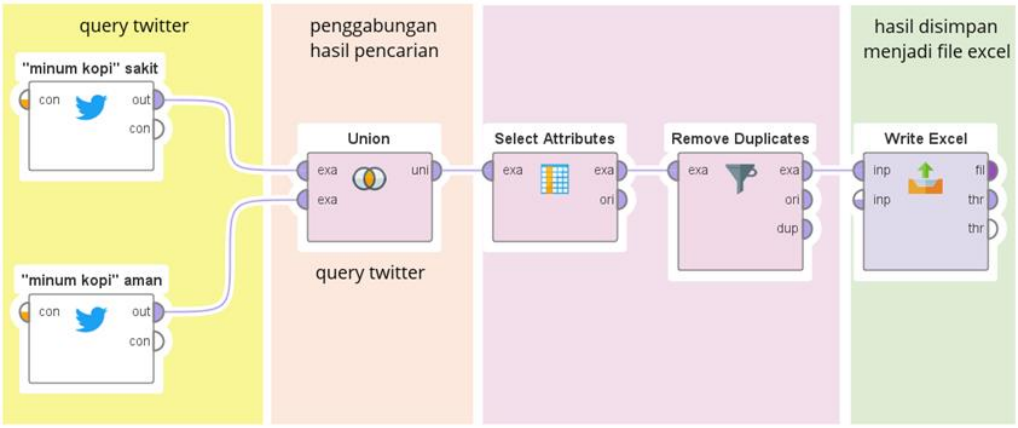

Fig. 2. Data mining process in RapidMiner

To obtain a data comparison, this study also conducted an online survey. Two options were provided in this survey system, a) Drinking coffee makes you sick, and b) Drinking coffee is healthy. The survey will be distributed by sharing a tweet link. This survey was conducted for seven days, in accordance with Twitter's rules [20].

The polling was conducted by creating a tweet containing a message introducing the purpose and objective of the survey. Then, the poll feature was selected, options for the survey were filled in, and the maximum duration of the survey, which could last up to 7 days, was determined. The process can be seen in Fig. 3.

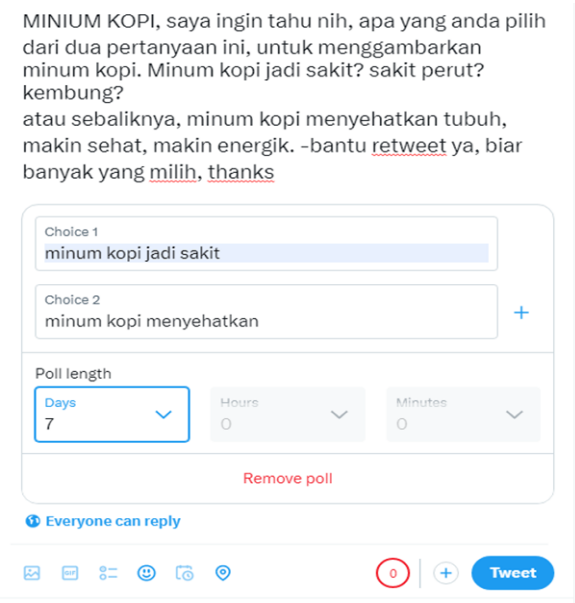

Fig. 3. The Process of Creating a Poll on a Twitter's Tweet

3. Results and Discussion

In Table 1, there is an opinionated statement that drinking coffee has a negative impact on health. The text part of the sentence explicitly states that the cause of illness is drinking coffee.

Table.1 Data Classification of Negative Opinions

ID	Text	Label
1585067228881641472	*jangan minum kopi pagi2 sakit perut wkwk*	Negative
1584980056065339392	*yaolo yaolo niat minum kopi biar ga ngantuk eh malah sakit perut wkwk*	Negative
1584976005495943168	*ih asik banget kakak bisa minum kopi. aku sekalinya nyeruput kafein langsung sakit perut masa, kak. ??*	Negative

Meanwhile, Table 2 shows that the sentences are categorized as neutral statements and are mostly dominated by questions.

Table.2 Data classification of neutral opinions

ID	Text	Label
1583559221081440256	*kalo bsk pagi gue minum kopi pait kr2 aman gak yaa,,*	Neutral
1583459347073445890	*maybe lebih aman makanan yaa. ga usah mahal mahal gapapa ko.*	Neutral
1583345106140397569	*soalnya blm tentu bisa minum kopi. iyaa, kadang malah minum kopi dari biji salak. kalo itu masih aman.*	Neutral

In Table 3, there are affirmations that drinking coffee is safe and categorized as positive statements.

Table.3 Data Classification of Positive Opinions

ID	Text	Label
1584774800689405953	*gue pasti sama dia sama-sama minum kopi terus? aman sejauh ini! lagi dan selalu berusaha diimbangi dengan makan terus.*	Positive
1584724633646817280	*pagiku aman kalau udah sarapan plus minum kopi.*	Positive
1584462034921754624	*yg pasti perut udh diisi dulu sm makanan si ceu biar lambung aman. abis itu mangga klo mau ngopi. biar gak insomnia jangan minum kopi diatas jam 6 sore si kyknya*	Positive

In Table 4, show the results of several modeling.

Table.4 The Results of Several Modeling

Model	Classification Error	Standard Deviation	Gains	Total Time	Training Time (1,000 Rows)	Scoring Time (1,000 Rows)
Naive Bayes	0.4	0.0	0.0	999.0	149.4	4297.7
Generalized Linear Model	0.4	0.1	0.0	4199.0	6957.3	6694.7
Fast Large Margin	0.4	0.1	-6.0	2464.0	122.0	6251.9
Decision Tree	0.4	0.0	0.0	1125.0	149.4	3404.6
Random Forest	0.4	0.0	-2.0	3661.0	280.5	3587.8
Gradient Boosted Trees	0.3	0.0	14.0	377175.0	3192.1	2511.5
Support Vector Machine	0.3	0.0	2.0	1945.0	332.3	3709.9

The Naïve Bayes sentiment analysis result can be seen in Fig. 4, where the positive opinion is minimal, below 10%, while the neutral opinion is 19%, and the negative opinion is 73%. The words “tiap, makin, and mood” are part of the support for the negative opinion. Meanwhile, contradicting negative opinion is obtained from the words “dada, asli, beli, tidur”. For example, it can be seen in the sentence "anjir tiap minum kopi malem2 pasti langsung mual + sakit perut??" from ID "1583154503926575105"., “Kalo stress minum kopi aja kali yah..biar makin sakit!!!” from ID ”1583567191580246016", become part of supports negative opinion, while those that become part of contradicts negative opinion, in the sentence "Hari-hari minum kopi malah mules, biasanya deg-degan, dada sakit, sekarang mules wkaakakaka" from ID "1582013788210876416", "udah tau ga bisa minum kopi pake ngide segala beli kopi, akhirnya sakit kan perutnya:(“ from ID “1582321855889043463”.

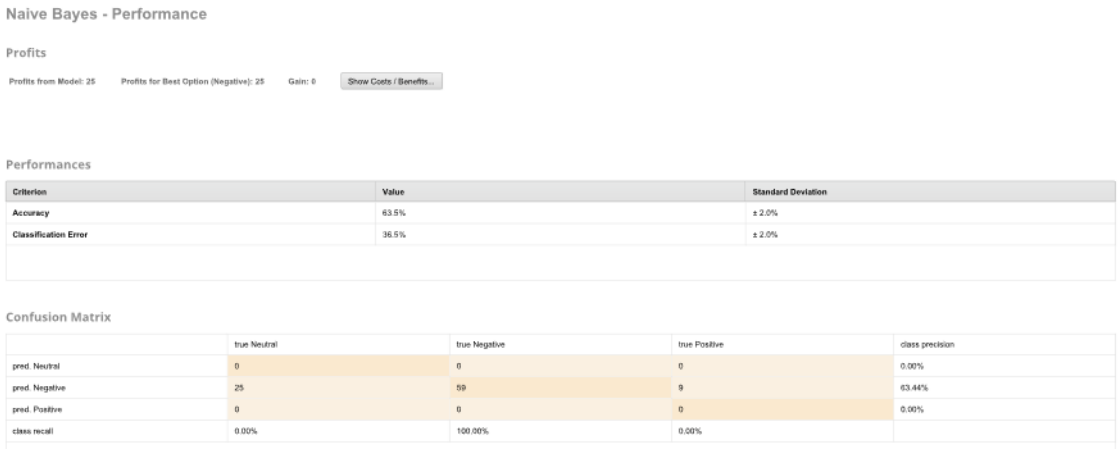

Fig. 4. Diagram and factors of negative opinion

The Naïve Bayes performance as seen in Fig. 5

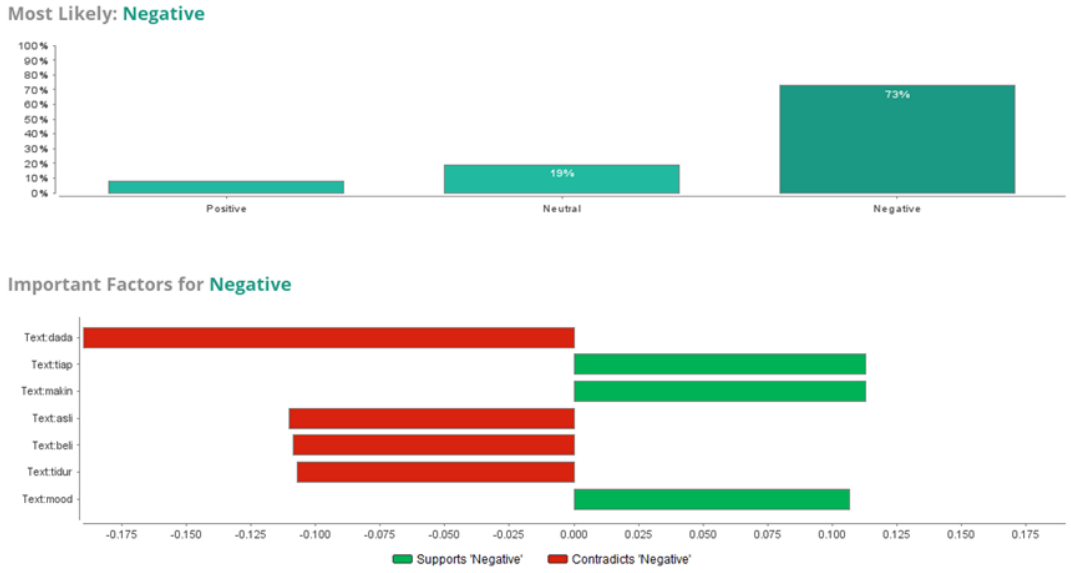

Fig. 5. Naïve Bayes – Performance

As a comparison, an online survey was conducted in this study. As seen in Fig. 6, the result of choosing drinking coffee as healthy obtained a value of 67.7%, much larger than the choice of drinking coffee causing illness at 32.3%. The survey participants amounted to 167 respondents, starting from December 13, 2022, until December 20, 2022.

Fig. 6. The results of the opinion poll on Twitter

These different results may be influenced by the limited scope of dissemination, which tends to be obtained from the account's circle of friends, as well as the respondents' subjective level towards the poll creator's profile.

4. Conclusion

The results obtained from this sentiment analysis provide an overview of the imbalance between coffee education, the promotion of coffee benefits, and the growing negative opinion regarding the impact of coffee on health. It may also be due to the increasing number of coffee shops that do not apply proper serving methods and the importance of coffee education by baristas to customers. These possibilities also serve as suggestions for further research development, and the results of online polls can be improved with wider dissemination.

References

[1] R. Poole, O. J. Kennedy, P. Roderick, J. A. Fallowfield, P. C. Hayes, and J. Parkes, "Coffee consumption and health: umbrella review of meta-analyses of multiple health outcomes," *BMJ*, vol. 359, p. j5024, Nov. 2017, doi: 10.1136/bmj.j5024.

[2] J. V. Higdon and B. Frei, "Coffee and Health: A Review of Recent Human Research," *Crit. Rev. Food Sci. Nutr.*, vol. 46, no. 2, pp. 101–123, Mar. 2006, doi: 10.1080/10408390500400009.

[3] F. Bastian *et al.*, "From Plantation to Cup: Changes in Bioactive Compounds during Coffee Processing," *Foods*, vol. 10, no. 11, p. 2827, Nov. 2021, doi: 10.3390/foods10112827.

[4] A. Samoggia and B. Riedel, "Consumers' Perceptions of Coffee Health Benefits and Motives for Coffee Consumption and Purchasing," *Nutrients*, vol. 11, no. 3, p. 653, Mar. 2019, doi: 10.3390/nu11030653.

[5] WHO, "Global action plan for the prevention and control of noncommunicable diseases 2013-2020.," *World Heal. Organ.*, p. 102, 2013, [Online]. Available at: https://www.who.int/publications/i/item/9789241506236.

[6] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, Aug. 2004, pp. 168–177, doi: 10.1145/1014052.1014073.

[7] G. Uddin and F. Khomh, "Automatic Mining of Opinions Expressed About APIs in Stack Overflow," *IEEE Trans. Softw. Eng.*, vol. 47, no. 3, pp. 522–559, Mar. 2021, doi: 10.1109/TSE.2019.2900245.

[8] M. Imran, P. Mitra, and C. Castillo, "Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages," *Proc. 10th Int. Conf. Lang. Resour. Eval. Lr. 2016*, pp. 1638–1643, May 2016, Accessed: Jun. 22, 2023. [Online]. Available: https://arxiv.org/abs/1605.05894v2.

[9] A. O. P. Dewi, A. Isnaini, and M. F. Lestari, "An Analysis of the post-COVID-19 Information Distribution on Social Media," *E3S Web Conf.*, vol. 359, p. 02037, Oct. 2022, doi: 10.1051/e3sconf/202235902037.

[10] N. Öztürk and S. Ayvaz, "Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis," *Telemat. Informatics*, vol. 35, no. 1, pp. 136–147, Apr. 2018, doi: 10.1016/j.tele.2017.10.006.

[11] C. J. Rameshbhai and J. Paulose, "Opinion mining on newspaper headlines using SVM and NLP," *Int. J. Electr. Comput. Eng.*, vol. 9, no. 3, p. 2152, Jun. 2019, doi: 10.11591/ijece.v9i3.pp2152-2163.

[12] S. Sendari, I. A. E. Zaeni, D. C. Lestari, and H. P. Hariyadi, "Opinion Analysis for Emotional Classification on Emoji Tweets using the Naïve Bayes Algorithm," *Knowl. Eng. Data Sci.*, vol. 3, no. 1, pp. 50–59, Aug. 2020, doi: 10.17977/um018v3i12020p50-59.

[13] Z. Ding, Z. Li, and C. Fan, "Building energy savings: Analysis of research trends based on text mining," *Autom. Constr.*, vol. 96, pp. 398–410, Dec. 2018, doi: 10.1016/j.autcon.2018.10.008.

[14] M. Bjaoui, H. Sakly, M. Said, N. Kraiem, and M. S. Bouhlel, "Depth insight for data scientist with RapidMiner « an innovative tool for AI and big data towards medical applications»," in *Proceedings of the 2nd International Conference on Digital Tools & Uses Congress*, Oct. 2020, pp. 1–6, doi: 10.1145/3423603.3424059.

[15] P. Ristoski, C. Bizer, and H. Paulheim, "Mining the Web of Linked Data with RapidMiner," *J. Web Semant.*, vol. 35, pp. 142–151, Dec. 2015, doi: 10.1016/j.websem.2015.06.004.

[16] Z. E. Rasjid and R. Setiawan, "Performance Comparison and Optimization of Text Document Classification using k-NN and Naïve Bayes Classification Techniques," *Procedia Comput. Sci.*, vol. 116, pp. 107–112, Jan. 2017, doi: 10.1016/j.procs.2017.10.017.

[17] K. Yadav and R. Thareja, "Comparing the Performance of Naive Bayes And Decision Tree Classification Using R," *Int. J. Intell. Syst. Appl.*, vol. 11, no. 12, pp. 11–19, Dec. 2019, doi: 10.5815/ijisa.2019.12.02.

[18] M. Singh, M. Wasim Bhatt, H. S. Bedi, and U. Mishra, "WITHDRAWN: Performance of bernoulli's naive bayes classifier in the detection of fake news," in *Materials Today: Proceedings*, Dec. 2020, p. 1, doi: 10.1016/j.matpr.2020.10.896.

[19] RapidMiner, *RapidMiner Studi o Manual*, p.116, 2012. [Online]. Available at: https://docs.rapidminer.com/downloads/RapidMiner-v6-user-manual.pdf.

[20] Twitter, "Twitter Polls – how to create and how to vote.". [Online]. Available at: https://help.twitter.com/en/using-twitter/twitter-polls.