JOTECHCOM Journal of Technology and Computer
Vol. 1, No. 2, May 2024, pp. 7~11
ISSN: 3048-0477 (Media Online)

Sentiment analysis on police brigadier shooting case using K-means clustering

Masri Wahyuni [1], Basyit Mubarroq Rambe [2], Yuni Franciska Br Tarigan [3], Karyawaty Gultom [4]
[1234] Akademi Manajemen Informatika dan Komputer (AMIK) Polibisnis
[1] masriwahyuni997@gmail.com, [2] boyrambe@gmail.com, [3] yuni.franciska@gmail.com, [4] gkaryawaty@gmail.com

Article Info

Article history:

Received April 18, 2024
Revised April 25, 2024
Accepted May 04, 2024

Keywords:

Sentiment analysis
Twitter
Topic modeling
Latent dirichlect allocation
K-means clustering.

ABSTRACT

As a medium that can be used to convey public criticism and aspirations in real time, Twitter is used as a data collection source using crawling techniques, to analyze public sentiment or response to the police brigadier shooting case. Latent dirichlet al-location (LDA) is used to determine the topics that appear in each of the collected tweets, and then used as a feature in grouping the contents of the tweets based on their respective sentiment values. The results of clustering using the k-means clustering algorithm obtained are: 11.9% of netizens gave a positive response, 18.9% of netizens gave a neutral response and 69.2% of netizens gave a negative response to the case. Thus, from the results of this study it can be concluded that netizens tend to give a negative response or reaction to the police brigadier shooting case, when viewed from the percentage of each type of response.

Corresponding Author:

Masri Wahyuni
Akademi Manajemen Informatika dan Komputer (AMIK) Polibisnis
Email: masriwahyuni997@gmail.com

1. INTRODUCTION

Twitter is a social media that can be utilized by its users to convey criticism or aspirations on a topic of social problems easily and in real time [1]. Twitter provides an application program interface (API) that can be used to crawl tweets on social media, using certain search keywords, so that a set of data is obtained in the form of a dataset of tweet content [2]. This set of tweets can be analyzed to see the public's response to the topic of the issue raised, which includes responses that are positive, neutral, or negative, based on the sentiment value in each tweet [3]. In this study, a sentiment analysis was conducted on the shooting case of a police brigadier, which was then classified based on the results of the sentiment value grouping, to see how netizens, especially Twitter social media users, responded to the case.

K-means is one of the clustering algorithms that works by partitioning data into several clusters based on the closeness of the characteristics of each data [4]. Several studies that apply this algorithm to sentiment analysis problems such as responses to PSBB policies [5], responses to BPJS services [6], and responses to opinions on beach tourism [7], show that the k-means algorithm can be used to group sentiment values in the content of tweets into groups of positive responses, neutral responses, and negative responses. The tweet dataset collected in this study is grouped based on its sentiment value, by adding other features obtained from the topic modeling results. Topic modeling is a technique for discovering, expressing and labeling the thematic structure of a document or text, using a model consisting of a set of algorithms [8]. One of the popular topic modeling methods used in sentiment analysis is latent dirichlet allocation (LDA) [9], which works by indexing semantic structures based on probability levels in the processed text. By using LDA, opinion classes can be formed based on topics that arise from the extraction of a set of texts [10] as shown in the sentiment analysis research on the level of contributions to BPJS Health [11], sentiment analysis on tropical diseases in Indonesia [12], and

sentiment analysis on Gojek customer complaints, where the topics generated are quite accurate as a representation of the content of the tweets collected. In this study, the results of topic modeling using the LDA method are used as a feature in grouping netizen responses to the police brigadier shooting case, based on positive, neutral and negative response categories.

2. METHOD

Dataset: This research uses datasets obtained from twitter social media. Data collection is done using the Twitter API which automatically crawls tweets containing the given search keywords. In this study, the search keyword "brigadier j" is used, according to the hashtag of this problem that has been trending on Twitter social media.

Sentiment Analysis: The sentiment analysis conducted in this study is to calculate the sentiment value of each tweet collected, determine the top 10 topics based on the content of each tweet, and group the results into a classification of netizen responses to the topic of the police brigadier shooting problem. In this study, three response categories were used to assess the sentiment of netizens [14]:

1. Positive sentiment: Sentiment where the content of the tweet is supportive or does not deny the topic of the problem.
2. Neutral sentiment: Sentiment where the content of the tweet does not contain an opinion or does not choose a partisan attitude towards the topic of the problem
3. Negative sentiment: Sentiment where the content of the tweet does not support or deny the topic of the problem. The initial stage in this sentiment analysis is pre-processing the text (tweet content), using four steps:
4. Transformation: The transformation step is carried out by converting all text content into lowercase letters, changing characters that have accent shapes, eliminating html format, and removing links or links from the contents of the tweet.
5. Tokenization: The tokenization step is carried out to break down the contents of the tweet into smaller components such as words or bigrams [15]. In this study, tokenization is carried out using the help of the regular expression model (Regexp).
6. Filtering: The filtering step is often referred to as stopwords removal, which is a process for selecting words by removing unnecessary words, so that only words that are important to the topic of the problem are retained. In this research, a TXT file is used which contains stopwords that will be removed from the contents of the tweet [16], and also the removal of words containing numbers or numeric.

Topic Modeling: As a feature that will be used in the clustering process, a model is formed to determine the topics contained in each tweet content. This process uses the latent dirichlet allocation (LDA) method that forms topics based on the words in the tweet content where the level is determined based on the level of cohesion of the words in the tweet content and calculated using equation (1) [17].

$$P(W, Z, \emptyset, \varphi; \alpha, \beta) = \prod_{i=1}^{k} p(\alpha_i; \beta) \prod_{j=1}^{m} P(\emptyset_i; \alpha) \prod_{t=1}^{n} P\left(Z_{j,t}|\emptyset_j\right) P\left(W_{j,t}|\alpha Z_{j,t}\right) \quad (1)$$

Where:

M : Number of tweets
N : Number of words in the tweet
A : Dirichlet prior parameter on topic distribution per tweet
B : Dirichlet prior parameter in the distribution of words per topic
∅_i : Distributions of topics per tweet
∅_j : Word distribution per topic
Z_(j,t) : Topic for jth word in tth tweet

From the calculation results using equation (1), the top 10 topics in each tweet content are taken as features in the clustering process using the k-means clustering algorithm.

K-Means clustering: K-means clustering is used in two processes, namely determining the type of netizen response based on the sentiment value of the tweet content and clustering the response category of each tweet content. For the process of determining the type of netizen response based on the sentiment value of the tweet content, the steps taken are as follows:

1. Determine the value of K: Since the desired response types are positive response, neutral response and negative response, the K value used in this stage is 3.
2. Randomly generate the centroid value for each cluster.
3. Calculate the distance between the sentiment value and the cluster center point

The distance calculation at this stage uses the euclidean distance formula, as shown in equation (2).

$$d_v = \sqrt{\sum_{i=1}^{n}(x_i y_i)^2} \quad (2)$$

4. Group each sentiment value based on the closest distance to the centroid.
5. Calculate the new centroid value based on the number of sentiment values in each cluster
6. Repeat the process 3 to 5 until the data in each cluster does not change.
7. Calculate the most sentiment value in each cluster

In this process, the number of each sentiment value (positive, 0 or negative sentiment) in each cluster is calculated. The sentiment with the highest number is used as a category of netizen response types in that cluster. Similarly, for the process of grouping response categories on problem topics based on 10 predetermined topics, the same process is used as determining the type of netizen response.

3. RESULTS AND DISCUSSION

From the crawling results using the Twitter widget with the keyword "Brigadier J", 1000 tweets were obtained which were used as datasets. These results are saved into a file using the Save Data widget with XLSX format. The saved file is then opened using the Corpus widget, and by using the Corpus Viewer widget, the contents of the tweets in the dataset can be seen.

Content:	Pengacara keluarga Brigadir Nofryansah Yosua Hutabarat atau Brigadir J, Kamaruddin Simanjuntak mengu #berita #menarik https://t.co/CK119xZvt7
Content:	Tunjukkan presisimu om @ListyoSigitP ,karena ini menyangkut tewasnya anak buahmu Brigadir J,..bukankah dalam suatu corp itu,anak buah sakit,sakitlah seluruh corp,.lah ini kan ampe tewas anak buahmu,.presisikan keadilanmu niscaya akan semakin kuat dan solid corp anda https://t.co/jjriVvkOOa
Content:	@DivHumas_Polri Sorry aku tidak lagi sobatmu. Tuntaskan dulu kasus brigadir j
Content:	@mandiminak @PRFMnews Isilop lagi fokus memecahkan kasus Brigadir J.
Content:	Para ahli ahli politik mulai mengeluarkan pendapatnya sejak kasus Brigadir J, aku iya iya kan aja laa
Content:	Brigadir J Ditembak Gara-Gara Kejadian di Magelang? Ada Cerita dan Fotonya https://t.co/okGfdM7DmS
Content:	@_melody_mellow @Love1Dhya_ @UmarBaagil5 @putra_jamparing @ComebackU25 @mamamamabiasa @AlbarraBack @Arini_Nat @Mantan_411 @Cherry_ijo Jalan panjang utk kasus Brigadir J msh berliku sepertinya. □□□□

Figure 1: Sample tweet content

Table I: Statistik number of likes

Statistik	Value
Total	2180
Mean	2.18
Dispersion	8,76
Min	0
Total Min	680
Max	554
Total Max	1

The obtained tweets contained unnecessary features, so of the many features in the tweets, only the Content feature, which contains comments from netizens, and the Number of Likes feature, which contains the number of positive responses from netizens to the tweets, were retained. The Number of Likes feature was analyzed to see the statistics of this feature on the dataset.

The Text Preprocessing widget provided by Orange 3 was used to process the dataset through the stages of transformation, tokenization, and filtering. The result of this process was a change in the number of

tokens from 27686 to 11656 tokens. The results of this process are the words used in sentiment analysis, in the form of a word cloud. From the resulting word cloud, 10 words with the largest number of weights were obtained, namely brigadier, ham, komnas, sambo, ferdy, police, autopsy, police, repeat, and death.

Table II: Pre-processed token

Stages	Token
Tranformasi – Lowercase	27686
Tranformasi – Remove accents	27867
Tranformasi – Parse html	27577
Tranformasi – Remove urls	21549
Tokenization – Regexp	18140
Filtering – Stopwords Bahasa Indonesia	11863
Filtering – Remove numbers	11656

The sentiment values in the following table are clustered using the k-means clustering algorithm, to determine the type of tweet response whether it is a positive, neutral, or negative response. The clustering results of these sentiment values are then analyzed to see the most sentiment in each cluster, which is used as a reference to determine the type of response from each cluster.

Table III: Determination of type *cluster*

Cluster	Positive Sentiment	Neutral Sentiment	Negative Sentiment	Response Type
C1	70	391	106	Neutral
C2	134	0	0	Positive
C3	0	0	299	Negative

The topic value 1 to topic 10 in each tweet content and the Number of Likes value in each tweet are used as features in the clustering process using the k-means clustering algorithm. In this clustering process, three clusters are used, namely C1, C2 and C3 to group each tweet based on its response to the police brigadier shooting issue.

Table IV: Final dataset

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topik 6	Topic 7	Topic 8	Topic 9	Topikc 10	Total Likes
0.887	0.012	0.012	0.012	0.012	0.012	0.012	0.012	0.012	0.0125	1
0.522	0.416	0	0	0	0	0	0	0	0	0
0	0.955	0	0	0	0	0	0	0	0	0
0.02	0.02	0.02	0.02	0.02	0.02	0.82	0.02	0.02	0.02	0
0	0	0	0	0	0	0	0.918	0	0	0
…	…	…	…	…	…	…	…	…	…	…
0	0	0	0	0	0	0	0.962	0	0	0
0.582	0	0	0	0	0	0.344	0	0	0	0
0	0.584	0	0	0	0	0	0.348	0	0	0
0.924	0	0	0	0	0	0	0	0	0	2
0	0.819	0	0	0	0	0	0.113	0	0	4

DISCUSSION

From the results of the sentiment value grouping shown in Table, it can be seen that 12.34% of cluster C1 members are positive sentiments, 68.96% are neutral sentiments, and 18.7% are negative sentiments. From the percentage of members in cluster C1, the neutral response type results are obtained, according to the percentage of the most sentiment values. For cluster C2 members, it can be seen that 100% are positive sentiments, so a positive response type is obtained. For members of cluster C3, it can be seen that 100% are negatig sentiments, so a negative response type is obtained.

From the topic modeling results shown in Table VIII, it can be seen that the word "brigadier" is a word that always appears as a keyword for all topics, followed by the words "police", "ferdy", and "sambo", which each appear in 5 topics. These results show that LDA succeeded in building a model that contains topics relevant to the problem, where the word "brigadier" is the keyword used in the crawling process, the word "police" is the institution where the police brigadier appointed in this case works, and the words "ferdy" and

"sambo" are the names of the shooting victim's superiors. From the results of grouping netizen responses, 11.9% of netizens gave a positive response to this case, 18.9% of netizens gave a neutral response, and 69.2% of netizens gave a negative response.

4. CONCLUSION

By using sentiment values obtained from the results of pre-processing and sentiment analysis, the appropriate type of response to a problem topic can be obtained, by grouping these values using k-means clustering and the largest number of sentiment values that become members of each cluster. The accuracy of the LDA method in this study in forming topics related to the content of the tweets collected is quite accurate, it can be seen from the 10 topics generated, the keywords with the highest frequency of occurrence are in accordance with the scope of the problem in question. These generated topics can then be used as features to group the dataset into analyzed netizen responses, with the result that most netizens have a negative reaction to the shooting of the police brigadier. Based on the overall results of the study, it can be concluded that the combination of topic modeling LDA method and k-means clustering algorithm proved to be able to answer the research problem in the form of the percentage of netizen responses to the shooting case that occurred.

REFERENCES

[1] S. Mandasari, B. H. Hayadi, and R. Gunawan, "Analisis Sentimen Pengguna Transportasi Online Terhadap Layanan Grab Indonesia Menggunakan Multinomial Naive Bayes Classifier," *J. Teknol. Sist. Inf. dan Sist. Komput. TGD*, vol. 5, no. 2, pp. 118–126, 2022.

[2] R. N. Fahmi, Nursyifa, and A. Primajaya, "ANALISIS SENTIMEN PENGGUNA TWITTER TERHADAP KASUS PENEMBAKAN LASKAR FPI OLEH POLRI DENGAN METODE NAIVE BAYES CLASSIFIER," *JIKO (Jurnal Inform. dan Komputer)*, vol. 5, no. 2, pp. 61–66, 2021, [Online]. Available: https://ejournal.akakom.ac.id/index.php/jiko/article/view/437.

[3] M. I. Zul, F. Yulia, and D. Nurmalasari, "Social media sentiment analysis using K-means and naïve bayes algorithm," *Proc. - 2018 2nd Int. Conf. Electr. Eng. Informatics Towar. Most Effic. W. Mak. Deal. with Futur. Electr. Power Syst. Big Data Anal. ICon EEI 2018*, no. February, pp. 24–29, 2018, doi: 10.1109/ICon-EEI.2018.8784326.

[4] R. Rahmati and A. W. Wijayanto, "ANALISIS CLUSTER DENGAN ALGORITMA K-MEANS, FUZZY C-MEANS DAN HIERARCHICAL CLUSTERING," *JIKO (Jurnal Inform. dan Komputer)*, vol. 5, no. 2, pp. 73–80, 2021.

[5] G. F. Santoso *et al.*, "Respon masyarakat terhadap kebijakan psbb sebagai penekan angka covid-19," *J. Inform. dan Komput.*, vol. 5, no. 2, pp. 38–48, 2021, [Online]. Available: https://ejournal.akakom.ac.id/index.php/jiko/article/view/401.

[6] A. B. Saputra, P. W. Cahyo, M. Habibi, and A. Priadana, "Analysis and visualization of BPJS on twitter using K-means clustering," *Int. J. Heal. Sci. Technol.*, vol. 3, no. 3, pp. 109–117, 2022, [Online]. Available: https://ejournal.unisayogya.ac.id/index.php/ijhst/article/view/2466.

[7] Y. W. Syaifudin and R. A. Irawan, "Implementasi Analisis Clustering Dan Sentimen Data Twitter Pada Opini Wisata Pantai Menggunakan Metode K-Means," *J. Inform. Polinema*, vol. 4, no. 3, p. 189, 2018, doi: 10.33795/jip.v4i3.205.

[8] P. Kherwa and P. Bansal, "Topic Modeling: A Comprehensive Review," *ICST Trans. Scalable Inf. Syst.*, vol. 7, no. 24, p. 159623, Jul. 2018, doi: 10.4108/eai.13-7-2018.159623.

[9] V. S. Anoop and S. Asharaf, "Aspect-oriented sentiment analysis: A topic modeling-powered approach," *J. Intell. Syst.*, vol. 29, no. 1, pp. 1166–1178, 2020, doi: 10.1515/jisys-2018-0299.

[10] N. L. P. M. Putu, Ahmad Zuli Amrullah, and Ismarmiaty, "Analisis Sentimen dan Pemodelan Topik Pariwisata Lombok Menggunakan Algoritma Naive Bayes dan Latent Dirichlet Allocation," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 123–131, 2021, doi: 10.29207/resti.v5i1.2587.

[11] T. D. Dikiyanti, A. M. Rukmi, and M. I. Irawan, "Sentiment analysis and topic modeling of BPJS Kesehatan based on twitter crawling data using Indonesian Sentiment Lexicon and Latent Dirichlet Allocation algorithm," *J. Phys. Conf. Ser.*, vol. 1821, no. 1, 2021, doi: 10.1088/1742-6596/1821/1/012054.

[12] D. Ridhwanulah and D. H. Fudholi, "Pemodelan Topik pada Cuitan tentang Penyakit Tropis di Indonesia dengan Metode Latent Dirichlet Allocation," *J. Ilm. SINUS*, vol. 20, no. 1, p. 11, 2022, doi: 10.30646/sinus.v20i1.589.

[13] Hariyady, S. Basuki, and L. Meidina, "IMPLEMENTASI ALGORITMA DETEKSI TOPIK KELUHAN PELANGGAN JASA OJEK ONLINE BERDASARKAN KOMENTAR MEDIA SOSIAL," *Semin. Nas. Teknol. dan Rekayasa*, vol. 5, pp. 160–166, 2019, [Online]. Available: http://research-report.umm.ac.id/index.php/sentra/article/view/3297.

[14] R. Sanusi, F. D. Astuti, and I. Y. Buryadi, "Analisis sentimen pada twitter terhadap program kartu pra kerja dengan recurrent neural network," *JIKO (Jurnal Inform. dan Komputer)*, vol. 5, no. 2, pp. 89–99, 2021.

[15] S. Chohan, A. Nugroho, A. M. B. Aji, and W. Gata, "Analisis Sentimen Pengguna Aplikasi Duolingo Menggunakan Metode Naïve Bayes dan Synthetic Minority Over Sampling Technique," *Paradig. - J. Komput. dan Inform.*, vol. 22, no. 2, pp. 139–144, 2020, doi: 10.31294/p.v22i2.8251.

[16] D. Haryalesmana and M. Wieriks, "ID-Stopwords," 2019. [Online]. Available: https://github.com/masdevid/ID-Stopwords.

[17] E. S. Negara and D. Triadi, "Topic modeling using latent dirichlet allocation (LDA) on twitter data with Indonesia keyword," *Bull. Soc. Informatics Theory Appl.*, vol. 5, no. 2, pp. 124–132, 2021, [Online]. Available: https://pubs.ascee.org/index.php/businta/article/view/455.

[18] T. I. Saputra and R. Arianty, "IMPLEMENTASI ALGORITMA K-MEANS CLUSTERING PADA ANALISIS SENTIMEN KELUHAN PENGGUNA INDOSAT," *J. Ilm. Inform. Komput.*, vol. 24, no. 3, pp. 191–198, Dec. 2019, doi: 10.35760/ik.2019.v24i3.2361.