Indonesian Journal on Data Science
ISSN 2987-7423 - Vol. 1, No. 1, Mei 2023, pp. 32-38

Sentiment Analysis Related National Social Security Agency for Employment in Indonesia: Hybrid Method Using Lexicon Based and Naive Bayes Classifier Approaches

Rizky Fauzi Akbar [a,1], Muhammad Habibi [b,2,*]

[a,b]Departement of Informatics, Universitas Jenderal Achmad Yani Yogyakarta, Yogyakarta, Indonesia
[1] fauziakbar28317@gmail.com; [2] muhammadhabibi17@gmail.com*

* corresponding author

ABSTRACT

The National Social Security Agency (BPJS) for Employment is the Social Security Administering Agency with the goal of ensuring that each participant or member of the family receives adequate necessities. In its implementation, there is information that is spread, particularly on Twitter, regarding the Ministry of Health's decision, namely regarding Old Age Security (JHT), which can only be distributed/taken after the participant turns 56 years old, causing both pros and cons among the public. Based on unanalyzed tweets on Twitter, it is necessary to do extensive research to collect relevant information based on netizens' viewpoints. This research describes sentiment analysis of tweets from Twitter using the terms JHT, BPJSTK, and BPJS, which yield 4154 data tweets. We employ two approaches in this study: Lexicon Based and Nave Bayes Classifier. According to this study, the accuracy of the testing data is 92% for the Lexicon Based and 95% for the Nave Bayes Classifier. This study concluded that the JHT at BPJS Employment received unfavorable attitudes and negative reactions among users who addressed the rejection of new restrictions where JHT, could only be dispensed or taken when participants at BPJS Employment were 56 years old.

ARTICLE INFO

Article history
Received: 8 Maret 2023
Revised: 28 April 2023
Accepted: 10 Mei 2023

Keywords
Sentiment Analysis
Lexicon Based
Naïve Bayes Classifier
BPJS Ketehagakerjaan
Text Mining

1. Introduction

The Republic of Indonesia's 1945 Constitution states in Article 28 H, paragraph (3), "Every person has the right to social security that allows him/her to fully develop into a human being with dignity." In addition, Article 34, Paragraph 2, of the 1945 Constitution states: "The state shares a social security system for all people and empowers people who are weak & unable in accordance with human dignity." National Social Security Agency (BPJS) for Employment was established by Law No. 24 of 2011 concerning Social Security Administering Bodies with the objective of implementing collateral awards to meet each participant's or family member's reasonable basic needs. Up to this point, PT Jamsostek (Persero) has offered all employees and their families 4 (four) protection programs, including the Work Accident Benefit Program (JKK), Death Benefit (JKM), Old Age Benefit (JHT), and Pension Benefit (JP) [1].

The decision of the Ministry of Health regarding JHT, which can only be disbursed/taken after the Labor BPJS participant turns 56 years old, is one of many things that are new problems for the community even though it has a positive goal. As a result, the pros and cons that exist among the public are poured in the form of tweets on Twitter. Twitter is a well-liked social media platform among the public since it allows users to share thoughts and receive up-to-date information. JHT, which can only be disbursed/withdrawn once the Employment BPJS participant turns 56 years old, is one example of information on an updated topic.

Based on this issue, research is required to produce opinions from the public in the form of tweets and retweets regarding BPJS Ketenagakerjaan JHT. Following analysis of the positive and negative feedback from the public, tweet data on BPJS JHT will provide insight into the sentiment on Twitter about BPJS Employment JHT. To categorize the positive and negative feelings towards JHT, this study will employ the Lexicon Based and Nave Bayes Classifier techniques. Using a confusion matrix, the accuracy of the categorization findings will be determined. The findings of this study are intended to aid in the investigation of both positive and negative feelings regarding JHT on Twitter.

There has already been study done on the Indonesian Social Insurance Administration Organization, such as research on the Public Sentiments Analysis of the Indonesian Social Insurance Administration Organization using Twitter data [2]–[4]. Furthermore, the lexicon-based method is a feature extraction method that has the potential to increase system performance. Lexicon-based research is also used in Mapping Twitter hate speech towards social [5] and sexual minorities and Sentiment Analysis for E-Commerce Product Reviews in Chinese [6]. The research was carried out using a combination of Lexicon-based methodologies to simplify the process of manually creating training data. It is hoped that by using this strategy, the data training process will be speedier and yield better training data.

2. Method

2.1. Data Collecting

Figure 1 depicts the research stages that were employed. Between February 14 and July 12, 2022, we used the data from the Twitter streaming API for this study. Using the keywords BPJSTK, BPJS, and JHT throughout that time, we were able to gather 30.000 tweets about the National Social Security Agency for Employment in Indonesia. The Twitter streaming API to access the full collection of tweet properties provides an interface.

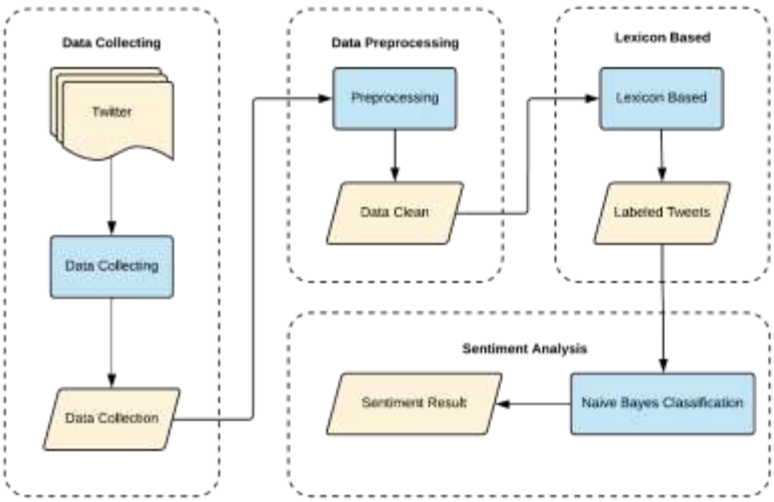

Fig. 1. Research Stages

2.2. Data Preprocessing

Data preparation comes next after data capture. Raw tweets typically contain a lot of noise, typos, and nonsensical phrases, as well as a variety of acronyms and slang terms. These words frequently conflict with the resulting tweet sentiment and hurt the categorization model's performance. Therefore, before extracting features from tweets, they must first be pre-processed. We execute the following preparation operations on the tweet data that will be processed [7]: (1) Tokenizing, dividing, or separating characters in a text, whether such letters are recognized as word separators. (2) Cleaning, remove unneeded elements like hashtags, digits, emoji, and urls from documents that have been received. (3) Case folding, changing from the original form to the conventional form (lowercase). (4) Remove the stopword Words with little information should be eliminated from texts. (5) Stemming, the conversion of all the words into their most fundamental forms. (6) Normalization, the conversion of abbreviated words into full words.

2.3. Lexicon Based

Following the completion of all pre-processing phases, 600 clean data must be collected and prepared for the sentiment testing stage utilizing a Lexicon-based dictionary. In classifying an organism, this stage is crucial. Because this study derives its emotion score from a word-level technique, in which each word's worth of data is examined. By computing each sentiment score, this lexicon-based system may be used to label data or answers, streamlining the categorization process. The method of figuring out each word in a sentence based on its usage in the positive and negative dictionaries, namely by adding up the worth of the opinion. A value of 1 or more and a value of -1 or more are the sums of the opinion values for positive and negative emotion, respectively [8].

The usage of lexicons based on Indonesian text has become common[9], [10]. We used the InSet Lexicon vocabulary from prior research in this study to classify sentiment towards Twitter data. This InSet Lexicon dictionary includes a collection of terms with positive and negative attitudes, as well as a weighted value for each word. There are 3609 positive words and 6609 negative terms in this lexical dictionary. In this Lexicon dictionary, the weight has a value with a score ranging from -5 to +5. We carried out several more words connected to the theme of Employment BPJS in this study [11].

2.4. Sentiment Analysis

Sentiment analysis, often known as opinion mining, is a branch of study that examines people's feelings, views, assessments, judgements, attitudes, and sentiments concerning product units, organizational services, individuals, topics, events, issues, and traits [12]. Sentiment analysis, also known as subjective analysis, opinion creation, judgment extraction, and so on, has various connections to emotional computing, including computer recognition and emotional expression [13].

2.5. Term Frequency-inverse document frequency

The Term Frequency-Inverse Document Frequency (TF-IDF) characteristic was used in this investigation. The TF-IDF measure is widely used in text categorization [14]. The TF-IDF method is a popular statistical strategy for determining the relevance of a term in a corpus document [15]. As stated in equation (1), the TF-IDF weighting system distributes weight to term t in document d. [16].

$$tf.idf_{t,d} = tf_{t,d} \times idf_t \tag{1}$$

The value of $tf_{t,d}$ is the weight of a term t in document d, while idf_t is the inverse document frequency of term t. Equation (2) is an equation for finding the value of idf_t. The value of idf_t is obtained from the result of the logarithm of N divided by df_t. is the total number of documents where df_t is the number of documents containing term t.

$$idf_t = \log \frac{N}{df_t} \tag{2}$$

2.6. Naïve Bayes Classifier (NBC)

The Nave Bayes Classifier is a classifier based on the Bayes theorem. This classifier assumes that a feature's presence in a class is unrelated to other features. [17]. Equation (3) is the Bayes theorem equation.

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \tag{3}$$

Where P(X|Y) is the probability of occurring X if it is known Y. P(Y|X) is the chance of occurring Y if it is known X. P(X) is the probability of occurring X and P(Y) is the probability of occurring Y.

2.7. Evaluation

The k-fold cross-validation approach is used to evaluate the constructed classifier model. This approach divides the data into k equal-sized parts. During the process, we choose one division for testing, while the others are employed for training. This techniwe choose one divisionque and is performed k times to ensure that someone use exactly once each partition for the test [18]. Total errors are calculated by adding the errors from all k processes. A Confusion matrix is also used to assess the accuracy of the classifier model created. A confusion matrix is an important tool in the machine learning visualization method, which usually has two or more categories [19].

3. Results and Discussion

3.1. Lexicon Based Evaluation Results

We evaluated the Lexicon Based technique utilizing data testing of 600 tweets that had been human identified and then compared it to automatic labeling using a lexicon-based method. The goal is to assess the accuracy of automatic labeling using a lexicon. Table 1 displays the results of the confusion matrix calculation.

Table 1. Confusion Matrix Result

Actual	Predicted	
	Positive	Negative
Positive	296	36
Negative	12	256

According to Table 1, the True Positive (TP) = 296; the False Positive (FP) = 36; the True Negative (TN) = 256; and the False Negative (FN) = 12. The lexicon-based accuracy value is 92%, the value precision is 89%, the recal value is 96%, and the F1-score is 92% based on these facts.

3.2. Naïve Bayes Classifier Evaluation Results

The classification model is evaluated using training data of 1000 labeled tweets using a lexicon, with each classified 500 negative and positive. The accuracy value was then determined using k-fold cross validation with 10 trials. The calculation of k-fold cross validation is repeated and has varying accuracy. The average k-fold validation accuracy is 81.60%, and the F1-score is 81.54%. Figure 2 depicts the percentage accuracy of k-fold cross validation results.

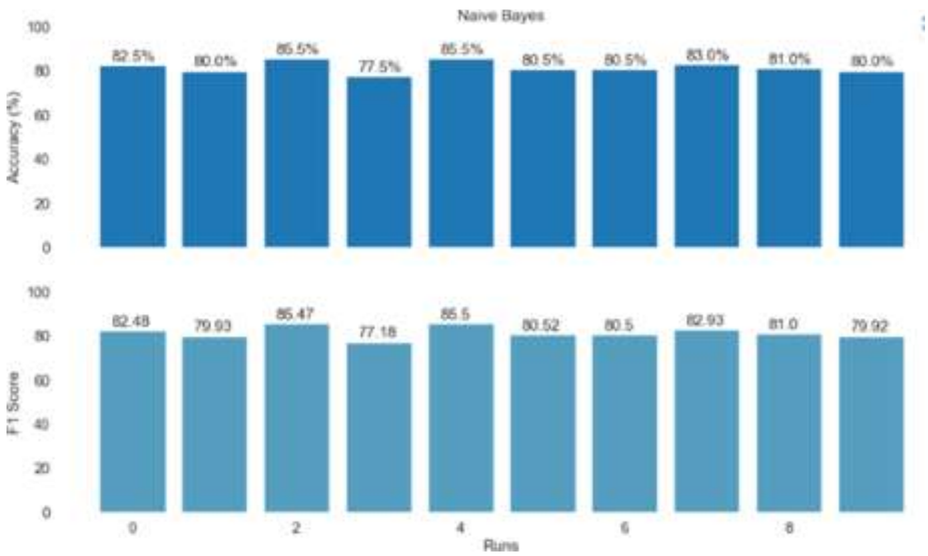

Fig. 2. Percentage of K-Fold Cross Validation Accuracy

The NBC test employs 500 labeled data testing samples. The test findings were as follows: True Positive (TP) = 82, False Positives (FP) = 5, True Negatives (TN) = 397, and False Negative (FN) = 16. The NBC test had an accuracy value of 95%, a precision value of 84%, a recall value of 83%, and an F1-Score of 88%. Table 2 shows the results of the confusion matrix calculation.

Table 2. Confusion Matrix Result

Actual	**Predicted**	
	Positive	**Negative**
Positive	82	5
Negative	16	397

3.3. Sentiment Analysis Results

Following the determination of the accuracy value, sentiment classification was performed on the total data, namely the 4,154 data labels by the machine. The aggregate data revealed that 2,746 were categorized negative and 1,408 were labeled positive. Figure 3 depicts this amount.

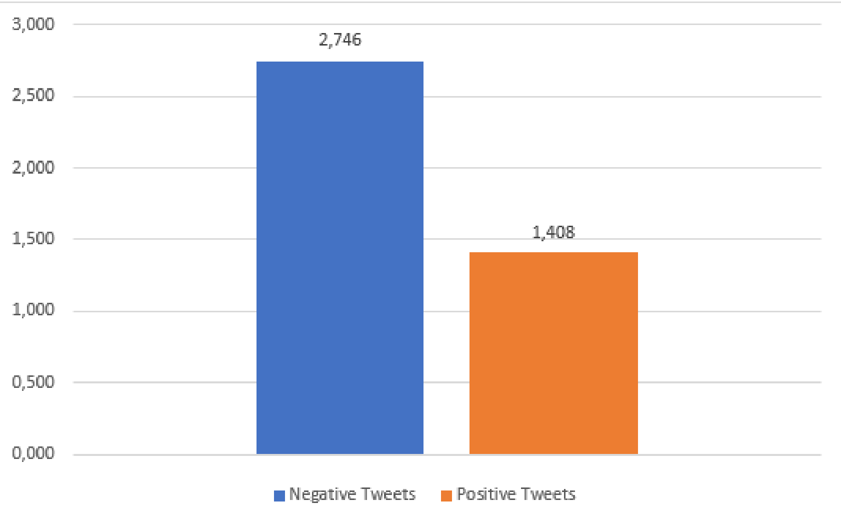

Fig. 3. Sentiment Analysis Results

Positive attitude in the tweet data discusses pay and intensive bonuses from health insurance for old age insurance, according to the analysis of tweet data. Meanwhile, the negative mood in the tweet data highlights the Ministry of Manpower's rejection of the latest law proposal in which the JHT can only be released after National Social Security Agency for Employment participants reach the age of 56. The overall data classification yielded more negative sentiments than positive sentiments, with more people refusing to accept or reject the Ministry of Health's planned draft law regarding JHT, which can only be disbursed when National Social Security Agency for Employment participants reach the age of 56.

4. Conclusion

Several conclusions can be drawn based on the results and discussion of sentiment analysis on Twitter media using the Lexicon Based and Nave Bayes Classifier methods with case studies related to Old Age Benefits at the National Social Security Agency for Employment in Indonesia, namely: (1) this research produces an accuracy value of 92% for the Lexicon Based method and 95% for the Nave Bayes Classifier method. (2) This study generates a sentiment analysis of old age security tweets with a higher proportion of negative attitudes than good sentiments, where the tweets debate the rejection of the Ministry of Health's bill plan.

References

[1] H. Sutrisno, "PENGARUH BPJS KETENAGAKERJAAN DALAM MENINGKATKAN KESEJAHTERAAN TENAGA KERJA," *PREPOTIF J. Kesehat. Masy.*, vol. 4, no. 1, pp. 78–84, Apr. 2020.

[2] S. Rahmawati and M. Habibi, "Public Sentiments Analysis about Indonesian Social Insurance Administration Organization on Twitter," *IJID (International J. Informatics Dev.*, vol. 9, no. 2, pp. 87–93, Dec. 2020.

[3] S. P. Kristanto, J. A. Prasetyo, and E. Pramana, "Naive Bayes Classifier on Twitter Sentiment Analysis BPJS of HEALTH," *Proc. - 2019 2nd Int. Conf. Comput. Informatics Eng. Artif. Intell. Roles Ind. Revolut. 4.0, IC2IE 2019*, pp. 24–28, Sep. 2019.

[4] A. T. J. Barron *et al.*, "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, p. 012054, Mar. 2021.

[5] V. Lingiardi, N. Carone, G. Semeraro, C. Musto, M. D'Amico, and S. Brena, "Mapping Twitter hate speech towards social and sexual minorities: a lexicon-based approach to semantic content analysis," *https://doi.org/10.1080/0144929X.2019.1607903*, vol. 39, no. 7, pp. 711–721, Jul. 2019.

[6] L. Yang, Y. Li, J. Wang, and R. S. Sherratt, "Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning," *IEEE Access*, vol. 8, pp. 23522–23530, 2020.

[7] M. Habibi, A. Priadana, A. B. Saputra, and P. W. Cahyo, "Topic Modelling of Germas Related Content on Instagram Using Latent Dirichlet Allocation (LDA)," pp. 260–264, Jan. 2021.

[8] C. Prianto and N. I. Hamka, "Sentimen Analisis Terhadap Pembelajaran Jarak Jauh Menggunakan Metode Naïve Bayes Classifier dan Lexicon Based," *J. Ilmu Komput.*, vol. 14, no. 2, pp. 79–87, Sep. 2021.

[9] B. T. Pratama, E. Utami, and A. Sunyoto, "A comparison of the use of several different resources on lexicon based Indonesian sentiment analysis on app review dataset," *Proceeding - 2019 Int. Conf. Artif. Intell. Inf. Technol. ICAIIT 2019*, pp. 282–287, Mar. 2019.

[10] F. T. Saputra, S. H. Wijaya, Y. Nurhadryani, and Defina, "Lexicon Addition Effect on Lexicon-Based of Indonesian Sentiment Analysis on Twitter," *Proc. - 2nd Int. Conf. Informatics, Multimedia, Cyber, Inf. Syst. ICIMCIS 2020*, pp. 136–141, Nov. 2020.

[11] Y. N. Yudha, D. Winarso, and syahril, "Analisis Sentimen Masyarakat Pada Twiter Terhadap Isu Covid-19 Menggunakan Metode Lexicon Based," *J. FASILKOM*, vol. 11, no. 2, pp. 97–103, Aug. 2021.

[12] B. Liu, "Sentiment Analysis and Opinion Mining," *http://dx.doi.org/10.2200/S00416ED1V01Y201204HLT016* , vol. 5, no. 1, pp. 1–184, May 2012.

[13] M. Habibi, A. Priadana, and M. R. Ma'arif, "Hashtag Analysis of Indonesian COVID-19 Tweets Using Social Network Analysis," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 15, no. 3, Jul. 2021.

[14] N. N. Amir Sjarif, N. F. Mohd Azmi, S. Chuprat, H. M. Sarkan, Y. Yahya, and S. M. Sam, "SMS spam message detection using term frequency-inverse document frequency and random forest algorithm," in *Procedia Computer Science*, 2019, vol. 161, pp. 509–515.

[15] A. Thakkar and K. Chaudhari, "Predicting stock trend using an integrated term frequency–inverse document frequency-based feature weight matrix with neural networks," *Appl. Soft Comput. J.*, vol. 96, p. 106684, Nov. 2020.

[16] A. A. Putri Ratna, A. Kaltsum, L. Santiar, H. Khairunissa, I. Ibrahim, and P. D. Purnamasari, "Term Frequency-Inverse Document Frequency Answer Categorization with Support Vector Machine on Automatic Short Essay Grading System with Latent Semantic Analysis for Japanese Language," in *ICECOS 2019 - 3rd International Conference on Electrical Engineering and Computer Science, Proceeding*, 2019, pp. 293–298.

[17] K. L. Priya, M. S. Charan Reddy Kypa, M. M. Sudhan Reddy, and G. R. Mohan Reddy, “A Novel Approach to Predict Diabetes by Using Naive Bayes Classifier,” in *Proceedings of the 4th International Conference on Trends in Electronics and Informatics, ICOEI 2020*, 2020, pp. 603–607.

[18] M. Habibi and E. Winarko, “Analisis Sentimen dan Klasifikasi Komentar Mahasiswa pada Sistem Evaluasi Pembelajaran Menggunakan Kombinasi KNN Berbasis Cosine Similarity dan Supervised Model,” Universitas Gadjah Mada, 2017.

[19] C. D. Manning, P. Raghavan, and H. Schutze, *An Introduction to Information Retrieval*. Cambridge, England: Cambridge University Press, 2009.