Journal of Physics: Conference Series

Sentiment analysis and topic modeling of BPJS Kesehatan based on twitter crawling data using Indonesian Sentiment Lexicon and Latent Dirichlet Allocation algorithm

To cite this article: T D Dikiyanti *et al* 2021 *J. Phys.: Conf. Ser.* **1821** 012054

Sentiment analysis and topic modeling of BPJS Kesehatan based on twitter crawling data using Indonesian Sentiment Lexicon and Latent Dirichlet Allocation algorithm

T D Dikiyanti[1,], **A M Rukmi**[1] **and M I Irawan**[1]

[1]Department of Mathematics, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

Abstract. In today’s era, company performance is influenced by quick and easy responses to interacting with users. Twitter is one of the social media which is believed that public opinion on Twitter can influence the government or companies to make policies. Public criticism on Twitter is more quickly responded than people who contacting customer service directly, this is because the companies or government do not want their image to be bad due to delays in responding to public complaints on Twitter. Studies related to opinion writing on social media can use the method of topic modeling and sentiment analysis in order to get what topics are currently being discussed and also the value of their sentiments. Modeling of the topic was carried out using Latent Dirichlet Allocation and sentiment analysis using the Indonesian Sentiment Lexicon. A case study of public opinion on BPJS Kesehatan using Twitter data for 3 months from February to April 2020, obtained 5 main topics with the BPJS Kesehatan's New Contribution Rate as a trending topic with a sentiment value of 61.7% positive and 38.3% negative.

1. Introduction

Sentiment analysis is a research topic that is still very important and is currently being used. This research will analyze a text document that can be obtained from online forums, blogs, social media and various sites that contain reviews. With this sentiment analysis, the information from a semi-structured text document is converted into a more structured data. Semi-structured data refers to data that has not been classified in a database, but contains important information that separates the individual elements in the data. Meanwhile, structured data is data that can be processed or stored. This sentiment analysis is used to analyze public opinion about a product, service, a brand, politics and several other topics. From this analysis, it can be used to provide feedback for a product or service so that it can be used as a reference for developing a better product or service. Topic modeling is also very much needed because from this stage you will find out what topics are currently being discussed by online media users.

Health is the most basic need for every society in all parts of the world. By having a healthy body, we will be able to carry out all our daily activities. If someone's health condition is in decline, it is not uncommon for us to feel a risk that has a material impact on the sufferer's finances, moreover this incident is uncertain, which indirectly makes the community always have to be financially ready to take medication if their health is declining. Health care insurance in Indonesia has existed since the Dutch colonial era. And after independence, in 1949, efforts to ensure the need for health services for the community, especially civil servants and their families, were continued by Prof. G.A. Siwabessy, as the Minister of Health who served at that time. In 2004, the government issued Law No. 40 of 2004 on the

National Social Security System (SJSN). Social security is a form of social protection to ensure that all people can fulfill their basic needs for a decent life. Then in 2011 the government enacted Law Number 24 of 2011 concerning the Social Security Administering Body (BPJS) and appointed PT. Askes (Persero) as the organizer of the social security program in the health sector, so that PT. Askes (Persero) also changed to BPJS Kesehatan. [1]

According to data published by BPJS Kesehatan on the website, it is explained that there are already around 208 million participants from BPJS Kesehatan in grades I, II, and III. With this large number of participants, BPJS has successfully fulfilled its commitment to serve the basic needs of public health through reliable, superior and trusted services. The results of the survey on participant satisfaction and health facilities towards BPJS Kesehatan conducted by one of the well-known survey institutions, namely Myriad Research Comitted, show that as many as 81% expressed satisfaction with BPJS Kesehatan. This figure exceeds the public satisfaction target set by the government, which is 75%. The 81% total percentage is a combination of the participant satisfaction index for services at the First Level Health Facilities, Advanced Referral Health Facilities, Branch Offices, and BPJS Health Centers. [2]

Quoted from the news on the kompas.com page on October 30, 2019, Twitter's 3rd quarter 2019 financial report, daily active users on the Twitter platform were recorded to have increased 17 percent, to 145 million users. Indonesia is claimed to be one of the countries with the largest growth in daily active Twitter users. From the TNS research which took place in the first year of 2016, it was said that Twitter can be trusted for product information. This research examines the daily activities of users to help marketers gain a better understanding of their personal values, behavior, how they view life and their various habits in social and digital life. Twitter has become part of the daily life of millennial generation in Indonesia. Twitter is also believed that public opinion on Twitter can influence the government or companies to make policy. Criticism on Twitter is more quickly responded to than the public contacting customer service, this is because the government or companies do not want their image to be bad because they are not quick to respond to public complaints on Twitter.

Therefore, based on the background explanation above, in this research an analysis of public sentiment in Indonesia regarding BPJS Kesehatan will be carried out as well as modeling the topics most discussed by Twitter netizens using the Latent Dirichlet Allocation method and sentiment analysis based on the Indonesian Sentiment Lexicon. LDA is development method from LSI, the research from Riko Wijayanto's Final Project on "Implementation of Social Media Mining for Decisionmaking in Product Planning Based on Topic Modeling and Sentiment Analysis" which uses the LSI method, namely the method before the LDA. [3]

This paper contains several things that will be discussed in this study, while the updates made in this paper are:

1. Modeling the topic using the modified Latent Dirichlet Allocation (LDA) method to determine the contribution rate.
2. Conduct a sentiment analysis using the Indonesian Sentiment Lexicon and project the sentiment results from modeling topics with LDA based on positive or negative polarity.

2. Literature review

There are several studies with themes related to this reasearch, namely the Final Project of ITS Information Systems written by Ari Agustina in 2013 which based on this research, LDA modeling has been proven to be able to identify topics well. The quality of the output from the resulting modeling of the topic is also quite good, indicated by the perplexity value of 34.92 with a standard deviation of 0.49 in 20 iterations. Accuracy by testing the resulting model is 83.7%. [4].

2.1. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative probabilistic topic modeling algorithm. This algorithm is used as a clustering unsupervised learning for unstructured data. LDA is able to find and categorize keywords in documents into their respective topics. The method chosen is able to classify each meaningful theme. It is generated based on the idea that documents consist of a mixture of latent topics and those topics are characterized by word distributions. Latent is defined as something that is present but not visible. LDA itself works with Gibbs sampling, which is to calculate the distribution of opportunities together by sampling each variable one by one based on the value of the other variables.

In 2017, Putri and Kusumaningrum, students from Dipenogoro University, proposed a sentiment analysis with topic modeling using the Latent Dirichlet Allocation (LDA) method which is applied to read the general trend of tourists' reviews based on certain topics classified into positive and negative sentiments. The results show that the average accuracy is at 60% and the best accuracy is 80%. It takes a balance in each part of the parameters to get the optimal level of accuracy. [5].

In 2019, Kaveh Bastani, Hamed Namavari and Jeffrey Shaffer published a journal on research on modeling topics about customer complaints using LDA which explained that LDA is an algorithm for unstructured data and is suitable for analyzing consumer behavior, product reviews and other research. [6]

2.2. Sentiment analysis

In terms of sentiment analysis, there are journals that discuss the sentiments of political parties participating in the election on Twitter in 2019, where this research was conducted by Ibnu Fanhar, Anisa Herdiani, and Widi Astuti from Telkom University, Bandung. This study obtained an average precicion of 40%, a recall of 42%, FI 35% and Accuracy of 61%. This result is quite good, but with a few notes so that later the system can get better precision, recall, FI, and accuracy results. [6]

InSet Lexicon, lexicon is a complete vocabulary collection of a language. This is a data dictionary-based method containing the words of sentiment in Indonesian which have been accompanied by a weight on each word. From the journal made by Fajri Koto and Gemala Y. Rahmaningtyas in their research entitled "InSet Lexicon: Evaluation of a Word List for Indonesian Sentiment Analysis in Microblogs" which was built to identify written opinions and group them into positive or negative opinions, which can be used to analyze public sentiment on a particular topic, event or product. Compiled using word collections from Indonesian tweets, InSet was created by weighing each word manually and enhanced by adding a set of stemming and synonyms. In this InSet there are 3,609 positive words and 6,609 negative words with scores ranging between -5 and +5. [10]

2.3. Natural language processing

Natural Language Processing is a set of tools, techniques, and algorithms for understanding natural language based on data which is mostly unstructured data such as text, images and videos which contain very large data. Natural language or natural language is a language developed and evolved by humans to make it easier for us to communicate rather than using language artificially, such as a computer programming language. The most important component in NLP is how to design and build applications and systems that allow interaction between machines and human-made natural languages. The NLP technique itself is designed to process and understand human natural language to develop maximum results.

2.3.1. Crawling from twitter API.

Crawling is a technique used to collect data contained in a web. Crawling works automatically, where the information obtained will match the keywords that have been entered by the user. The process is to visit each document on a website, starting from listing all the urls of the website, tracking them one by one, then the data will be collected. In this case study, the used data is given from Twitter API. API or what is commonly called Application Programming Interface is a program / application provided by the developer so that certain parties can more easily access or retrieve data from Twitter and process it.

From here you will get some account information, view mentions, retweets, favorites, followers, friends from a specific Twitter account. In this study using the REST API, which is used to retrieve historical data from Twitter.

2.4. Topic modeling

Topic modeling is an algorithm that is used to find main topics based on documents or sets of words that are very large or large and are unstructured documents. Topic modeling can organize word sets according to the topics found. Modeling this topic is one of the unsupervised learning methods that applies grouping to find latent variables or black holes from a number of text or document data sets. Topic modeling algorithms can be adapted to many types of data. This algorithm has been used to find patterns in generic data, images, and social networks. [8]

In this study, topic modelling used in this paper is as stated in LDA section.

3. Method and methodology

The object used in this research is tweet data from Twitter with the keyword "BPJS Kesehatan". In this study, the aspects to be studied are topic modeling and sentiment analysis regarding BPJS Kesehatan during certain intervals. The software used is Anaconda Navigator with Spyder (Phyton-based).

3.1. Preprocessing data

From the data obtained, there are still many HTML tags that are useless in sentence assessment so that this stage is the stage for removing HTML tags. Then the elimination of punctuation marks, such as periods, commas, colons and other punctuation marks. Besides that, it also removes numbers because in this research it only focuses on text data. Followed by changing the capital letter to lowercase.

Stemming aims to change a word into its root word by eliminating all word affixes including word prefixes, word insertions, word endings and / or removing word prefixes and endings in derivative words. In doing this research, the writer did stemming using Literature library which is a simple word library in Indonesian. Then the elimination of unnecessary words, which have absolutely no high word value so are usually deleted from the sentence when processing them so as to retain the words that have the maximum value. Examples of words like, "and", "like", and others.

Finally, tokenization is the process of breaking or dividing textual data into smaller and more meaningful components, which will later be called tokens. In this stage, two stages of tokenization are carried out, namely sentence tokenization and word tokenization. This technique is used to break a document into a sentence and a sentence into a word.

3.2. Data modeling with Latent Dirichlet Allocation

At this stage, an analysis of the results of the distribution of topics is carried out using the Latent Dirichlet Allocation which proves that the resulting distribution of topics is compatible with existing documents. At the same time, we got some of the topics most frequently discussed by Twitter users regarding BPJS Kesehatan.

In LDA, the user needs to determine the number of topics the algorithm will classify. The algorithm will then find the latent relationship between the words and the topic the words are associated with. LDA implements a bag-of-words model that ignores order of words. The bag-of-words model is a model that tokenizes each word regardless of their grammar and word order. However, LDA will look at the words independently and the occurrence of each word used in the classifier training process.

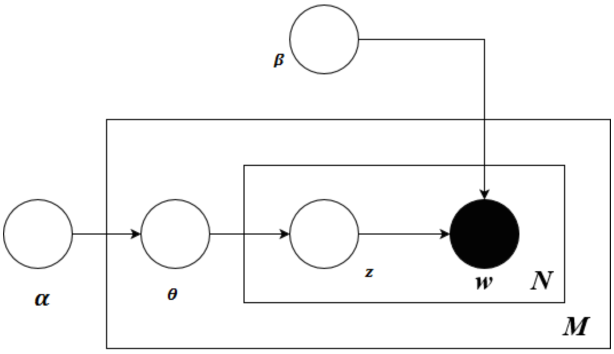


Figure 1. Graphic model for Latent Dirichlet Allocation

M represents the document used, N represents the word set in the document, w represents the words represented in the document, z represents the topic for the word symbolized, θ represents the topic distribution for the document, α represents the parameters defined for the distribution of documentary topics. The greater the α value that is owned by a document, which indicates the more mix of topics discussed in the document. The lower the α value indicates that the document only addresses a certain number of topics. Meanwhile, β is the parameter used to calculate the distribution of words in the topic. The higher the β value, the more words will be in the topic. The smaller the β value, the fewer words in the topic so that it is more specific.

In general, the iteration of the LDA algorithm itself is as follows:

1. Initialize α and β parameters
2. Initialize random assignment of topics
3. Iteration, For each word in the document, Resample topics for words. Given all other words and determining the assignment of the topic.
4. Get results
5. Model evaluation

First, the k dimension of the Dirichlet distribution is assumed to be known and fixed. Second, the word probability is parameterized by a matrix β, k × V where $\beta_{ij} = p\left(w^j = 1 \middle| z^j = 1\right)$, which is currently treated as a fixed quantity that must be estimated. Finally, the Poisson assumption is less used and a more realistic document Length distribution can be used as needed. Next, note that N is independent or independent of all other generated variables (θ and z). The random variable of the dimensional Dirichlet k can take the value to the simplex (k-1) (a k vector θ which lies in the simplex (k-1) if $\theta_i \geq 0, \sum_{i=1}^{k} \theta_i = 1$) and has the following probability densities:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{a_1 - 1} \dots \theta_k^{a_k - 1}, \qquad (1)$$

where parameter α is a vector-k with component $a_i > 0$,, and where Γ (x) is a gamma function. Given the parameters α and β, the joint distribution of the topic mix θ, the N set of z topics, the N set of w words is indicated by:

$$p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta), \qquad (2)$$

where $p(z_n|\theta)$ is θ_i for unique i such that $z_n^i = 1$. Integral to θ and add to z, we get the marginal distribution of a document: [8]

$$p(\boldsymbol{w}|\alpha, \beta) = \int p(\theta|\alpha) (\prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta)) \, d\theta, \qquad (3)$$

Finally, by taking the product of the marginal probability from a single document, we obtain the probability of a corpus:

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right) d\theta_d, \tag{4}$$

3.3. Sentiment analysis on modeled data

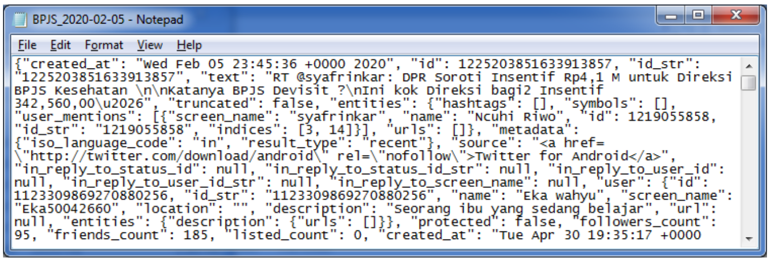
At this stage, after obtaining several topics that are often discussed, a sentiment analysis will be carried out using the Indonesian Sentiment Lexicon which will later get the results of the opinions of Twitter users in positive or negative points.

4. Program running results

The data is obtained by performing a crawling technique using the tweepy library provided by python which requires a token to access the Twitter API. Tokens are obtained after submitting to the Twitter application developer. The data needed in this study are tweets in Indonesia with a span of time from February 2020 to April 2020.

4.1. Preprocessing results

The keywords used to get tweets are “bpjs” and “health bpjs”. In retrieval this tweet uses the "geocode" parameter based on the latitude, longitude and radius values of the user's location, namely Surabaya. These latitude, longitude and radius values are obtained from Google Maps. Retrieval of this tweet is limited by time, a maximum of 8 days before the time of retrieval so it must be crawled every week. Tweets that have been successfully obtained will then be saved into a file with the json format.



```
{"created_at": "Wed Feb 05 23:45:36 +0000 2020", "id": 1225203851633913857, "id_str":
"1225203851633913857", "text": "RT @syafrinkar: DPR Soroti Insentif Rp4,1 M untuk Direksi
BPJS Kesehatan \n\nKatanya BPJS Devisit ?\nIni kok Direksi bagi2 Insentif
342,560,00\u2026", "truncated": false, "entities": {"hashtags": [], "symbols": [],
"user_mentions": [{"screen_name": "syafrinkar", "name": "Ncuhi Riwo", "id": 1219055858,
"id_str": "1219055858", "indices": [3, 14]}], "urls": []}, "metadata":
{"iso_language_code": "in", "result_type": "recent"}, "source": "<a href=
\"http://twitter.com/download/android\" rel=\"nofollow\">Twitter for Android</a>",
"in_reply_to_status_id": null, "in_reply_to_status_id_str": null, "in_reply_to_user_id":
null, "in_reply_to_user_id_str": null, "in_reply_to_screen_name": null, "user": {"id":
1123309869270880256, "id_str": "1123309869270880256", "name": "Eka wahyu", "screen_name":
"Eka50042660", "location": "", "description": "Seorang ibu yang sedang belajar", "url":
null, "entities": {"description": {"urls": []}}, "protected": false, "followers_count":
95, "friends_count": 185, "listed_count": 0, "created_at": "Tue Apr 30 19:35:17 +0000
```

Figure 2. Twitter crawling result view

Table 1. Before and after data preprocessing comparison.

Data before *preprocessing*	Data after *preprocessing*
"" gagal nalar pe ama adalah ada berapa banyak contoh negara yang menjadikan corona sebagai penyebab menurunnya pariwisata ke	gagal nalar pe ama contoh negara jadi corona sebab turun pariwisata
"" mn uud negara wajib menghormati melindungi amp memenuhi hak atas jaminan kesehatan dan jg menjamin penyediaan faske	mn uud negara wajib hormat lindung penuh hak jamin sehat jg jamin sedia faske
"" nik asli nama asli heddy setya permadi tgl lahir	nik asli nama asli heddy setya madi lahir nunggak iur bpjs sehat rp juta

nunggak iuran bpjs kesehatan rp juta	
bpjs kesehatan sambung sugiyanto juga menyampaikan belasungkawa mendalam atas meninggalnya kedua orangtua ali mardani	bpjs sehat sambung sugiyanto belasungkawa dalam tinggal orangtua ali mardani
"" salah satu cara kurangi kasus intoleransi antaragama di ri adalah mengadopsi sistem bpjs kesehatan soal distribusi setiap kk	salah kurang intoleransi antaragama ri adopsi sistem bpjs sehat distribusi kk

Furthermore, the coherence of the model will be calculated so that the resulting topic modeling is a stable model result and can represent the entire document. In this stage, we have to determine some input such as the minimum and maximum number of topics, alpha and beta values, and also the number of iterations. This test will be carried out with an experiment with a number of iterations which later will be selected which iteration is the best by considering the coherence value. In theory, the higher the coherence value, the better it is to represent the model.

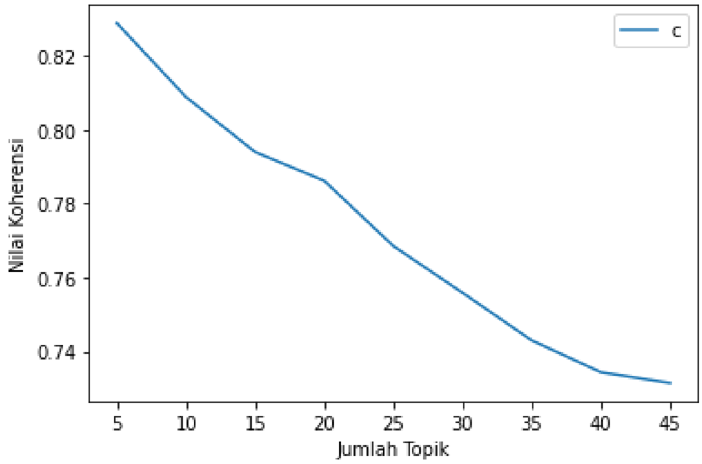


Figure 3. Coherence value graphical result view in 100 iterations

The highest coherence value is when the number of topics is 5 topics, so the model with the 100 iterations test will be more stable when the number of topics is initialized as many as 5 topics.
From the picture above, it can be seen that there are still 3 topics that overlap or intersect, but not many parts are cut off.

4.2. Topic modelling with Latent Dirichlet Allocation result

The result given from the methods different for each iteration with the better results obtained when the iteration value is higher. At some point the result will not change in an amount of iteration value (steady-state condition). The testing will be start at 100 iterations, with increase and decrease at will.

Then, after evaluating for any iteration value of 50,100, 200, 500, 1000, 1500. Given the best solution for the topic modelling is done in 1500 iterations as it shown in picture below.

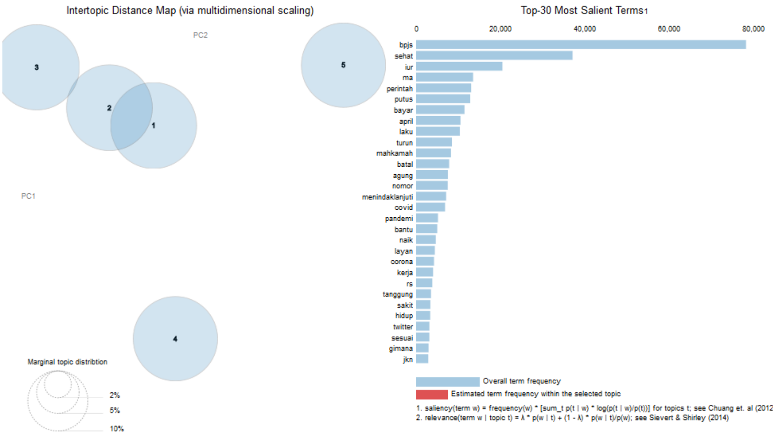


Figure 4. Visualization results of modeling topics tested in multiple iteration value

4.3. Sentiment analysis with Indonesian Sentiment Lexicon result

The number of words that were not counted or words that were not in the Lexicon list were 567,488 out of 971,741 words counted in the program. Here is the polarity score find it values.

Table 2. Polarity calculation model

Sentence	Word in lexicon	Score	Conclusion
contoh	jadi	1	Positive = 14
negara jadi	sebab	3	Negative = -8
corona sebab	turun	1	
turun	gagal	-4	Sentiment
pariwisata	jadi	1	Score = 6
gagal nalar	sebab	3	
pe ama	turun	1	
contoh	gagal	-4	
negara jadi	jadi	1	
corona sebab	sebab	3	
turun			
pariwisata			
gagal nalar			
pe ama			
contoh			
negara jadi			
corona sebab			

Table 3. Polarity value calculation table

MONTH	POLARITY		PERCENTAGE	
	POSITIVE	NEGATIVE	POSITIVE	NEGATIVE
February 2020	129.994	-78.339	62.39%	37.61%
March 2020	300.188	-213.047	58.49%	41.51%
April 2020	262.464	-146.302	64.20%	35.80%
	AVERAGE		**61.70%**	**38.30%**

From the results of the sentiment analysis using the Indonesian Sentiment Lexicon, public sentiment regarding BPJS Kesehatan is in quite good results, which is positive at 62%. Adjusted to the results of modeling the predetermined topic, the results of negative sentiment resulted from public complaints about the news about BPJS Health premium rates which were still being debated among decision makers at that time. Meanwhile, the positive sentiment results resulted from public sentiment regarding the cancellation of the BPJS Health premium rate increase.

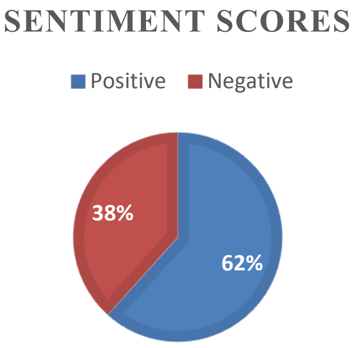


Figure 5. Sentiment analysis result image

5. Conclusion and recommendation

Modeling of the topic was carried out using the Latent Dirichlet Allocation method and sentiment analysis using the Indonesian Sentiment Lexicon. A case study on BPJS Kesehatan by utilizing Twitter data for 3 months from February to April 2020 found 5 main topics which, if outlined by the public, often discuss about the BPJS Health New Contribution Rate with a sentiment value of 61.7% positive and 38.3% negative.

It is better to do the steps where changing non-standard words or words that are often used by Twitter users freely to be converted into standard words according to the Big Indonesian Dictionary or using data that public sentiment has indeed been expressed using formal language. And also the use of lexicons which have even more vocabulary so that all words can be weighted and produce a more effective sentiment analysis.

Hopefully this research can help further research related to the field of mathematical modeling and data analysis, especially large-scale data processing.

6. References

[1] Humas BPJS Kesehatan, "Sejarah Perjalanan Jaminan Sosial di Indonesia," https://bpjs-kesehatan.go.id/bpjs/pages/detail/2013/4, Jakarta, 2018.
[2] Humas BPJS Kesehatan, "Indeks Kepuasan Peserta dan Faskes terhadap BPJS Kesehatan Sukses Lampaui Target," https://bpjs

kesehatan.go.id/bpjs/dmdocuments/46b07615fad0343451d7860b65909610.pdf, Jakarta, 2015.

[3] M. I. Irawan, R. Wijayanto, M. L. Shahab, N. Hidayat, A. M. Rukmi, “Implementation of Social Media Mining for Decision Making in Product,” Journal of Physics: Conference Series, p. Conf. Ser. 1490 012068, 2020.

[4] A. Agustina, "Analisis dan Visualisasi Suara Pelanggan pada Pusat Layanan Pelanggan dengan Pemodelan Topik Menggunakan Latent Dirichlet Allocation (LDA) Studi Kasus: PT. Petrokimia Gresik," Tugas Akhir S1 Sistem Informasi ITS, Surabaya, 2017.

[5] I. R. Putri and R. Kusumaningrum, "Latent Dirichlet Allocation (LDA) for Sentiment Analysis Toward Tourism Review in Indonesia," *Journal of Physics,* vol. 801, 2017.

[6] I. F. Nur. F, A. Herdiani and W. Astuti, "Analisis Sentimen Berbasis Leksikon Inset Terhadap Partai Politik Peserta Pemilu 2019 Pada Media Sosial Twitter," *e-Procceeding of Engineering,* vol. 6, no. 3, p. 10397, 2019.

[7] K. Bastani, H. Namavari and J. Shaffer, "Latent Dirichlet Allocation (LDA) for Topic Modeling of CFPB Consumer Complaints," Expert System With Applications, vol. 127, pp. 256-271, 2019.

[8] A. F. Hidayatullah, E. C. Pembrani, W. Kurniawan, G. Akbar and R. Pranata, "Twitter Topic Modeling on Football News," *3rd International Conderence on Computer and Communication Systems,* 2018.

[9] D. M. Blei, Y. A. Ng and I. M. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research 3 (2003) ,* pp. 993-1022, 2003.

[10] F. Koto and G. Y. Rahmaningtyas, "InSet Lexicon: Evaluation of a Word List for Indonesian Sentiment Analysis in Microblogs," jakarta, 2017