Sentiment Analysis and Topic Modeling of the Socio-Economic Registration 2022

I Simbolon[1], **N H Manurung**[1], **S Andini**[1], **L H Suadaa**[1,*]

[1]Statistical Computing Department, Politeknik Statistika STIS, Indonesia

*Corresponding author's e-mail: lya@stis.ac.id

Abstract. Socio-Economic Registration or Regsosek is an activity of Statistics Indonesia (BPS) that aims to collect data related to the profile, social and economic conditions, and welfare levels of all residents in 514 regencies/cities in Indonesia. One indicator of the success of Regsosek 2022 is the response and opinion from the community regarding the activity. The response and opinion can provide an overview of the implementation of Regsosek 2022 so that the picture can be used as a lesson learned to carry out the following population data collection. This study uses several methods to analyze the results of community responses and opinions on Regsosek activities, especially on Twitter social media. The method used in this research is sentiment analysis classification with four techniques: Naïve Bayes, Nearest Centroid, K-Nearest Neighbors, and Support Vector Machine. Then, the performance of the four techniques will be compared. In addition, the topic modeling method will also be used with two techniques, namely Latent Semantic Analysis and Latent Dirichlet Allocation. Data is collected using web scraping techniques. The results obtained from the sentiment analysis classification are that the Nearest Centroid method provides the best results with a relatively high and balanced f1-score value in positive and negative sentiments, which are 59% and 66%, respectively. Moreover, LDA modeling results are better than the LSA method for topic modeling results.

1. Introduction

Socio-Economic Registration (Regsosek) is an activity of the Central Bureau of Statistics (BPS) and was held on October 15 - November 14, 2022, which collected data related to the profile, social conditions, economy, and welfare level of the entire population in 514 cities/regencies in Indonesia. Regsosek itself was carried out at the behest of the president through the submission of the RUU APBN TA 2023, which aims to improve the social protection program reform database and accelerate the elimination of extreme poverty. With the implementation of Regsosek, it is hoped that the government can implement its programs in an integrated, more efficient way to improve the quality of government services, ranging from health, education, and social assistance to population administration.

The data collection method still uses paper and pencil interviewing (PAPI) door-to-door and involves 441 thousand data collection officers. Due to the massive scale and magnitude of the activities carried out, different views and responses from all parties related to the implementation of this Regsosek.

All the censuses and surveys conducted by BPS are carried out in accordance with international standards for implementing statistical activities, namely the General Statistical Business Process Model (GSBPM). Every stage of GSBPM have been passed in the process of completing the Regsosek and have now reached the final stage, namely Evaluate. One new method that available to be used in

evaluating Regsosek is to analyze public sentiment regarding this series of census implementations. Therefore, sentiment analysis and topic modeling need to be carried out as a method in the process of evaluating the success of Regsosek 2022. The result of this research can be a source of knowledge for the organizers, in this case, the BPS, as well as related institutions such as the National Development Planning Agency, the Ministry of Finance, the Ministry of Home Affairs, the Ministry of Villages, and the Ministry of Communication and Information to describe the situation and conditions as well as public views regarding the implementation of Regsosek. In addition, the responses and opinions of the community regarding Regsosek can also provide an overview and understanding of the steps for carrying out population registration data collection that is more comprehensive so that it can become a lesson in carrying out population data collection in the future.

Responses and public opinion related to the implementation of Regsosek can be divided into three: positive, negative, and neutral. Due to the rapid progress of the times, data about public responses and opinions can already be found through social media as a popular forum for expressing public responses and opinions regarding various matters, including responses and opinions related to Regsosek. The development and progress of information technology, the availability of the fast Internet, and the significant increase in social media users have changed media habits and people's habits [1]. New communication technologies expand the possibilities for transmitting and receiving information. People use social media to find information, spread stories, and discuss concerns [2].

Social media is a term that describes various technologies used to bring people together to collaborate, exchange information, and interact through message content through a web application network. As the Internet continues to evolve, so do the technologies and features available to users. One notable advantage of social media is disseminating information that can be done anonymously and encrypted [3]. In this era, information exchange occurs quickly and imperceptibly, especially with the growing influence of social media in spreading this information and leaving historical data stored in social media server databases [4].

One of the pioneers of social media that is very capable is Twitter. Twitter only allocates 280 characters in one tweet raised by each user. This allows information to be condensed into several sentences, even a few words. Because of the system offered by Twitter, many data analysts do sentiment analysis and topic analysis on this application [5]. Using public tweets on the Twitter application platform can collect responses and opinions from the public regarding the implementation of Regsosek and is an easy and inexpensive way to collect data through conventional surveys. However, because Twitter is an extensive application and many tweets are spread on social media, it will be challenging to manually do sentiment analysis and topic modeling [6].

Data Mining is a new method which used for collecting and learning large amounts of data quickly and efficiently. Data mining comes to be a solution of finding new insights or unknown associations and relations from a commercial data which often used to predict the future as well [7]. Web scraping is a way to extract information from websites using computer software. Computer software has an ability to imitate humans behaviour in exploring the internet by running a complete web browser [8]. In other words, web scraping provides a way to complete data collection more quickly and efficiently than the manual method of copy-paste information from websites into a database [9].

To analyze public sentiment on the Twitter platform, there are several methods that can be used. These methods have been used and provided good results in previous research [10]. For this reason, this research will use this classification technique and then compare its performance in classifying public sentiment towards Regsosek. The classification techniques are Naïve Bayes, Nearest Centroid, K-Nearest Neighbors, and Support Vector Machine. On the other hand, the techniques that will be used in modeling this topic also refer to previous research, namely Latent Semantic Analysis and Latent Dirichlet Allocation, whose modeling results will also be compared with each other.

Based on the explanation above, this study aims to apply big data processing methods to help analyze tweets spread across the Twitter platform. This data can produce knowledge related to the implementation of the Regsosek 2022, in this case, analyzing the sentiments generated by the community regarding the Regsosek 2022. The sentiments that arise can be either positive, negative, or

even neutral. These sentiments can be used by the government, in this case, the BPS and related institutions such as Bappenas, the Ministry of Finance, the Ministry of Home Affairs, the Ministry of Villages, and the Ministry of Communication and Information to identify problems that arise during the holding of the Regsosek 2022 data collection activities. This information will serve as an evaluation for the future in carrying out other data collection activities going forward.

By paying attention to the background, we can draw out the problems discussed in this study, namely the steps to identify and analyze Twitter tweet data in text. Text mining is the field of science that specifically discusses text data processing methods. Text mining is extracting patterns/information through existing data, in this case, text. Text mining is the development of data mining or knowledge retrieval through structured databases [11].

So far, researchers have not found previous studies that identify community sentiment regarding the holding of the Regsosek 2022. Therefore, considering that the community's responses are crucial to the success of the Regsosek 2022 activities, it is essential to identify and analyze the sentiments generated by the community through tweets on Twitter that appeared during the Regsosek 2022 activities. Therefore, the following problems can be identified:

1. Community sentiment regarding the Regsosek data collection, which is difficult to collect and interpret using manual methods.
2. Topics that arise and are discussed by the community related to the data collection on Regsosek varied and were difficult to collect and interpret using manual methods.

2. Research Objective

The objectives of this study are as follows:

1. Conduct experiments on the four sentiment analysis techniques, including Naïve Bayes, Nearest Centroid, K-Nearest Neighbors, and Support Vector Machine on data tweets about Regsosek.
2. Analyze topics tweets of Regsosek using topic modeling techniques, including Latent Semantic Analysis and Latent Dirichlet Allocation.

3. Related Researchs

Research by Hendrawan et al. [12] aims to firstly, the study sought to apply topic modeling techniques to the BPS Knowledge Management System, enabling the identification of document topic groups. By employing Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) methods, the research aimed to determine the most suitable clustering technique for BPS's knowledge management systems. Secondly, the research aimed to provide valuable recommendations regarding the optimal topics within the system. By analyzing the topic groups generated through the topic modeling, the study sought to identify the most relevant and beneficial topics for BPS Statistik Indonesia. Finally, the research aimed to develop a practical and efficient topic modeling web service for BPS. This encompassed the creation of a data preprocessing function and a topic group recommendation function. The goal was to implement a RESTful web service that would facilitate seamless integration of the LDA model and provide valuable services to users, enabling them to preprocess data and receive topic recommendations for documents entered the BPS Knowledge Management System. Upon completion of the research, the results indicated that the Mallet LDA model, with 25 topic groups and a coherence score of 0.4803, outperformed other models. Consequently, the study confirmed that LDA was the most effective modeling method for BPS Statistik Indonesia's knowledge management needs. Subsequently, the successful implementation of the LDA model in the RESTful web service fulfilled the research's objective of providing practical and accessible services for data preprocessing and topic recommendations within the organization's knowledge management system.

In their research, Jannah et al. [13] carry out research aimed to to examine public sentiment and psychology as seen through public emotional states during the COVID-19 pandemic from March to July 2020. Twitter data was used as research data and then analyzed using sentiment and emotion analysis with a lexicon-based approach. The study results show that negative sentiments are more widely

expressed, and fear is the emotion most felt by the community. This can be used as a recommendation for the government to pay more attention to the people's emotional state.

Research by Illia et al. [14] was conducted to get public sentiment toward the application using Twitter data. The data collection period was from 31 August to 7 September 2021, and this period was chosen because of the emergence of news regarding vaccine data leaks related to data leaks in the PeduliLindung application. Sentiment analysis was performed using the TextBlob and VADER libraries. The results of this sentiment analysis are sufficient to display public opinion, and it is hoped that decision-makers can improve applications based on these opinions. Then, it was found that the VADER library is better at conducting sentiment analysis in research because the lexicon approach is based on social media.

This research is distinct in that it evaluates community responses to a specific government activity (Regsosek) and uses a combination of sentiment analysis and topic modeling techniques. The study builds on the existing research by using similar sentiment analysis techniques (as seen in research [13] and [14]) and topic modeling methods (as seen in research [12]) to understand how the community perceives and responds to a government initiative. The use of various classification techniques and the comparison of their performance further contribute to the field of sentiment analysis. Furthermore, this research aims to assess public sentiment on Twitter related to the Regsosek activity held from October 15 to November 14, 2022. During that time frame, there had been no research on public sentiment regarding Regsosek activities. Another thing that sets this research apart from the previous studies is that it compares the performance of the four sentiment analysis techniques used to find the best Machine Learning method to classify tweets, namely Naïve Bayes, Nearest Centroid, K-Nearest Neighbors, and Support Vector Machine. In addition, this research not only conducts sentiment analysis but also performs and compares different topic modeling methods namely Latent Dirichlet Allocation and Latent Semantic Analysis to identify key topics of opinion summarization.

4. Methods

The scope of the research conducted by the authors is to apply sentiment analysis and topic modeling methods to Twitter tweets from the Indonesian community containing the keyword "regsosek". As sentiment analysis models, we used Naïve Bayes, Nearest Centroid, K-Nearest Neighbors, and Support Vector Machine methods on tweets generated by the community regarding Regsosek 2022, specifically in terms of positive and negative sentiment, then compared the results. The explanations of each method are provided below:

1. Naïve Bayes

Naïve Bayes is based on Bayes' theorem and makes the "naïve" assumption that the features (words or terms) in a document are conditionally independent given the class label (sentiment category). Naïve Bayes calculates the probability that a given text document belongs to a particular sentiment class (e.g., positive, negative, or neutral). It uses the conditional probabilities of words in the document given each sentiment class. By comparing these probabilities, the algorithm assigns the document to the sentiment class with the highest likelihood [15].

$$y = argmax_y P(y) \prod_{i=1}^{d} P_{(x_i|y)}$$

Through the model equation, it is known that the dataset is divided into two parts, namely the feature matrix and the response/target vector. The Feature Matrix (X) contains all the vectors (rows) of the dataset, where each vector consists of dependent feature values. The number of features is d, which is X = (x1, x2, x3, xd). The response/target vector (y) contains the class/group variable values for each row of the feature matrix.

2. Nearest Centroid

In nearest centroid classification, each class is represented by a centroid (a vector of features). To classify a new document, the algorithm calculates the distance between the document's feature

vector and the centroids of each class. The class with the nearest centroid is assigned as the predicted sentiment class for the document. Nearest Centroid can be used to classify text documents into sentiment categories. It relies on the similarity between the features of the document and the centroids of sentiment classes to make predictions [16].

$$\hat{y} = argmin_{l \in Y} ||\vec{\mu}_l - \vec{x}||$$

$\hat{y}$ This symbol typically represents the predicted or estimated class or cluster for a given data point.
l Represents a class or cluster label from the set Y of possible classes or clusters.
$\vec{\mu}_l$ Indicates the centroid or representative point for class l.
$\vec{x}$ Represents a data point or vector that you want to classify or cluster.

3. K-Nearest Neighbors
K-NN classifies a document based on the majority class of its k-nearest neighbours in the feature space. It measures the distance between the test document and the training documents, and the most common class among the k-nearest training documents is assigned as the predicted sentiment. K-NN can be used to find the k-nearest documents in the training data for each test document and then predict sentiment based on the majority sentiment class of those neighbours [17].

$$y^{(x)} = \arg\max_j \sum_{i=1}^{k} w(i,j) \cdot I(y(i) = j)$$

$y^{(x)}$ is the predicted class label for the data point x.
k is the number of nearest neighbours.
i is an index that iterates through the neighbours of x.
j is an index that represents each class labels.
$w(i,j)$ is a weighting factor for the contribution of neighbor i to the class j.
$I(y(i) = j)$ is an indicator function that equals 1 if the class label of the i-th neighbour is j and 0 otherwise.

4. Support Vector Machine
SVM aims to find a hyperplane that best separates the data points of different sentiment classes while maximizing the margin between the classes. It can be applied to text classification by transforming text documents into numerical feature vectors (e.g., using TF-IDF) and then finding the optimal hyperplane to separate the sentiment classes. SVM is a powerful and widely used method in sentiment analysis. It can handle both binary and multiclass sentiment classification tasks. SVM seeks to find a decision boundary that maximizes the margin between different sentiment classes, leading to good generalization performance [18].

$$f(x) = sign(w \cdot x + b)$$

f(x) is the decision function that predicts the class label for a given input vector x.
w is the weight vector that represents the coefficients of the features.
x is the input vector (the feature vector of the data point to be classified).
b is the bias term or intercept.
The sign function (sign) returns either +1 or -1, indicating the predicted class label.

On the other hand, topic modeling aimed to identify the topics discussed by the community regarding the implementation of Regsosek 2022 using two different methods: Latent Semantic Analysis and Latent Dirichlet Allocation which are explained in these paragraphs below:

1. Latent Semantic Analysis (LSA)
LSA is based on linear algebra and aims to discover the underlying semantic structure of a collection of documents. It achieves this by performing singular value decomposition (SVD) on a term-document matrix. This process reduces the dimensionality of the data while capturing the latent semantic relationships between terms and documents. By reducing the dimensionality of the data, it helps in capturing the essential meaning of words and documents. In sentiment analysis, LSA can be used to improve feature extraction, where the lower-dimensional representations can be employed as features in machine learning models [19].

$$X = USV^T$$

X is the term-document matrix, where each row corresponds to a term (word) and each column corresponds to a document. The elements of this matrix typically represent word frequencies (e.g., TF-IDF values).

U is the left singular vectors (term-topic matrix), where each column is a vector that encodes the relationships between terms and latent topics.

S is a diagonal matrix of singular values. It represents the strength of each latent topic.

V^T is the right singular vectors (document-topic matrix), where each row is a vector that encodes the relationships between documents and latent topics.

2. Latent Dirichlet Allocation (LDA)
LDA assumes that documents are mixtures of topics, and topics are mixtures of words. The algorithm iteratively identifies topics within a collection of documents by assigning words to topics and adjusting the topic mixtures to optimize the likelihood of generating the observed documents. LDA is utilized in sentiment analysis to uncover the main topics or themes within a corpus of text. After identifying these topics, sentiment analysis can be performed for each topic separately, allowing for a more focused sentiment assessment [20].

$$P(w,z,\theta,\varphi \mid \alpha,\beta) = \prod_{i=1}^{M} P(\varphi_i \mid \beta) \prod_{j=1}^{N} P(\theta_j \mid \alpha) \left(\prod_{k=1}^{K} P(z_{i,j} \mid \theta_j) P\left(w_{i,j} \mid \varphi_{z_{i,j}}\right) \right)$$

α is a Dirichlet prior concentration parameter that represents the document's topic density, with a higher α, the document is assumed to consist of more topics and produces a more specific topic distribution per document.

β is the same previous concentration parameter that represents the word density of the topic, with a high β, the topic is assumed to consist of most words and produces a more specific distribution of words per topic.

M is number of documents

W is vector which contains the vocabulary corpus

N is number of words

K is number of topics

Ψ is word distribution for each K topics

Φ is topic division for each document i

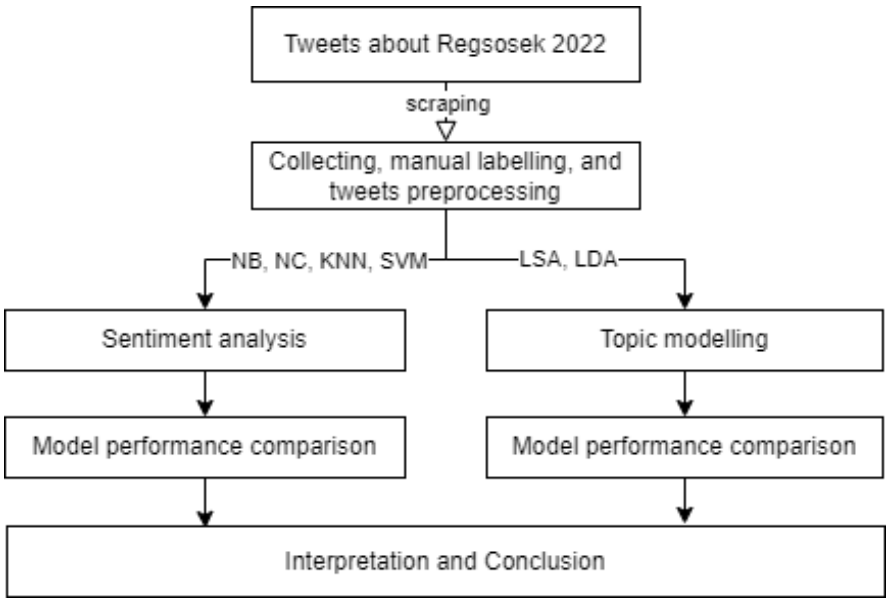

Figure 1. Research Flowchart

Figure 1 depicts the analysis methods used in this study. It begins with the implementation of the Socio-Economic Registration (Regsosek) activity for the entire population of Indonesia in 514 districts/cities. Subsequently, topics related to public opinions on the Regsosek 2022 data collection are discussed in Twitter posts from October 15 to November 14, 2022.

Tweets in the Indonesian language related to the "regsosek" topic were retrieved using web scraping. After cleaning the data, manual labeling was performed by classifying the tweets into two polarity groups: positive and negative.

Once the data was divided into training and testing sets, sentiment analysis was conducted using four models: Naïve Bayes, Nearest Centroid, K-Nearest Neighbors, and Support Vector Machine. The resulting confusion matrix provided Precision, Recall, and F1-Score values, which were used to compare the performance of the developed models.

Next, topic modeling was performed using Latent Semantic Analysis and Latent Dirichlet Allocation to examine the distribution of tweets across their respective topics. Coherence graphs were used to determine the number of topics discussed within the retrieved Twitter corpus. The final step involves interpreting and drawing conclusions from the collected topics as evaluative material for stakeholders involved in the implementation of Regsosek 2022.

5. Results

After collecting data using web scraping techniques, we obtained the results of tweets about the society responding to the implementation of Regsosek 2022 through social media Twitter. The tweets obtained were 539 tweets using the keyword “regsosek” during the period 15 October 2022 to 14 November 2022. At first, we did hand-labeling or labeled the data manually by researchers. In this research, we divided into two labels, negative and positive. The purpose of this labeling is to separate the positive and negative sentiments of the society. From this labeling step, positive sentiment was obtained at about 70.84%, and negative sentiment at about 29.16% from all the training data. The sentiment statistic is shown in Table 1.

Table 1. Total and Percentage of public tweets about Regsosek 2022

Sentiment	Total	Percentage
Positive	311	70,84%
Negative	128	29,16%

The data was prepared through text preprocessing before being processed and analyzed. This stage is an essential stage before doing further data analysis steps. The purpose of doing this preprocessing is to clean and prepare the data corresponding to the research objectives so that good results will be obtained.

5.1. Sentiment Analysis with Naïve Bayes, Nearest Centroid, K-Nearest Neighbors, and Support Vector Machine

Before modeling, the data was divided into training and testing data. We use the training data to learn about the data and build a classifier to map the data to a particular class. On the other hand, we use the testing data to evaluate the classifier model that has been formed before. Testing data used in this study amounted to 18.55% of the entire data.

Data analysis is continued by conducting classifier modeling to map the data to a particular class. This modeling is done using four classification methods: Naïve Bayes, Nearest Centroid, K-Nearest Neighbors, and Support Vector Machine. The results of modeling with the four methods will be compared to see the best method that produces the best F-1 Score after evaluation using testing data.

Table 2. Confusion Matrix Text Classification with Naïve Bayes, Nearest Centroid, K-Nearest Neighbors, dan Support Vector Machine.

Sentiment	Method	Precision	Recall	Accuracy	F1-Score
Positive	*Naive Bayes*	1	0,04	0,51	0,08
	Nearest Centroid	0,68	0,53	0,63	0,59
	K-Nearest Neighbor	0,73	0,22	0,56	0,33
	Support Vector Machine	0,81	0,25	0,59	0,39
Negative	*Naive Bayes*	0,5	1	0,51	0,67
	Nearest Centroid	0,6	0,73	0,63	0,66
	K-Nearest Neighbor	0,53	0,92	0,56	0,67
	Support Vector Machine	0,55	0,94	0,59	0,54

After evaluating the model using testing data, precision, recall, accuracy, and f1-score values were obtained from each model that has been presented in the table above. The Nearest Centroid method gives the best results with a high f1-score that balanced on positive and negative sentiment, about 59% and 66%, respectively. Thus, it can be stated that the Nearest Centroid method is the most suitable method with the best success compared to the three other methods in classifying public sentiment into negative and positive sentiment using the available data.

5.2. Topic Modeling with Latent Semantic Analysis and Latent Dirichlet Allocation

In this section, modeling or forming a topic from words will be carried out to find the most common topic from society's response to Regsosek. In this section, modeling uses two methods, which will be compared based on the results obtained. The two methods are Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA).

Table 3. Topic Modeling with LSA dan LDA

Sentiment	Topic	Words	Interpretation
Positive (LSA)	1	emas, cipta, saya, enak, tanyain, maaf *gold, create, me, delicious, ask, sorry*	-
	2	tanyain, memori, web, mendem, mumeett, nator *ask, memory, web, harbored, dizzy, nator*	-

Sentiment	Topic	Words	Interpretation
	3	genshinku, bekas, cucu, tau, juga, bilang, ngulang, engkas *my genshin, former, grandson, know, too, said, repeated, finished*	-
Negative (LSA)	1	regsosek, tugas, udah, ga, data, aja, bgt, orang, gila, ya *regsosek, tasks, already, no, data, just, really, people, crazy, yeah*	The data collection task is too heavy
	2	gedong, kawasan, sulit, tembus, rumah *building, area, difficult, entry, house*	So difficult to meet respondents at home
	3	gara, gila, sensus, data, trauma, orang *because, crazy, census, data, trauma, people*	Complaints about the difficulty of data collection process
Positive (LDA)	1	regsosek, data, bps, tugas, bupati, sensus, sosial *regsosek, data, bps, tasks, regent, census, social*	The Regent supports the success of Regsosek
	2	regsosek, tugas, rumah, bilang, maksud, kerja, semangat *regsosek, tasks, home, say, mean, work, enthusiasm*	Regsosek staff carry out their duties with enthusiasm
	3	regsosek, ppl, selesai, tugas, alhamdulillah *regsosek, ppl, finished, task, alhamdulillah*	Regsosek data collection is almost complete
Negative (LDA)	1	regsosek, tugas, isi, data, rumah,responden, gak, tidur *regsosek, task, fill, data, home, respondent, no, sleep*	Fatigue officers taking care of respondent data
	2	regsosek, tugas, ppl, sakit, sabar, mental, pusing *regsosek, task, ppl, sick, patient, mentally, dizzy*	The job is too heavy and affects health
	3	regsosek, jam, kerja,tugas, sakit, sibuk,gaji, ppl *regsosek, time, work, task, sick, busy, salary, ppl*	Salary does not match the weight of the work

In this section, modeling or forming a topic from words will be carried out to find the most common topic from society's response to Regsosek. In this section, modeling uses two methods, which will be compared based on the results obtained. The two methods are Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA).

Topic modeling is done on both positive and negative sentiments separately, and for each sentiment in each method, we built on three topics that are most widely discussed by society related to Regsosek. Topic modeling results from both types of methods, LDA and LSA, are then presented in Table 3. Based on the results presented in the table, the results of modeling with LDA are better because the different topics built on the corresponding words are easier to interpret compared to topics built with the LSA method. In fact, a topic for positive sentiment built using the LSA method is hard to interpret by the author because the words presented are considered incompatible with each other or come from a different theme/topic.

6. Conclusions

From the results of the analysis that has been done, it can be concluded that:

1. The best method in the classification of sentiment analysis of people's Twitter tweets about the Socio-Economic Registration 2022 is the Nearest Centroid method, with a relatively high and balanced f1-score on positive and negative sentiment, 59% and 66%, respectively.
2. The LDA method is the best for modeling people's Twitter tweets about Socio-Economic Registration 2022. This is because the topic results from modeling with LDA are more straightforward to interpret than topics built with the LSA method.

In classifying sentiment analysis with various techniques, several techniques have very low F1-score values. This can happen because the preprocessing stage is not good enough. The same thing applies to the output generated by topic modeling using the LDA method on positive sentiment. It cannot be interpreted because the words presented are incompatible or come from the same theme/topic. This can also happen because the preprocessing stage is not good enough. Therefore, for further research on this topic, pay attention and improve it again for the preprocessing stage.

In addition, the results of the classification of sentiment analysis and topic modeling regarding the Socio-Economic Registration 2022 can be used as a reference/input for the government to evaluate the future of other community data collection activities.

References

[1] Zhong, B. (2021). Social Media Communication: Trends and Theories. United Kingdom: Wiley.

[2] Westerman, G., Bonnet, D., McAfee, A. (2014). Leading Digital: Turning Technology Into Business Transformation. United Kingdom: Harvard Business Review Press.

[3] Nadaraja, R., & Yazdanifard, R. (2013). Social media marketing: advantages and disadvantages. Center of Southern New Hempshire University, 1-10.

[4] Fountain, J. E. (2004). Building the virtual state: Information technology and institutional change. Rowman & Littlefield.

[5] SEBASTIAN, T. (2012). Sentiment Analysis for Twitter (Doctoral dissertation).

[6] Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., & Li, X. (2011). Comparing twitter and traditional media using topic models. In Advances in Information Retrieval: 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings 33 (pp. 338-349). Springer Berlin Heidelberg.

[7] Maksood, F. Z., & Achuthan, G. (2016). Analysis of data mining techniques and its applications. International Journal of Computer Applications, 140(3), 6-14.

[8] Birnbaum, M. H. (2004). Human research and data collection via the Internet. Annu. Rev. Psychol., 55, 803-832.

[9] Khder, M. A. (2021). Web Scraping or Web Crawling: State of Art, Techniques, Approaches, and Application. International Journal of Advances in Soft Computing & Its Applications, 13(3).

[10] Han, E. H., & Karypis, G. (2000, September). Centroid-based document classification: Analysis and experimental results. In European conference on principles of data mining and knowledge discovery (pp. 424-431). Berlin, Heidelberg: Springer Berlin Heidelberg.

[11] Tan, A. H. (1999, April). Text mining: The state of the art and the challenges. In Proceedings of the pakdd 1999 workshop on knowledge discovery from advanced databases (Vol. 8, pp. 65-70).

[12] Hendrawan, M. Y., & Projo, N. W. K. (2021). Topic Modeling in Knowledge Management Documents BPS Statistics Indonesia. In Proceedings of The International Conference on Data Science and Official Statistics (Vol. 2021, No. 1, pp. 119-130).

[13] Jannah, Y. A. N., & Prasetyo, R. B. (2022, November). Analisis Sentimen dan Emosi Publik pada Awal Pandemi COVID-19 Berdasarkan Data Twitter dengan Pendekatan Berbasis Leksikon. In Seminar Nasional Official Statistics (Vol. 2022, No. 1, pp. 597-608).

[14] Illia, F., Eugenia, M. P., & Rutba, S. A. (2021). Sentimen Analysis on PeduliLindungi Application Using TextBlob and VADER Library. Proceedings of The International Conference on Data Science and Official Statistics, 2021(1), 278–288.
[15] Webb, G. I., Keogh, E., & Miikkulainen, R. (2010). Naïve Bayes. Encyclopedia of machine learning, 15(1), 713-714.
[16] McIntyre, R. M., & Blashfield, R. K. (1980). A nearest-centroid technique for evaluating the minimum-variance clustering procedure. Multivariate Behavioral Research, 15(2), 225-238.
[17] Peterson, L. E. (2009). K-nearest neighbor. Scholarpedia, 4(2), 1883.
[18] Brereton, R. G., & Lloyd, G. R. (2010). Support vector machines for classification and regression. Analyst, 135(2), 230-267.
[19] Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse processes, 25(2-3), 259-284.
[20] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.