The 6th International Seminar on Science and Technology (ISST) 2020
July 25th 2020, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

Media Sentiment Analysis of East Java Province: Lexicon-Based vs Machine Learning SVM

Ikhwan Rustanto, and Nur aini Rakhmawati
Department of Information System, Institut Teknologi Sepuluh Nopember, Surabaya
e-mail: ikhwanrustanto@gmail.com

***Abstrak*—Indonesian Ministry of Communication and Informatics reported internet users in Indonesia reached 150 million with a penetration of 56% in January 2019. This indicates the era of information disclosure; therefore, information on government performance is more easily obtained by all levels of society. Society is becoming more sensitive to government performance, and more feedback is being given to the government. This large amount of feedback has stimulated research on public sentiment analysis. This study compares the public sentiment analysis by two different approaches to the government performance of East Java Province. The study was comparing the lexicon-based method approach and the Support Vector Machine (SVM) from the machine learning approach. This study uses Twitter and Instagram datasets, and also the online news media web that reports on East Java. This study found that by using a combined data source of social media and online media, the lexicon-based approach produced an accuracy value of 57.7%; while the SVM machine learning method approach produces an accuracy of 44.7%.**

***Kata Kunci*—East Java, Lexicon-Based, Sentiment Analysis, Support Vector Machine.**

I. INTRODUCTION

BEING a country that has a lot of administrative division, making Indonesia has its challenges in the management of government in the area. As many as 34 provinces divided into 416 districts and 98 cities are in the administrative area of Indonesia [1]. These regions have the authority to regulate their administrative management, according to the characteristics and resources of each region.

Data from the Ministry of Communication and Information Technology noted that as of February 2018, internet users in Indonesia reached 142.36 million [2]. An increase of 54.68% compared to the beginning of 2016 increasingly shows the era of information disclosure. The openness of information makes information on government performance more easily obtained by all levels of society.

This information disclosure makes the public more sensitive to government performance. So the more information they get, the more feedback people give the government. The number of media that used to access the internet makes it easier for the public to convey their feedback. When the internet was not yet widespread, the people could only use the voice of readers in the mass media to express their feedback. Now, community feedback conveyed through social media. Both social media from mass media and official social media owned by local governments. Also, Law No. 14 of 2008 concerning public information disclosure requires every government agency and local government to provide information related to policies, plans and realization in government performance.

Almost all governments have media, both in the form of online news media and social media, to disseminate information on government performance. In addition to disseminating information, these media can be used to get feedback from the public. Through this feedback, the government can measure the level of satisfaction of the community towards government performance.

This research will look at community sentiment towards the Government of East Java Province (East Java Provincial Government). East Java Provincial Government itself has an official account to disseminate information and receive feedback from the public, namely the @jatimpemprov Twitter account. There are also some Twitter accounts from the work units below, such as the @humasprovjatim Public Relations twitter account, the @bkdjatim Regional Personnel Twitter account, etc. Also, the East Java Provincial Government have accounts on other social media, namely Instagram with @humasprovjatim and @jatimpemprov accounts. News about the East Java Provincial Government will also be obtained from different online media including Tribunnews, Detik and Kompas.

In this study, feedback from the community will be divided into three. First is satisfaction, which can be measured from positive opinions. Second is dissatisfaction, which can be measured by negative opinion. While the third neutral, which is not a positive or negative opinion. Positive, negative and neutral opinions can be extracted from feedback and reports entered into these accounts, through Sentiment Analysis (SA) techniques. Sentiment Analysis is one technique to analyze opinions, sentiments, appreciation, emotions towards a product, service, organization, individual and their attributes [3].

Based on a research survey of SA techniques, Medhat et al. (2014) categorize SA into two approaches. Namely, the machine learning method and the lexicon-based method [4]. This study will use a lexicon-based method and compare with Support Vector Machine (SVM), which is a machine learning approach. The lexicon-based method is used because it can be able to share document sentiments in real-time without being limited to the topics discussed in the document being assessed. The SVM comparison is used because it is a method widely used in SA with machine learning.

Table 1.
Example of Sentiment labelling results

#	Source	Text	Sentiment
1	Twitter	Malang banjir, mohon untuk menyelesaikan masalah banjir @KemensosRI @jokowi @KemenPU @OmbudsmanRI137 @DPR_RI @PDemokrat @Pak_JK @JatimPemprov @PemkotMalang @PemkabMalang	Negative
2	Twitter	Min update terkini banjir yg melanda madiun, ngawi & magetan dong	Negative
3	Twitter	Selamat berkarya melayani masyarakat SDA prov jatim , semoga banjir segera redah	Positive
4	Twitter	Gubernur Jawa Timur, Khofifah Indar Parawansa menghimbau seluruh masyarakat mewaspadai potensi fenomena hidrometeorologi hingga pertengahan Maret. Hidrometeorologi dimaksud meliputi banjir, rob, banjir bandang, tanah longsor, dan angin puting beliung. @KhofifahIP @bpbd_jatim	Positive

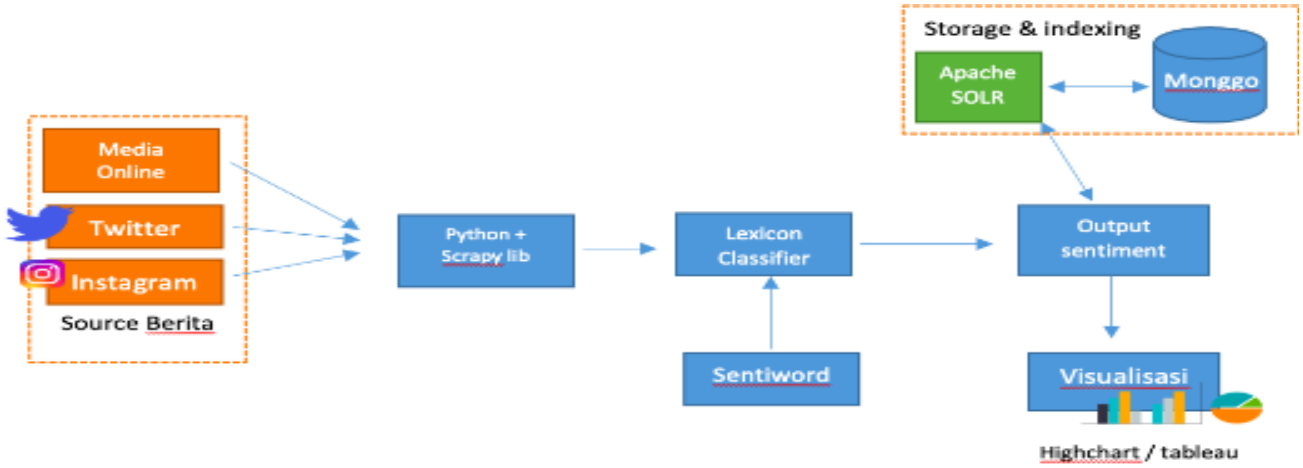

Figure 1. Schematic sentiment analysis process.

Systematics of writing in this study consists of 4 (four) parts, namely: (1) introduction; (2) methodology; (3) results and discussion; and (4) Conclusion.

II. METHODOLOGY

The research process is divided into several stages. Stage (1) Literature study (2) data acquisition from predetermined news sources. Stage (3) labelling the initial data (4) sentiment analysis process using the lexicon-based method. Stage (5) displays the visualization of results using web media. More details will be discussed in each section.

A. Literatute Study

The literature study stage contains the stages of searching research references, theories and concepts used as a basis for research. This stage also includes previous research on sentiment analysis.

1) Sentiment Analysis

Sentiment analysis is one part of opinion mining [5] which is used to extract polarity from existing opinions. Existing polarity can be positive, negative or neutral polarity. Medhat et al. (2014) In his survey, categorized sentiment analysis methods into two techniques. Namely, the machine learning method and the lexicon-based method.

In machine-learning methods, existing models must be trained with the initial dataset. The dataset used must have an initial polarity that can be assessed by manual assessment. This initial dataset already has an opinion label for each of its inputs, so that it will be predefined data training for machine learning algorithms. After training the algorithm is complete, the algorithm can be used to analyze the test dataset or the actual data.

In the lexicon-based method, analysis is carried out using a sentiment dictionary to assess the sentiments of text data. In the lexicon-based approach, sentiment assessment is based on a dictionary of words and word phrases. Are a word and expression a positive value or a negative value [6]. The lexicon-based method itself can be further categorized into prior-polarity and contextual polarity. Sentiment dictionary, which is generally used for sentiment analysis with the lexicon-based approach is SentiWordNet. This dictionary is the result of notes automatically created in the English language lexical database [7].

2) Sentiment Analysis

Many studies discuss sentiment analysis. However, research using the lexicon-based method is less than using machine learning. This lack of research can be caused by the lack of a sentiment dictionary with Indonesian. SentiWordNet itself uses English as the primary language in its dictionary.

Kusumawati's research on the analysis of Indonesian-language sentiments was conducted by Kusumawati to examine public sentiment for the increase in cigarette prices. In this study, researchers used Twitter data to assess public sentiment [8]. Rosdiansyah uses the k-NN method combined with a lexicon to evaluate the polarity of Twitter data on the Jokowi keyword [9]. A combination of lexicon-based and double-propagation was investigated by Buntoro to assess the polarity of the Twitter data with the Jokowi and Prabowo keywords. [10]. Also, research with backpropagation with

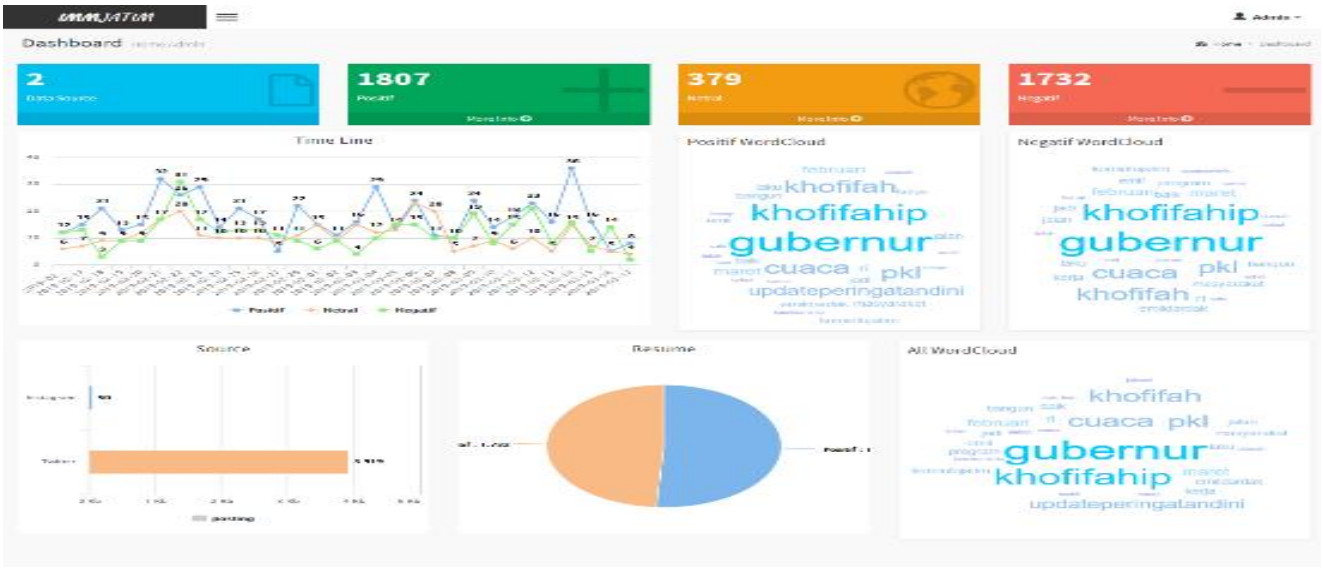

Figure 2. Visualization of Statistical Sentiments in Social Media.

Table 2.
Example result of sentiment

#	Source	Text	Sentiment
1	Twitter	Malang banjir, mohon untuk menyelesaikan masalah banjir @KemensosRI @jokowi @KemenPU @OmbudsmanRI137 @DPR_RI @PDemokrat @Pak_JK @JatimPemprov @PemkotMalang @PemkabMalang	Negative
2	Twitter	Giat penanganan banjir Kali Jatiroto, Selasa 25/12, mobilisasi alat berat dan bahan banjiran ke lokasi @bpbd_jatim @JatimPemprov pic.twitter.com/2IuZna4YYL	Negative
3	Twitter	Neng Ita Prioritaskan Infrastruktur Penanggulangan Banjir dan UKM @JatimPemprov @humas_kota #inilahmojokerto #mojokerto https:// inilahmojokerto.com/10/12/2018/nen g-ita-prioritaskan-infrastruktur-penanggulangan-banjir-dan-ukm/ …	Negative
4	Twitter	Selamat berkarya melayani masyarakat SDA prov jatim , semoga banjir segera redah ,	Positive
5	Twitter	Gubernur Jawa Timur, Khofifah Indar Parawansa menghimbau seluruh masyarakat mewaspadai potensi fenomena hidrometeorologi hingga pertengahan Maret. Hidrometeorologi dimaksud meliputi banjir, rob, banjir bandang, tanah longsor, dan angin puting beliung. @KhofifahIP @bpbd_jatim	Positive

lexicon-based features and Bag of Word was conducted by Munir et al. to identify hate speech on Twitter. [11].

Research using the Support Vector Machine (SVM) was conducted by Indrayuni to analyze hotel review sentiments [12]. This research uses SVM based on Particle Swarm Optimization as a tool for feature selection. A study by Faradhillah et al using SVM compared to Naïve Bayes to analyze the level of community satisfaction with the Surabaya City Government [13].

Another study conducted by Devi et al. which uses SVM based on Feature Selection and Semantic Analysis. In her research, Devi used a film review dataset. In his research, Devi used SVM as the main algorithm compared to Naïve Bayes. They use RBF kernels from SVM with modified based on hyperparameters, where soft margin constants C and gamma γ [14].

B. Data Acquisition

The collection of initial data precedes the sentiment analysis process. The data is taken from social media which is carried out using the Intelligent Media Monitoring (IMM) system. News sources come from Twitter data, Instagram of East Java Provincial Government, and online media web that preach East Java Provincial Government. The dataset is crawled using Python and the Scrapy library.

The keywords used for crawling from Twitter are the hashtags #jatimpemprov and #humasprovjatim. Also from the account @jatimpemprov and @humasprovjatim which is the official account of the East Java Provincial Government. While from Instagram, it is taken with the keyword 'teak provincial government'. For online media, news data will be taken from the media Detik, Tribunnews and Kompas. News will be crossed from the official website of the media using the keyword 'East Java'.

C. Initial Data Labeling Process

The next step is to label the initial data. The data labelling process is done manually. Data assumed to have the intention of praise, appreciation, motivation and gratitude will be marked as positive sentiment. While data showing disasters, complaints, and disagreements will be marked as a negative sentiment. Examples of data labelling results can be seen in the table 1:

D. Sentiment Analysis Process

In this study, the process of identifying opinions and sentiments from the dataset is done using sentiment analysis with the lexicon-based method and compared with the SVM

Table 3.
Performance of lexicon classification results for online media

Value	Classification		
	Positive	Neutral	Negative
Recall	0.731722429	0.227310575	0.655172414
Precision	0.705917513	0.731903485	0.347773032
F-Measure	0.718588378	0.346886912	0.454364289

Table 4.
Performance of lexicon classification results for social media

Value	Classification		
	Positive	Neutral	Positive
Recall	0.870431894	0.2	0.648464164
Precision	0.715846995	0.285714286	0.822510823
F-Measure	0.785607196	0.235294118	0.72519084

Table 5.
Performance of lexicon classification results for the combination of the two data source

Value	Classification		
	Positive	Neutral	Positive
Recall	0.753524804	0.227085054	0.653482373
Precision	0.707699853	0.723684211	0.406417112
F-Measure	0.729893778	0.345694532	0.501153973

Table 6.
Performance of SVM classification results for the combination of the two data source

Value	Classification		
	Positive	Neutral	Positive
Recall	0.715698393	0.391608391	0.505030181
Precision	0.672473867	0.626865671	0.414191419
F-Measure	0.693413173	0.482065997	0.455122393

method of the machine learning approach.

1) Lexicon Sentiment Analysis

In sentiment analysis using a lexicon-based method, sentiment analysis is carried out using a dictionary as a word assessor. This method relies on the word opinion or sentiment, where the words in a post express positive, negative or neutral sentiments. This method determines the sentiment or polarity of opinion through several functions of the word opinion in a document or sentence. This approach can also be called a dictionary-based approach.

The words from posting data will be judged to determine the opinion of the words that are there. Words that indicate pleasure or approval, such as extraordinary and kind, show positive sentiment. Whereas words that indicate things that are not desired, such as bad or evil, have negative sentiments or negative polarity. Usually, sentiments are found in adjectives and adverbs, but some verbs and nouns contain sentiment values. So this method uses a dictionary containing opinion words to determine a positive or negative sentiment from a post. This dictionary is commonly called the opinion lexicon. In this study, the opinion dictionary used to provide the value of a sentiment is to use SentiWordNet.

2) Analysis of SVM Sentiments

For data analysis using Support Vector Machine (SVM) data that has been given positive and negative labels are processed using python language tools. Data is divided into training data and testing data, in SVM algorithm training data uses 40% of the total data. SVM calculations use linear kernels with C = 100 and = 0.01 using the Sklearn library in Python. The results of the SVM algorithm are validated with ten fold-cross validations to find the accuracy of the algorithm.

E. Visualization of Results

The results of sentiment data will be displayed using one of the tools available on the web, namely, by using the highchart plugin. Statistics shown include the resume of each polarity, daily polarity timeline, wordcloud of each polarity and total data from each data source. Overall, from data acquisition, then sentiment analysis and data visualization in the sentiment analysis process in the IMM application can be seen in figure 1

III. RESULT AND DISCUSSION

In the results and discussion section, the results of the analysis of media sentiment analysis of the East Java Provincial Government will be explained. This section will also discuss research results and validate research results.

A. Research Results

From the results of this sentiment analysis research, based on data taken from 30 November 218 to 17 March 2019, data obtained through the crawling process included 3918 social media posts and 664 online news media data. From these results, obtained 1807 social media posts have positive polarity, 379 neutral polarity and 1732 posts have a negative polarity. While from online news media data, 396 news gets positive polarity, 15 news with neutral polarity and 253 get negative polarity. Data will continue to change because crawling will continue to run. Crossed data is entered into the MongoDB database, and indexed with Apache Solr. With the Apache Solr indexing, it will speed up data searching in the

Mongo DB database. Data that has received the sentiment polarity value will then be displayed using highchart. Visualization will be divided into two, namely visualization for social media and visualization for online media. And for each visualization will display statistics on resume data, both positive and negative polarity, the amount of data each day, data entry data sources and wordcloud for each polarity. To see posts related to wordcloud and resume data, you can do this by pressing the resume or wordcloud you are looking for. For visualization of sentiments in social media can be seen in Figure 2. At the same time, the visualization of sentiments in social media can be seen in Figure 2.

In table 2, you can see at a glance the data obtained from Twitter that is displayed to IMM application users. The data has a rating of positive or negative sentiment. The data collected can be seen based on the data source, and the date the data was posted.

From the experimental results, the combined prediction results obtained from social media and online media data sources with the lexicon approach produce 58% accuracy. This is because the accuracy of the classification with social media data sources is very low, which is 55%. In comparison, the classification with online media sources is quite high at around 75%. This shows the classification with online media sources has better accuracy than with social media data source. While using the SVM approach, the accuracy obtained from the combined data was 44.7%. Predictions with the SVM approach are not separated because the amount of online media data is too small to be used as training data.

B. Research Validation

Results of sentiment analysis obtained excellent results using online media data source. But despite having optimal accuracy, it does not always indicate that the analysis has good performance. A confusion matrix is used to calculate recall, precision and f-measure to validate the performance of sentiment analysis.

The recall, precision and f-measure values obtained from each data source with the lexicon approach can be seen in table 3 for online media, table 4 for social media and table 5 for combined data sources. From this value, for classification with online and social media data source sources, positive class shows the highest performance. While the neutral class shows the value with the lowest performance but has the highest precision value. Whereas the SVM machine learning approach can be seen in table 6, where positive classes dominate the values of Precision and Recall. For the machine learning approach SVM only uses a combined data source because the total online media data is too little to be used as training data.

The combined data also shows positive class classification performance, which has the highest performance. And the neutral class that has the highest precision. In the neutral class, precision is best because a lot of data are classified as neutral according to the actual class, compared to the total predictions for the neutral class. While neutral has a low recall because of actual neutral class data, very few neutral data can be classified through the classification process. This is due to a lot of data from positive and negative classes which turns into neutral classes.

IV. CONCLUSION

This study is a sentiment analysis using a lexicon-based method compared to using SVM from a machine learning approach regarding people's perceptions of the government of East Java. Community sentiments are classified into three classes, positive, neutral and negative. From the results of the study showed the classification performance with the lexicon-based method still could not be called good, but it had a better accuracy value than the classification using the machine learning approach. By using a combined data source of social media and online media obtained an accuracy value of only 58%, with the highest precision 72% for neutral classes and the highest recall of 75% for positive classes. For the machine learning SVM approach, the accuracy value is 44.7%. Highest precision in the positive class of 67.2% and the highest recall in the positive class also with a value of 71.56%.

In the lexicon approach, the neutral class performance value is the lowest compared to other classes, even though it has high precision. This is because a lot of data that is neutral has not been successfully classified as neutral by the lexicon-based method. The absence of preprocessing processes can be a significant trigger for the emergence of misclassification. Because in this study, a lot of data or posts that are not related to the delivery of opinions or news related to the East Java Provincial Government, because it uses target crawling keywords. In this study, the data is included in neutral opinion, while the results of the classification can be positive or negative depending on the final sentiment value. In the classification using SVM a lot of data is deleted because it only has content in the form of a URL, and because of lack of training data, so the classification results are not very good. So for further research, can use preprocessing to delete data that is not related to classification with the lexicon approach and the addition of datasets to the machine learning approach.

REFERENCES

[1] "OTDA Kemendagri," *Jumlah Provinsi, Kab, Dan Kota*, 2018. https://otda.kemendagri.go.id/seputar-otda/jumlah-provinsi-kab-dan-kota/ (accessed Nov. 02, 2018).

[2] "Kementerian Komunikasi dan Informatika.," *Jumlah Pengguna Internet 2017 Meningkat, Kominfo Terus Lakukan Percepatan Pembangunan Broadband*, 2018. https://kominfo.go.id/index.php/content/detail/12640/siaran-pers-no-53hmkominfo022018-tentang-jumlah-pengguna-internet-2017-meningkat-kominfo-terus-lakukan-percepatan-pembangunan-broadband/0/siaran_pers (accessed Nov. 02, 2018).

[3] O. Shepelenko, "Opinion Mining and Sentiment Analysis using Bayesian and Neural Networks Approaches," 2017.

[4] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.* vol. 5, no. 4, pp. 1093–1113, 2014.

[5] D. Osimo and F. Mureddu, "Research challenge on opinion mining and sentiment analysis," *Univ. Paris-Sud Lab. LIMSI-CNRS Bâtim.*, vol. 508, 2012.

[6] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," 2005.

[7] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.,"

in *LREC*, 2010, vol. 10, pp. 2200–2204.

[8] I. Kusumawati, E. W. Pamungkas, S. Kom, and M. Kom, "Analisa Sentimen Menggunakan Lexicon Based Untuk Melihat Persepsi Masyarakat Terhadap Kenaikan Harga Rokok Pada Media Sosial Twitter," PhD Thesis, Universitas Muhammadiyah Surakarta, 2017.

[9] D. Rosdiansyah, "Analisis Sentimen Twitter Menggunakan Metode K-Nearest Neighbor dan Pendekatan Lexicon," PhD Thesis, Universitas Islam Negeri Sultan Syarif Kasim Riau, 2014.

[10] G. A. Buntoro, T. B. Adji, and A. E. Purnamasari, "Sentiment Analysis Twitter Dengan Kombinasi Lexicon Based Double dan Propagation," *CITEE 2014*, pp. 39–43, 2014.

[11] M. M. Munir, M. Fauzi, and R. Perdana, "Implementasi metode backpropagation neural network berbasis lexicon based features dan bag of words untuk identifikasi ujaran kebencian pada twitter," 2018.

[12] E. Indrayuni, "Analisa Sentimen Review Hotel Menggunakan Algoritma Support Vector Machine Berbasis Particle Swarm Optimization," *EVOLUSI-J. Sains Dan Manaj. AMIK BSI Purwok.*, vol. 4, no. 2, 2016.

[13] N. Y. Faradhillah, R. P. Kusumawardani, and I. Hafidz, "Eksperimen Sistem Klasifikasi Analisa Sentimen Twitter pada Akun Resmi Pemerintah Kota Surabaya Berbasis Pembelajaran Mesin," *SESINDO 2016*, vol. 2016, 2016.

[14] D. N. Devi, C. K. Kumar, and S. Prasad, "A feature-based approach for sentiment analysis by using support vector machine," in *Advanced Computing (IACC), 2016 IEEE 6th International Conference on*, 2016, pp. 3–8.