Contents lists available at GrowingScience

International Journal of Data and Network Science

homepage: www.GrowingScience.com/ijds

Sentiment analysis of Saudi e-commerce using naïve bayes algorithm and support vector machine

Mohammed Shenify[a*]

[a]*College of Computing and Information, Albaha University, Saudi Arabia*

C H R O N I C L E

Article history:
Received: November 20, 2023
Received in revised format: January 2, 2024
Accepted: March 6, 2024
Available online: March 6, 2024

Keywords:
Sentiment analysis
Social Media
e-Commerce
Naïve Bayes
Support Vector Machine

A B S T R A C T

The Covid-19 pandemic which has spread across all countries, including Saudi Arabia, has caused the government to create limited curfew policies in the country that affected the economy. This policy has given rise to a new trend in society, namely the habit of shopping online. The trend of purchasing online via e-commerce increases. However, people's opinions and attitudes towards this trend vary. Therefore, this research was conducted with the aim of determining the subjectivity of public opinion or sentiment on the e-commerce activities using probability and statistical approaches, i.e.: the Naïve Bayes (NB) and Support Vector Machine (SVM) classifiers. Three experimental scenarios of dataset splitting for training and testing; 90%:10%; 80%:20%; and 70%:30%. The comparison of accuracy values was carried out using an automatic labeling method. Experimental results show that the 70%:30% split scenario provides the best result, with 89% of accuracy, 99.7% of Precision, 88% of Recall and 93.5% of F1-score for the SVM classifier.

1. Introduction

In March 2020, the World Health Organization (WHO) stated that the world was facing the Covid-19 pandemic which had a direct impact on many aspects of society. In Saudi Arabia, to prevent the spread of the Covid-19 virus, the government chose a new policy for limited curfew during the beginning of the pandemic. As a result of this policy, social dialogue and geographical constraints are limited and have an impact on reducing the community's ability to carry out economic activities. This government policy has also given rise to new trends or habits in society, namely the habit of online shopping. People are starting to buy online with the aim of reducing outside activities and avoiding social interactions. Apart from being more convenient and easier, most people believe that purchasing on the internet is cheaper. Moreover, currently many e-commerce players do not hesitate to provide discounts and free shipping during certain periods.

There are many opinions in the community about online shopping in e-commerce, both positive and negative opinions conveyed via social media Twitter. These views or opinions can be examined to determine the subjectivity of opinions, and the findings from this analysis are referred to as sentiment analysis. Sentiment Analysis is the activity of analyzing someone's opinions, opinions, attitudes or emotions regarding a particular product, topic or problem so that it can be seen whether this is a positive, negative or neutral sentiment (Afshoh & Pamungkas, 2017).

In this research, the author hopes that the model that has been created for classifying public opinion sentiment on Twitter towards e-commerce in Saudi Arabia will be able to gain insights for considering strategies to manage e-commerce policy during pandemic in the future. This research uses the Naïve Bayes Classifier (NBC) and Support Vector Machine (SVM) classifier because it uses probability methods, the chance of words appearing and is suitable for text data. Meanwhile, the SVM classifier is suitable for text classification and how the algorithm works can deal with outlier data. NBC is a classification

* Corresponding author.
E-mail address maalshenify@bu.edu.sa (M. Shenify)

ISSN 2561-8156 (Online) - ISSN 2561-8148 (Print)

doi: 10.5267/j.ijdns.2024.3.006

method that eliminates Bayes' theorem. The classification method used is a probability and statistical method by predicting future opportunities based on previous experience, so it is known as Bayes' Theorem (Tuhuteru & Iriani, 2018). Meanwhile, Support Vector Machine (SVM) is a two-class classification technique, which tries to predict each class given a set of input data (features) (Ayumi & Fanany, 2016). The classification concept with SVM is looking for the best hyperplane that functions as a separator of two data classes. The SVM algorithm works by optimally separating data using hyperplane distance measurements from the closest points rather than finding the maximum point to maximize the distance between class labels based on class collection limit measurements (Ramayanti & Salamah, 2018).

This research emphasizes binary classification which consists of two classes, namely positive or negative. The Classification Method is used to determine, group, or categorize an unstructured document into one or more previously known groups automatically based on the contents of the document (Indriani, 2014).

The paper is arranged in five parts: Section 1 provides an introduction followed by Section 2 that contains related research. Section 3 discusses the dataset and method, then Section 4 presents the experimental results and discussion. Lastly, Section 5 provides conclusions.

2. Related Works

Sentiment analysis in this research is the analysis of a person's opinions, attitudes, and emotions into written language. One of the media that can be used in sentiment analysis is mass media, where one of the data mining techniques in the form of text is carried out by means of text mining which aims to process text to become information obtained from forecasting patterns and trends through statistical patterns (Luqyana et al., 2018). There has been a lot of previous research carried out regarding sentiment analysis, both using one algorithm, and comparing the performance of one algorithm with another algorithm as researchers will do. Sghaier and Zrigui (2016), discuss sentiment analysis for Arabic e-commerce websites. The authors conducted various experiments using different techniques of stemming and the performance of some classification algorithms to have the combination that gives us satisfactory results.

Several previous studies compared the Support Vector Machine (SVM) algorithm with the Naïve Bayes Classifier (NBC) which found that the SVM algorithm had better accuracy than NBC (Arora et al., 2020; Al-Barznji & Atanassov, 2018; Rama et al., 2020; Al-barznji & Atanassov, 2018; Rama et al., 2015; Indriyani & Wibowo, 2022; Frizka et al., 2021; Petiwi et al., 2022; Setiawan & Utami, 2021; Saepulrohman et al., 2021). However, there are also previous studies that have produced experiments with better accuracy of the Naïve Bayes algorithm than SVM (Ardianto et al., 2020; Tamrakar et al., 2020; Wardhani et al., 2018; Fikri et al., 2020; Supriyadi et al., 2022).

3. Materials and Methods

The process of analyzing sentiment on comments on Twitter regarding a person's rights to E-commerce in Saudi Arabia based on public comments on Twitter can be seen in Fig. 1.

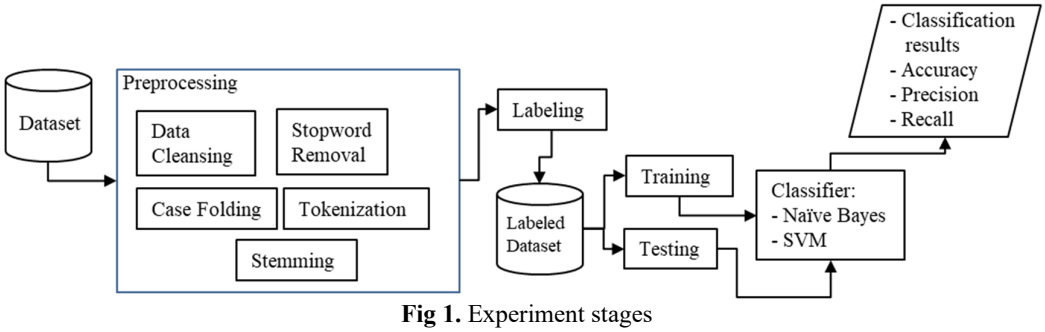

Fig 1. Experiment stages

The initial stage starts from collecting the dataset which continues to the preprocessing stage to clean the raw dataset. After completing preprocessing, enter the labeling stage to find out the positive or negative class. Next, we enter the TF-IDF stage to give weight to each word contained in the dataset. Then enter the classification stage using the Naïve Bayes algorithm and Support Vector Machine.

3.1 Data Preprocessing

Preprocessing is the initial processing of data to prepare text data for the classification process, i.e.: cleaning, case folding, tokenizing, stopword removal and stemming are carried out. The preprocessing stage is done directly in the Rapid Miner software. The whole process includes several steps: cleaning the online text, removing spaces, expanding abbreviations,

stemming, removing stop words, processing negations, and finally feature selection. The following is an explanation of the preprocessing stages.

a. Data Cleaning is the stage where unnecessary characters and punctuation are removed from the text. Cleaning aims to reduce interference/noise in the dataset. The stages in this cleansing are Delete URLs, Delete Usernames, Delete RT tags and Delete Hashtags.
b. Case Folding = Transform case is an operator for case folding. Case folding is changing the form of a word to its basic form so that a character can be uniform (lower case) 4 . Case folding is done to make searching easier. Not all text documents are consistent in their use of capital letters. Therefore the role of case folding is needed in converting the entire text in a document into a standard form (usually lowercase).
c. Tokenization is the process of breaking down sentences into words that are more meaningful and significant. Tokenization separates each word that makes up a sentence.
d. Stopword Removal is the process of removing words that do not describe something that should be removed. The words that are omitted are words that are in the dictionary which contains a list of words (stopword list).
e. Stemming is a stage for carrying out the process of changing words containing infixes or suffixes into basic words that will contain more meaning to obtain information so that comments will be more specific in categorization.
f. Remove Duplicate: Remove Duplicate in this research aims to delete data that has duplicates.
After carrying out the preprocessing stage, the data that has been preprocessed becomes 1324 for E-commerce data in Saudi Arabia.

3.2 Automatic Labeling

At this stage, data that has been cleaned during preprocessing will be given a data class label. Data class labeling is divided into two, positive class and negative class. The labeling process was carried out using the R language via Google Colab tools. The labeling process using the R language requires an Arabic dictionary that is integrated with Google Drive. The Arabic dictionary used consists of two dictionaries, namely Arabic dictionary with positive words and Arabic dictionary with negative words. The automatic labeling stage process is carried out as shown in Fig. 2.

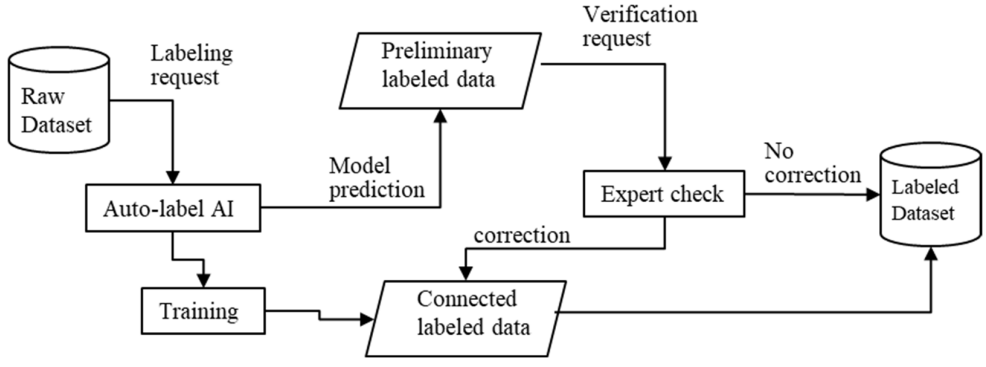

Fig. 2. Process of Automatic Labeling

At the labeling stage, an assessment is carried out on each sentence. Each detected word is given a score to assess the sentiment class. For positive words a value of 1 is given, while for negative words a value of -1 is given. If there are words that are not in the positive or negative dictionary, they are given a score of 0. Assessment is done by counting the number of scores for each word in one sentence. If the value ≥0 then it is labeled as positive tweet sentiment, conversely if the value <0 then it is labeled as negative tweet sentiment. If the score is assessed at 0 then it is labeled as a positive tweet. The positive data class will be given a value of 1, while the negative data class will be given a value of 0. An expert of Arabic language from Albaha University Language Center assists in checking the labeling results.

3.3 Term Frequency Inverse Document Frequency (TF-IDF)

The TF-IDF method is a method for calculating the weight of each word which is most used in information retrieval. This method is also famous because it is efficient, easy and has accurate results. It calculates the term frequency (TF) and inverse document frequency (IDF) values for each token (word) in each document in the corpus (Maarif, 2015).

3.4. Naïve Bayes Classifier

Naïve Bayes classifier is a machine learning algorithm that uses Bayes' theorem to predict the class label of a given input based on the conditional probabilities of the features. It is called naive because it assumes that the features are independent of each other, which may not always be true (Jurafsky & Martin, 2009).

3.5. Support Vector Machine (SVM)

SVM is a type of supervised machine learning algorithm that can perform data analysis and classification. It works by finding the best hyperplane that splits the data points into different classes in the feature space. The hyperplane is selected to have the largest margin between the nearest points of different classes that makes the SVM classifier resistant to outliers and efficient in high-dimensional spaces (Wang & Wang, 2023).

4. Experimental Set up and Results

4.1. Coding and Dataset Creation

The source code stages start from reading the data to validating the two algorithms using *Google Colab* using the R language and Python programming. The R language is used during the labeling process. This research uses source code starting from Preprocessing, Labeling and Classification for 1 case study, i.e.: E-commerce in Saudi Arabia. This research uses a sampling process in the form of a sampling method on the dataset. The source code in this research consists of preprocessing dataset with automatic label classification and Naïve Bayes and Support Vector Machine algorithm classification.

In this research, the dataset is crawled from Twitter during the period of September to December 2023. The data crawling process is assisted by using the Rapidminer tool with the Twitter API. The data taken are tweets that are relevant to the topic; in this case, the top five E-commerce companies in Saudi Arabia is a relevant keyword. The keywords used in this research are "#e-commerce", "#Amazon", "#haraj.com.sa", "#noon.com", "#aliexpress.com", "#temu.com", "#opensooq.com". The data is saved in .csv format with a total of 1324 data records on the Saudi Arabi E-commerce. Table 1 presents the examples of the captured data. Due to the privacy matter, the identity of the company accounts are made anonymous.

Table 1
Examples of records in the Dataset

Date & Time	Text in Arabic	English Translation	Manual labelling
16/09/2023 11:04:43	مرة واحدة ولن تكون هناك مرة ثانية لاستخدام @Xxxxxxxx يكفى استخدام إذا كنت ترغب فى التسوق عبر التجارة الإلكترونية وكانت .Xxxxxxxx @ بعثة هى الرحلة الاستكشافية، فمن الأفضل أن تنتقل إلى رحلة استكشافية Xxxxxxxx أخرى أو إذا كان هذا هو الخيار الوحيد، فمن الأفضل عدم التسوق هههه	It's enough to use @Xxxxxxxx once and there won't be a second time using @ Xxxxxxxx expedition. If you want to shop on e-commerce and the Xxxxxxxx is the expedition , it's better to just change to another expedition or if that's the only choice, it's better not to shop hahaha	Positive
16/09/2023 17:44:01	من الخطورة فتح التجارة الإلكترونية في هذه الساعة حتى الساعات الأولى من الصباح... عرضة للتحقق دون تفكير	It's dangerous to open e-commerce at this hour until the early hours of the morning... prone to checking out without thinking	Positive
19/09/2023 04:17:38	@xxxx cares إلا xxxxالتجارة الإلكترونية لا تقل حجمًا عن شكاوى عملاء ، أليس كذلك؟ لا تستطيع تحمل تكاليف توظيف الموظفينchatbotعبر برنامج	@xxxx cares ecommerce is as big as xxxx customer complaints only via chatbot, right? Can't afford to hire employees?	Negative
26/09/2023 04:18:40	@xxxx @xxxx @xxxx_cares في xxxx ، كيف تم إلغاء تنشيط حسابي ما مشكلتي؟ على الرغم من وجود معاملات جارية، كلما استمر هذا الأمر، قل صحته اليوم. #التجارة_الالكترونية #التجارة_الالكترونية	@xxxx @xxxx @xxxx_cares how come my xxxx account was deactivated, what's wrong with me? Even though there are transactions going on, the more this goes on, the less true it is today. #ecommerce #ecommerce	Negative

The division of data presentation is the separation of training and testing data based on percentages, for example 90%:10% means 90% is training data and 10% is testing data. In this study, the percentage division was divided into three experimental scenarios, i.e.: the first experiment uses a data division of 90%:10%, the second experiment uses a data division of 80%:20% and the third experiment uses a data division of 70%:30%. This research compares the accuracy of the recognition based on the automatic labeling method.

4.2. Results

Table 1 shows the results of the labeling process. During the experiment, the automatic labeling process was unable to determine128 tweets, thus, only 1196 tweets are considered.

Table 2
Automatic labeling results

Sentiment	# of Data
Positive	942
Negative	254
Undefined	128
Total	1324

This research uses Confusion matrix to evaluate the performance of the classification. A confusion matrix is a matrix that displays how accurately a classification model predicts on a testing dataset. It shows the comparison between the true labels and the predicted labels and counts the number of right and wrong predictions for each class. Performance metrics include: accuracy, precision, recall, and F1-score are defined in (1) – (4) (Ting, 2011).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$F1\_Score = \frac{2*Precision*Recall}{Precision+Recall} \tag{4}$$

where, TP is True Positive, FP is False Positive, TN is True Negative, and FN is False Negative. The Confusion matrix from the observation for all scenarios during the experiment is shown in Table 3.

Table 3
Confusion matrix from the experiments

	90%:10%		80%-20%		70%-30%	
	SVM	NBC	SVM	NBC	SVM	NBC
TP	753	684	895	778	938	808
FP	188	257	46	163	3	133
TN	51	69	98	67	126	17
FN	204	186	157	188	129	238

Table 4 shows the experimental results on accuracy, precision, recall, and F-Score of the proposed system. The best result for model performance evaluation using the automatic labeling method is when the split data training vs testing of 70%:30%, with the accuracy value of 66% for the Naïve Bayes classifier and 83% for the SVM classifier.

Table 4
Experimental results

Dataset Splitting	**Precision**		**Recall**		**F1-Score**		**Accuracy**	
	SVM	NBC	SVM	NBC	SVM	NBC	SVM	NBC
90%:10%	81%	72.7%	78%	78.6%	84.5%	77.3%	80%	63%
80%::20%	91%	82.6%	85%	80.5%	87.9%	81.5%	83%	66%
70%:30%	99.7%	86%	88%	77.2%	93.5%	81.4%	89%	69%

5. Discussion

From the overall results of the experiments, it was found that the 70%:30% split scenario SVM classifier provides the best accuracy value during the testing, with an accuracy value of 89%. Meanwhile, Naïve Bayes classifier only achieves 69%. The scenario gives better results because it provides evaluation values that are close to balance after the tuning process. These results are in line with research by Hakami (2023) that conducted an analysis of public sentiment regarding e-commerce conditions in Saudi Arabia during the 2020 pandemic. Using machine learning techniques, this study explores how different aspects affect the online shopping preferences of consumers for popular apps in Saudi Arabia. This study also provides useful insights for the e-commerce sector to enhance their services, customer satisfaction and revenue.

6. Conclusions

From the experiment, it can be concluded that although the number of training datasets is relatively small. The performance evaluation of the Support Vector Machine classifier is considered good, for the 70%:30% split scenario with 89% of accuracy, 99.7% of Precision, 88% of Recall and 93.5% of F1-score for the SVM classifier. The results showed that people positively accept e-commerce, especially during and after the pandemic. As one of the top five countries whose control and management well the pandemic, Saudi Arabia adopted e-commerce very well.

References

Afshoh, F., & Pamungkas, E.W. (2017). Analisa Sentimen Menggunakan Naïve Bayes Untuk Melihat Persepsi Masyarakat Terhadap Kenaikan Harga Jual Rokok Pada Media Sosial Twitter. *Jurnal Muhammadiyah Surakarta, 1*(1), 1-11.

Al-Barznji, K., & Atanassov, A. (2018, May). Big data sentiment analysis using machine learning algorithms. In *Proceedings*

of 26th International Symposium" Control of Energy, Industrial and Ecological Systems, Bankia, Bulgaria.

Ardianto, R., Rivanie, T., Alkhalifi, Y., Nugraha, F. S., & Gata, W. (2020). Sentiment analysis on E-sports for education curriculum using naive Bayes and support vector machine. *Jurnal Ilmu Komputer dan Informasi*, *13*(2), 109-122.

Arora, A., Patel, P., Shaikh, S., & Hatekar, A. (2020). Support vector machine versus naive bayes classifier: A juxtaposition of two machine learning algorithms for sentiment analysis. *International Research Journal of Engineering and Technology*, *7*(7), 3553-3563.

Ayumi, V., & Fanany, M. I. (2016). A comparison of SVM and RVM for human action recognition. *Internetworking Indonesia Journal*, *8*(1), 29-33. doi: 10.13140/RG.2.1.3986.0560.

Fikri, M.I., Sabrila, T.S., & Azhar, Y. (2020). Comparison of Naïve Bayes and Support Vector Machine Methods in Twitter Sentiment Analysis. *SMATIKA Jurnal, 10*(02), 71-6.

Frizka, F., Utami, E., & Al Fatta, H. (2021). Analysis of Opinion Sentiment towards the Covid-19 Vaccine on Twitter Social Media Using Support Vector Machine and Naïve Bayes. *Komtika Journal, 5*(1), 19-25.

Hakami, N. A. (2023). Identification of Customers Satisfaction with Popular Online Shopping Apps in Saudi Arabia Using Sentiment Analysis and Topic modelling. In Proceedings of the 2023 7th International Conference on E-Commerce, E-Business and E-Government (ICEEG '23). Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3599609.3599610

Indriani, A. (2014). Classification of Forum Data using Naïve Bayes Classifier. Pp. 5–10 in *National Seminar on Information Technology Application (SNATI)*.

Indriyani, E. R., & Wibowo, M. (2022). Comparison of the Naïve Bayes Method and Support Vector Machine for Sentiment Analysis towards the Astrazeneca Vaccine on Twitter. *Media Informatika Budidarma Journal, 6*(3), 1545-53.

Jurafsky, N. I. A., & Martin, J. H. (2009). *Speech and Language Processing*. 2nd Ed., Pearson Prentice Hall.

Luqyana, W. A., Cholissodin, I., & Perdana, R.S. (2018). Cyberbullying Sentiment Analysis in Instagram Comments Using the Support Vector Machine Classification Method. *J-PTIIK Journal, Brawijaya University, 2*(11), 4704–13.

Maarif, A. A. (2015). *Application of the TF-IDF Algorithm for Searching for Scientific Works*. B.Sc, Thesis, Dian Nuswantoro University, Semarang, Indonesia.

Petiwi, M.I., Triayudi, A., & Sholihati, I.D. (2022). Go-food Sentiment Analysis Based on Twitter Using Naïve Bayes Method and Support Vector Machine. *Media Informatika Budidarma Journal, 6*(1), 542-50.

Rama, G., Reddy, R., & Mamidi, R. (2015). Resource Creation Towards Automated Sentiment Analysis in Telugu (a Low Resource Language) and Integrating Multiple Domain Sources to Enhance Sentiment Prediction. *Proceedings of The Eleventh International Conference on Language Resources and Evaluation (LREC), European Language Resource Association (ELRA), Miyazaki, Japan. pp. 627-34*.

Ramayanti, D., & Salamah, U. (2018). Complaint Classification Using Support Vector Machine for Indonesian Text Dataset. *International Journal Science Resource Computational Science Engineering and Information Technology, 3*(7), 179–84.

Saepulrohman, A., Saepudin, S., & Gustian, D. (2021). Analysis of WhatsApp Application User Satisfaction Sentiment Using the Naïve Bayes Algorithm and Support Vector Machine*. The Best: Accounting Information Systems and Information Technology Business Enterprise*, *6*(2), 91-105.

Setiawan, H., & Utami, E. 2021). Post-Covid-19 Online Lecture Twitter Sentiment Analysis Using Support Vector Machine and Naïve Bayes Algorithms, *Komtika Journal*, *5*(1), 43-51.

Sghaier, M. A., & Zrigui, M. (2016). Sentiment analysis for Arabic e-commerce websites. *2016 International Conference on Engineering & MIS (ICEMIS), Agadir, Morocco*, 2016, pp. 1-7, doi: 10.1109/ICEMIS.2016.7745323.

Supriyadi, R., Maulidah, N., Fauzi, A., Nalatissifa, H., & Diantika, S. (2022). Application of the Naive Bayes Algorithm and Support Vector Machine in Predicting Autism. *SWABUMI, 10*(1), 55-9.

Tamrakar, S., Bal, B. K., & Thapa, R. B. (2020). Aspect Based Sentiment Analysis of Nepali Text Using *Support Vector Machine* and *Naive Bayes*. *Technical Journal, 2*(1), 22-29.

Tuhuteru, H., & Iriani, A. (2018). Sentiment Analysis of the Ambon Branch of the State Electric Company Using Support Vector Machine and Naive Bayes Classifier Methods. *JIP-IT*, *3*(3), 394–401. doi: 10.30591/jpit.v3i3.977

Wang, N. Li, G., & Wang, Z. (2023). Fast SVM Classifier for Large-Scale Classification Problems, *Information Sciences*, 642.

Wardhani, N. K., Rezkiani, S. K., Setiawan, H. E. N. D. R. A., Gata, G. R. A. C. E., Tohari, S., Gata, W. I. N. D. U., & Wahyudi, M. O. C. H. A. M. A. D. (2018). Sentiment analysis article news coordinator minister of maritime affairs using algorithm naive bayes and support vector machine with particle swarm optimization. *Journal of Theoretical and Applied InformationTechnology*, *96*(24), 8365-8378.