Sentiment Analysis of Community Opinion on Online Store in Indonesia on Twitter using Support Vector Machine Algorithm (SVM)

Sentiment Analysis of Community Opinion on Online Store in Indonesia on Twitter using Support Vector Machine Algorithm (SVM)

H Syahputra

Mathematics Department, Universitas Negeri Medan, Jl. Willem Iskandar Pasar V Medan Estate 20211, Indonesia

Corresponding Author: hsyahputra@unimed.ac.id

Abstract. Public opinion on the Online Shop service company in Indonesia can be seen through comments sent via tweeter. The purpose of this study is to see the public opinion sentiment towards online stores in Indonesia via Twitter using the Support Vector Machine algorithm. Public opinion and assessment on Twitter can be classified into 3 classes, namely negative, neutral, and positive. The approach used is Multiclass One Vs Rest SVM by using Kernel Sigmoid, Linear, and RBF to classify public tweets about Indonesian Online Shop services, namely Shopee, Tokopedia, Bukalapak, and JDid. The results obtained are the Multiclass (One Vs Rest) Support Vector Machine kernel algorithm which has the best accuracy is the Sigmoid kernel with 82% accuracy on the Shopee online store dataset, 94.7% on the Tokopedia dataset and 75.3% for the Bukalapak dataset, while the Linear kernel Jdid dataset provides better accuracy, namely 78%. Clarification errors can occur due to overfiting, namely the model adjusts the training data very well (accuracy can reach 100%) so that the model cannot generalize well on the testing data.

1. Introduction

In this modern era, the use of the internet has become a natural thing for the world community to do whatever they want. In an era like today, humans like all things that are practical and automatic to carry out their survival, especially in terms of carrying out buying and selling transactions. Lately, Online Stores have often colored cyberspace trade. Online shop is an electronic commerce where consumers directly buy goods from the seller through a website on the internet where transactions are carried out without intermediary services [1].

The development of technology and information has also resulted in social media becoming the most popular means of communication, so that currently people tend to provide opinions, criticisms, and suggestions through social networking media and one of them is Twitter. According to statistics, Twitter is the fastest growing social network since 2006. According to the MIT Technology Review (2013), Indonesia is the third largest contributor to tweeting with 1 billion tweets, behind the United States (3.7 billion) and Japan (1.8 billion) [2].

Tweet, which is the status text of a Twitter media account user, can generally contain information about the user's identity, conversation, and user feelings. For example, on the Indonesian Online Store service, it is through this tweet that the public can convey an opinion or assessment of the services provided by the Online Store. Therefore, it can be used the application of machine learning methods,

namely text mining to classify the polarity of the opinion. The polarity found is a pattern in unstructured textual data in a document [3]. One of the analyzes in text mining is data sentiment analysis. Data sentiment analysis or opinion mining is a field of study to analyze opinions, sentiments, evaluations, judgments, human attitudes, and emotions towards entities such as products, services, organizations, individuals, problems, events, topics, and their attributes [4].

Much research has been done on sentiment analysis before. Some machine learning techniques that can be used include the Naive Bayes Classifier, Decision Trees, and Support Vector Machine. Research on sentiment analysis using a dataset from twitter was conducted by Nugroho [5]. In 2018, Kautsar Ramadhan also conducted research on Twitter sentiment analysis entitled "Sentiment Analysis on Online Stores Using Naive Bayes on Twitter Social Media", this research analyzed public opinion sentiment regarding Tokopedia and Lazada Online Stores using the Naive Bayes Classifier algorithm.

The next research that became the author's reference in compiling this research was the research conducted by Athoillah [6]. This research discusses the classification using the Support Vector Machine (SVM) algorithm to classify images of two-wheeled transportation objects with four or more wheeled transportation objects. Another study conducted by Vijayarani [7] explained that the SVM algorithm has a better classification accuracy than the Naive Bayes algorithm.

Based on previous studies, the authors are interested in conducting further sentiment analysis about Online Stores using the SVM algorithm related to public opinion on services provided by Online Stores in Indonesia, namely Shopee Online Shop, Tokopedia, Bukalapak, and Jdid on social media using the Multiclass Support Vector Machine algorithm.

2. Methods

2.1. Data collection

The data to be used is secondary data obtained directly from social media twitter, in the form of tweets or posts regarding comments and opinions of the Indonesian people to services provided by the Indonesian Online Shop. The data taken are tweets or posts in Indonesian on social media twitter.

2.2. Data processing

Data processing is carried out through the following processes:

1. The pre-processing is case folding, data cleaning, language normalization, stop word removal, stemming, and tokenization of all data so that feature extraction can be carried out.
2. Feature extraction is done to get the features (terms) of each tweet for use in the classification model. The feature extraction process is weighting with TF-IDF and feature scaling with the Min-Max Scaler.
3. Classification.
   The classification process carried out in this study is to apply the Multiclass Support Vector Machine (SVM) algorithm to classify data into three sentiments, namely data with negative, positive or neutral sentiment towards services provided by Online Stores in Indonesia.
4. After the classification results are obtained, a model evaluation will be carried out to determine the level of accuracy and error of the system in predicting the sentiment class of the test data.

3. Results and Discussion

3.1 Data Acquisition

The data acquired is a dataset related to tweets about online shop services Shopee, Tokopedia, Bukalapak, and JD.id in Indonesia. The total data for each online store is 600 tweets with 150 training data for each negative, neutral and positive class, and 150 testing data. Examples of tweet data for Shopee's online shop obtained can be seen in Table 1.

Table 1. Example of a Data Acquisition Result Tweet

Data	*Username*	*Tweet*	Label
Training	Upium	Oke fix mungkin mulai sekarang mendingan beli pulsa/ paket data di @ShopeeID lebih murah, dapat koinnya gila2an! Keren.	positive
	Asfahany Hendry	Voucher gratis ongkir di @ShopeeID ilang padahal baru dipake 1. Sedih loh aku kalo dapet barang bagus dan murah tapi harus kena ongkir.	netral
	Kartoki	Kecewa banget gw sama @ShopeeID @ShopeeCare ada pesanan gw yang blm gw trf, jadi males belanja lagi di @ShopeeID #SHOPEEMENGECEWAKAN #shopeexblackpink	negative
Testing	Riski P Dewi	Udah 3x belanja di @ShopeeID dan kusuka banget .Lengkap, banyak pilihan, harganya jauh lebih murah, barang cepat dikirim. Free ongkir lagi	-

3.2 Pre-Processing Data

Data pre-processing is the initial stage of text mining to convert data according to the format required so that it can be processed to the next stages. This research will carry out several pre-processing steps for text data, namely: the case foldings process, the data cleaning process, the language normalization process, the stop word removal process, the stemming process, and the data tokenization. Pre-Processing results can be seen in the Table 2.

Table 2. Pre-processing Results on the Sample Dataset

Data	Tweet Sebelum Pre-Processing	Hasil Pre-Processing	Label
X_1	Oke fix mungkin mulai sekarang mendingan beli pulsa/ paket data di @ShopeeID lebih murah, dapat koinnya gila2an! Keren.	Oke Fix Mending Pulsa Paket Murah Koin gila Keren	Positif
X_2	Voucher gratis ongkir di @ShopeeID ilang padahal baru dipake 1. Sedih loh aku kalo dapet barang bagus dan murah tapi harus kena ongkir.	Voucher Gratis Ongkir Hilang Pakai Sedih Bagus Murah Kena Ongkir	Netral
X_3	Kecewa banget gw sama @ShopeeID @ShopeeCare ada pesanan gw yang blm gw trf, jadi males belanja lagi di @ShopeeID	Kecewa Banget Pesan Transfer Malas Belanja	Negatif

$X_{testing}$	Udah 3x belanja di @ShopeeID dan kusuka banget .Lengkap, banyak pilihan, harganya jauh lebih murah, barang cepat dikirim. Free ongkir lagi	Belanja Suka Banget Lengkap Murah Cepat Kirim Free	-

3.3 Feature Extraction

3.3.1 Weighting of Features (words)

In the initial stage, it is done by counting the terms (words) in each document, in order to obtain the frequency of terms. Then calculate the df value which is the number of documents in which a term (word) appears. Furthermore, the calculation is carried out by finding the idf value for each term with Equation (1). After that, the TF-IDF weight calculation will be carried out for each term.

For example, the word "disappointed", the total number of documents (N) = 4, and the frequency of the appearance of the word "disappointed" in all documents (df) = 1.

a. Calculate idf:

$$idf_{aplikasi} = \log\left(\frac{4}{1}\right) = \log(4) = 0{,}602$$

b. Calculate the weight (w) on document X3:

$$w_{X3,kecewa} = 1 * 0{,}602 = 0{,}602$$

The results of the weight calculation for the *X1, X2, X3, Xtesting* documents can be seen in the Table 3.

Table 3. TF-IDF Weighting Conclusion

		TF				*Df*	*N/df*	*Idf*	Bobot (TF-IDF)			
No	*Terms*/Fitur	X_1	X_2	X_3	$X_{testing}$				X_1	X_2	X_3	$X_{testing}$
1	oke	1	0	0	0	1	4,000	0,602	0,602	0,000	0,000	0,000
2	voucher	0	1	0	0	1	4,000	0,602	0,000	0,602	0,000	0,000
3	kecewa	0	0	1	0	1	4,000	0,602	0,000	0,000	0,602	0,000
4	belanja	0	0	1	1	2	2,000	0,301	0,000	0,000	0,301	0,301
5	fix	1	0	0	0	1	4,000	0,602	0,602	0,000	0,000	0,000
6	mending	1	0	0	0	1	4,000	0,602	0,602	0,000	0,000	0,000
7	gratis	0	1	0	0	1	4,000	0,602	0,000	0,602	0,000	0,000
8	banget	0	0	1	1	2	2,000	0,301	0,000	0,000	0,301	0,301
9	ongkir	0	2	0	0	1	4,000	0,602	0,000	1,204	0,000	0,000
10	pesan	0	0	1	0	1	4,000	0,602	0,000	0,000	0,602	0,000
11	suka	0	0	0	1	1	4,000	0,602	0,000	0,000	0,000	0,602
12	hilang	0	1	0	0	1	4,000	0,602	0,000	0,602	0,000	0,000
13	lengkap	0	0	0	1	1	4,000	0,602	0,000	0,000	0,000	0,602
14	murah	1	1	0	1	3	1,333	0,125	0,125	0,125	0,000	0,125
15	malas	0	0	1	0	1	4,000	0,602	0,000	0,000	0,602	0,000
16	cepat	0	0	0	1	1	4,000	0,602	0,000	0,000	0,000	0,602
17	kirim	0	0	0	1	1	4,000	0,602	0,000	0,000	0,000	0,602

18	free	0	0	0	1	1	4,000	0,602	0,000	0,000	0,000	0,602
19	koin	1	0	0	0	1	4,000	0,602	0,602	0,000	0,000	0,000
20	pakai	0	1	0	0	1	4,000	0,602	0,000	0,602	0,000	0,000
21	gila	1	0	0	0	1	4,000	0,602	0,602	0,000	0,000	0,000
22	sedih	0	1	0	0	1	4,000	0,602	0,000	0,602	0,000	0,000
23	keren	1	0	0	0	1	4,000	0,602	0,602	0,000	0,000	0,000
24	bagus	0	1	0	0	1	4,000	0,602	0,000	0,602	0,000	0,000
25	pulsa	1	0	0	0	1	4,000	0,602	0,602	0,000	0,000	0,000
26	paket	1	0	0	0	1	4,000	0,602	0,602	0,000	0,000	0,000
27	kena	0	1	0	0	1	4,000	0,602	0,000	0,602	0,000	0,000

3.3.2 Feature Scaling (Term)

The process of scaling the feature weights is carried out in order to keep each weight of each term in the value range 0-1, with the aim of avoiding numerical difficulties during the calculation process. Feature scaling is done using the Min Max Scaler method. An example of scaling a feature in the term (word) "disappointed" for an X3 document if known,$x_{X3,kecewa} = 0{,}602$, $x_{min} = 0{,}125$ dan $x_{max} = 1{,}204$ and namely:

$$x'_{X3,kecewa} = \frac{0{,}602 - 0{,}125}{1{,}204 - 0{,}125} = 0{,}442$$

Next, the results of scaling the overall weight of the X1, X2, X3, Xtesting documents can be seen in the Table 4.

Table 4. Feature Scaling Result

Normalisasi Bobot			
X_1	X_2	X_3	$X_{testing}$
0,442114	0	0	0
0	0,442114	0	0
0	0	0,442114	0
0	0	0,163171	0,163171
0,442114	0	0	0
0,442114	0	0	0
0	0,442114	0	0
0	0	0,163171	0,163171
0	1	0	0
0	0	0,442114	0
0	0	0	0,442114
0	0,442114	0	0
0	0	0	0,442114
0	0	0	0
0	0	0,442114	0
0	0	0	0,442114
0	0	0	0,442114
0	0	0	0,442114
0,442114	0	0	0
0	0,442114	0	0
0,442114	0	0	0
0	0,442114	0	0

0,442114	0	0	0
0	0,442114	0	0
0,442114	0	0	0
0,442114	0	0	0
0	0,442114	0	0

After scaling the weights for all the features (words) that exist, the results of the weighting (normalization) will then be used in the formation of vectors for the SVM (Support Vector Machine) classification model.

3.4 Classification with Support Vector Machine with Python Program

In this study, a non-linear SVM algorithm was used using several kernels, namely the Gaussian Radial Basis Function kernel, the Linear kernel and the Sigmoid kernel. The parameters in the SVM were determined by trial and error.

The number of features resulting from feature extraction in each training data for each online store is 1394 features in Shopee's online shop training data, on Tokopedia's training data totaling 1419 features, 1366 features on Bukalapak training data, and on Jdid's training data totaling 1374 features.

3.5 Classification Accuracy

In the Table 5, you can see the highest accuracy given by each SVM kernel for the Shopee online store dataset. The highest accuracy is given by the sigmoid kernel with an accuracy of the testing data of 82% but the accuracy score for the training data with the sigmoid kernel is only 80.20%, which means that the sigmoid kernel can predict the testing data more accurately by avoiding overfiting.

Table 5. Highest Accuracy of Each Kernel (Shopee)

Kernel	Akurasi Klasifikasi	
	Data *Training*	Data *Testing*
Rbf	98,00%	78,00%
Linear	99,80%	74,00%
Sigmoid	80,20%	82,00%

Next, for the accuracy of the classification results of the Tokopedia dataset, it can be seen in the Table. The highest accuracy is also given by the sigmoid kernel with an accuracy of the testing data of 94.7%, but the accuracy score for the training data with the sigmoid kernel is only 84.90%. The kernel that provides the lowest accuracy for the Tokopedia dataset is the Linear kernel with an accuracy of 89.30% but has the greatest training data accuracy compared to the RBF and Sigmoid kernels, which is 99.80%.

Table 6. Highest Accuracy of SVM Kernel (Tokopedia)

Kernel	Akurasi Klasifikasi	
	Data *Training*	Data *Testing*
Rbf	99,30%	92,70%
Linear	99,80%	89,30%
Sigmoid	84,90%	94,70%

Furthermore, for the accuracy of the classification results of the Bukalapak online store dataset, it can be seen in the Table. The highest accuracy is given by the sigmoid kernel with an accuracy of the testing data of 75.30% and the accuracy score for the training data with the sigmoid kernel is 78.7%.

Table 7. Highest Accuracy of SVM Kernel (Bukalapak)

Kernel	Akurasi Klasifikasi	
	Data *Training*	Data *Testing*
Rbf	100,00%	74,00%
Linear	100,00%	68,00%
Sigmoid	78,70%	75,30%

In the Table 8, you can see the highest accuracy given by each SVM kernel for the online store dataset JD.id. The highest accuracy is given by the linear and sigmoid kernels with the same accuracy on the testing data, which is 78% and for the accuracy score on the training data with a linear kernel of 99.1% and a sigmoid kernel of 79.60%. In the online store dataset, JD.id linear kernel with high training accuracy can perform the same classification as the sigmoid which has less training accuracy so that it can be concluded that the linear kernel can better classify testing data on the JD.id online store dataset.

Table 8. Highest Accuracy of SVM Kernel (JD.id)

Kernel	Akurasi Klasifikasi	
	Data *Training*	Data *Testing*
Rbf	99,10%	76,70%
Linear	99,10%	78,00%
Sigmoid	79,60%	78,00%

4. Conclusion

Based on the results of the discussion in the previous chapter, the following conclusions can be drawn:

1. Multiclass (One Vs Rest) Support Vector Machine can classify tweets into Shopee, Tokopedia, Bukalapak, and JDid Online Shop accounts obtained from Twitter into three classes, namely negative, neutral, and positive with the highest accuracy above 75%.
2. In the Multiclass (One Vs Rest) Support Vector Machine algorithm, the kernel that has the best accuracy is the Sigmoid kernel with 82% accuracy on the Shopee online store dataset, 94.7% on the Tokopedia dataset and 75.3% for the Bukalapak dataset, while at The Linear kernel Jdid dataset provides better accuracy of 78%.
3. Clarification errors can occur due to overfiting, namely the model adjusts the training data very well (accuracy can reach 100%) so that the model cannot generalize properly to the testing data. Then in a tweet there is a word that has a greater weight in the class that should not be, which results in the wrong classification of the data.

References

[1] Ramadhan, K., Muslim, K., (2018): Analisis Sentimen terhadap Toko Online mengguakan Naive Bayes pada Media Sosial Twitter, Jurnal E-Proceeding of Engineering, 5(3).
[2] MIT, T., (2013): Language Data Reveals Twitters Global Reach, MIT Technology Review, diakses melalui http://www.technologyreview.com.
[3] Feldman, R., dan Sanger, J., (2006): The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data, Cambrige University Press, Cambridge.
[4] Liu, B., (2012): : Sentiment Analysis and Opinion Mining, Morgan and Claypool Publishers.
[5] Nugroho, D. G., Chrisnanto, Y. H., dan Wahana, A., (2016): Analisis Sentimen pada Jasa Ojek *Online* Menggunakan Metode Naive Bayes, Prosiding SNST, FT, Universitas Wahid Hasyim Semarang.

[6] Athoillah, M., Irawan, M.I., dan Imah, E.M., (2015): Support Vector Machine untuk Image Retrieval, Prosiding Seminar Nasional Matematika dan Pendidikan Matematika,.
[7] Vijayarani, S., dan Dhayanand, S., (2015): Liver Disease Prediction using SVM and Nave Bayes Algorithms, International Journal of Science, Engineering and Technology Research (IJSETR), 4(4).