Event Classification in Surabaya on Twitter with Support Vector Machine

Drajad Bima Ajipangestu
Department of Information Technology Management
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
drajad.19092@mhs.its.ac.id

Riyanarto Sarno
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

***Abstract*— Twitter is a social media that is often used by many people in the world. The information is spread and obtained through social media. For example, there is a company that is organizing a new event that many people need to know. This allows the creation of a system that supports the presentation of user information by detecting certain events from Twitter's social media data. In this study, tweet data will be retrieved using Twitter API and stored in JSON format. Furthermore, there will be a pre-processing which includes the deletion of characters, number, URL, stemming, and lower case. Furthermore, feature extraction is performed using Global Vector for Word Representation. we will classify into four classes, which are Competitions, Seminars, Festivals, and Other events. The classification is using SVM to predict the type of event. There are three experimental methods used, there is SVM C, SVM linear, and SVM Nu. SVM Nu was conducted with changes in the SVC parameters in the form of kernel and Nu to produce the best accuracy. Based on the experiments we have done, the best results are obtained with an accuracy of 85.2% by classification using the NuSVC method with an RBF kernel and nu parameter of 0.2.**

***Keywords*—SVM, Twitter, Text Classification**

I. INTRODUCTION

One of the important characteristics of Twitter is its simple service and is known as a fairly popular news distribution media[1]. Twitter is a social media that is still trending among the people with a total of 321 million active users[2]. The main feature of Twitter is the tweet where users can communicate with other users and tell whatever they want[3]. Social media is a place for users to say something that can be seen by people around the world[4]. So as more and more accounts are trying to make interesting tweets in order to get a lot of accounts that follow or are called followers. The popularity of Twitter causes this social media has been used for a variety of covert purposes, for example as individual protests, opinions, events, activities, distribution of information media.

Twitter is often used as a medium for sharing information among the public. This has caused several communities and individuals who have finally created Twitter accounts to deliver news about events to be held around them. Events can be in the form of information about a competition, seminar, workshop, festival event and others. Twitter is flexible so it makes Twitter is an ideal medium to be used for various event detection. The tweet data obtained can be processed into a basis for detecting an event that occurred in Surabaya. To facilitate the representation of information in Twitter, we can classify and classify each tweet in Surabaya. for example, when we need information for a competition or seminar that is held in Surabaya.

In this paper, it discusses the identification of local events that occurred in Surabaya based on tweet data. Tweets are taken from the Twitter API and saved in JSON format. then proceed with pre-processing which includes the elimination of special characters, casefolding, stemming and stopword removal. In addition, labeling is also done to determine the class of the tweet. Then feature extraction is performed using NLP Standford and for classification, we use Support Vector Machine method.

II. RELATED WORK

A. Support Vector Machine

Nowadays SVM has succeeded in solving real world problems, and providing better solutions compared to conventional methods such as artificial neural networks. besides that many researchers use SVM as a reference method. That's why we try to explore using SVM in this study. SVM works well for unstructured and semi-structured data such as images and text. And changes to the kernel also greatly affect the power of SVM, so here the kernel and some parameters are available for testing data accuracy. We use three methods from SVM to find the best accuracy for our Twitter data.

The SVM concept can help find the best hyperplane that works as a separator for two classes in the input space[5]. The problem that can be discussed is the effort to find a line that addresses the two groups. SVM is one of the best classification methods for unseen data samples.

In this paper, classification uses the SVM multiclass, a middle class separated by more than two classes[7]. SVM tries to find the best hyperplane in the input space. The basic method of SVM is a linear, and subsequently developed so that it can be used in non-linear problems[8]. by incorporating the concept of kernel tricks in high-dimensional workspaces[9]. This development has stimulated research interest in the area of pattern recognition for the full investigation of the potential capabilities of the SVM in terms of application.

B. Kernel

SVM is used by the kernel as a way to find out the right transformation function[9]. Choosing the right kernel function is very important, because this kernel function will determine the feature space where the classifier function will be searched for. As long as the kernel functions are compatible, SVM will operate correctly[12]. Kernels that will be used in this SVM classification include linear, polynomial, RBF, and sigmoid kernels. Each kernel has a different equation[13]. In this tweet classification, the SVM method is used in the scikit-learn library. There are many types of SVM libraries with a variety of different kernels, including the SVC library, NuSVC and linear SVC. SVC and NuSVC have almost the same method, but have different sets of parameters and have different mathematical formulas.

C. Regularization Parameters

Regularization parameters are parameters that determine the amount of penalty due to errors in data classification [14]. In the SVM method, this parameter is known as C. Parameter C is in charge of controlling the tradeoff between margin and classification error which thus helps in increasing output accuracy[15].

The greater the value of C, the greater the penalty imposed for each classification error. C values range from 0 to infinity. The modification for this case is the introduction of the parameter nu which has a ranexpoge between 0 to 1 and represents the lower and upper limits of the number of examples that support vectors and is on the wrong side of the hyperplane[16]. NuSVC Lsibrary is a library for classifying using the SVM method, but by using its regularization parameter, nu. So, for this research NuSVC is used because for the trial, the regularization parameter that will be used is Nu. Nu is the upper limit on the margin error section and the lower limit of the total support vector.

D. Global Vector for Word Representation

GloVe or the abbreviation of Global Vectors for World Representation is a learning algorithm to get vector representations of words produced by Stanford University[17]. GloVe was developed after the appearance of word2vec and reformulated and optimized the word appearance factor in the matrix[18]. An example of the results of the GloVe model train with dimension 50 using the Indonesian language Wikipedia case looks like the one shown in Fig. 1.

The metrix similarity is used to evaluate the nearest neighbor to produce a single scalar that calculates the relation of two words. This simplicity can be a problem because the two words given almost always indicate a more complicated relationship than can be captured by one number[19]. The simplest candidate for a series of numbers is the vector difference between two word vectors, and then the result is showcase interesting linear substructures of the word vector space[17]. GloVe uses a collection of texts using a Wikipedia corpus, in which a collection of texts will be built into a vocabulary and each word in the vocabulary produces a vector which amounts to hundreds of dimensions.

```
388584 50
dan 0.444982 0.010489 -0.237920 0.106180
1.796543 -0.321136 -1.159101 1.506741
0.402018 -1.576963 -0.561928 1.600467 0.227854
1.381892 -1.111252 0.879198 1.598972 -0.473417
0.373547 -1.649145 0.658694 -1.242055 0.320449
0.620280 0.037984 0.29247 0.309581 -2.177434
-1.718750 -0.035637 0.087594 0.635854 -0.671529
0.151672 -0.638005 -0.944322 -0.069826 0.109005
0.584862 0.990185 2.855315 -0.260770 -0.294730
0.072384 0.057275 0.328155 0.975686 0.166217
0.135363 0.151103
yang 0.878525 -0.336221 0.181657 -0.468781
2.026934 -0.326302 -0.582394 2.133491
0.220227 -1.172880 -0.030570 1.753427 -0.142790
1.118214 -1.037040 1.142790 0.957231
-0.167193 -0.126756 -1.301438 0.826168 -1.051016
0.985847 0.584295 -0.610685 0.138893 -0.376630
-2.436371 -1.334595 0.130661 0.636875 0.148972
-0.736501 0.464860 -0.299370 -1.621045 0.525350
0.031106 -0.316752 0.415075 2.299269 0.684297
-0.055856 -0.648499 -0.429378 0.961799 1.052839
0.191329 0.370311 0.348548
```

Fig. 1. Glove with 50 dimentions

E. Performance Evaluation

Confusing matrices help provide information about how correct and how wrong the classification[21]. Table I is the explanation table and the confusion matrix formula in the classification by comparing the predicted class and the actual class[22]. True Positive (TP) is the positive data that is properly classified, True Negative (TN) is the negative data that is correctly classified, False Positive (FP), the negative data but classified incorrectly by the system, and False Negative (FN), the positive data but classified incorrectly by the system.

TABLE I. CONFUSION MATRIX

Actual	**Predicted**	
	Positive	***Actual***
Positive	TP	FN
Negative	FP	TN

Where recall precision and accuracy are formulated as follows:

$$\text{Recall} = \frac{TP}{(TP+FN)} \quad (1)$$

$$\text{Precision} = \frac{TP}{(TP+FP)} \quad (2)$$

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+TN+FN)} \quad (3)$$

III. PROPOSED SYSTEM

The system flow starts when the crawl module that uses the timeline API requests Twitter data to get tweets that contain events that are around Surabaya. Then it will be processed to the next stage in the form of preprocessing, feature extraction with GloVe, classification using SVM, performance testing. The flowchart of general system design can be seen in Fig. 2.

A. Data

The data used for this research is the tweet data that obtained through the @eventsurabaya account timeline using the Twitter API. Twitter data used in this thesis is a tweet data. It can be seen in Fig. 3. The data obtained can be seen in the picture. Tweets that have been taken, will be preprocessed and given a Label, amounting to 4 classes, namely Festival, Seminar, Competition and Others. From a total of 600 tweets that have been recorded manually that going to be divided into 2 parts, which is data train and data testing with a ratio of 8: 2. 480 total data for data train and 120 data for testing data that are randomly balanced.

B. Preprocessing

Pre-processing in text processing is important so that the data is ready to be processed which will also improve the results of the research and eliminate the noise word of the sentence[23]. Twitter data can be overcome by applying several types of preprocesses that match the characteristics of the data. Examples of preprocessing and labeling in tweets can be seen in Table II. The preprocesses used in this system include; lower case, Remove Stopword, Stemmer, Remove all number, Remove URL, remove character.

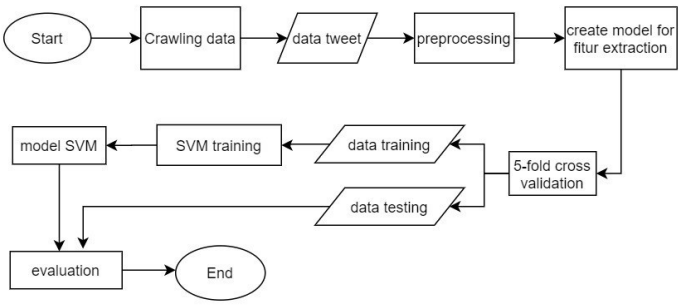

Fig. 2. Flowchart system in general

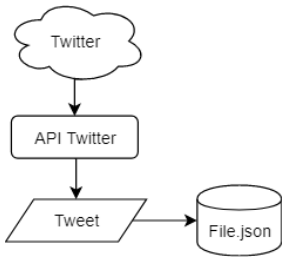

Fig. 3. Obtained data flow

TABLE II. EXAMPLES OF PREPROCESSING AND LABELING TWEETS

Tweet	Preprocessing	Label
Pentas Religi Festival Anak Soleh pada tanggal 10 Juni 2020 di Royal plaza Surabaya| 15.00-22.00 WIB | RSVP Festival Anak Soleh : 0878-5484-4045	pentas religi festival anak soleh tanggal royal plaza surabaya wib rsvp festival anak soleh	0
Free Seminar "Bagaimana meningkatkan kecerdasan buah hati kita?" | Senin, 20 Mei 2020 | kampus c unair, mulyorejo 09.00-12.00 WIB | RSVP : Ms Elly (031) 7325994 | Whatsapp 082142096381	free seminar bagaimana tingkat cerdas buah hati kita senin kampus c unair mulyorejo wib rsvp ms elly whatsapp	1
English Competitions and Seminar 2020 | Sabtu, 29 Juni 2020 | Student Center Universitas Ciputra Surabaya pukul 18.00-21.30 WIB	english competitions seminar sabtu student center universitas ciputra surabaya pukul wib	2
Pada tanggal 28 Mei 2020 ini, UNIQLO (PT Retailing) kembali membuka lowongan kerja melalui program UMC (Uniqlo Manager Candidate). Follow @custommice \/ Informasi : WA 081370443124	tanggal uniqlo pt retailing kembali buka lowongan kerja lalu program umc uniqlo manager candidate follow informasi wa	3

C. Term Frequency

Term Frequencies (TF) measures how often a word appears in a document[25]. The data is come from 822 tweets starting from 1 Jan 2020 until 23 Mar 2020. The following TF is shown in Fig. 4. By counting the number of words that appear after the preprocessing stage, we can determine approximately what events often appear in Surabaya. Based on the term frequency, we decide to divide into four classes, which are Competition, Seminars, Festivals and others for classes which is not included in the previous three classes, where the label is 0 for Festivals, 1 for Seminars, 2 for Competitions and 3 for Others. For example of tweet after preprocessing; "marvell city school fun fest mading competition marvell city surabaya info" will be classified into Competitions class.

D. Features Extraction

After going through the preprocessing stage, the next stage is taking features from each tweet. The feature is obtained by dividing or cutting each word in each tweet to get the value of GloVe per word, then averaging the results of each piece of words, into one feature value in each tweet.

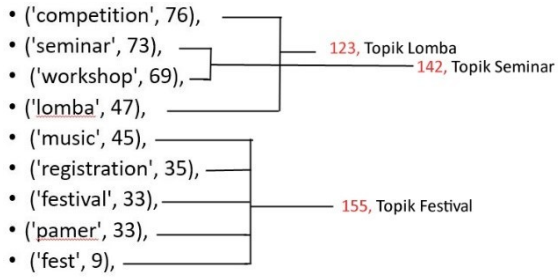

Fig. 4. Term frequency to get the topic

E. Classification with Support Vector Machine

We are using C-Support Vector Classification, Linear Support Vector Classification and Nu-Support Vector Classification. We use K = 5 for separating data. The first steps that need to be done in the classification using SVM are training, that is learning the text patterns on the tweet into a model. Prediction results can contain bias or errors. In addition, there are many kernels that can be used when classifying using SVM, including linear kernels, rbf, poly, and sigmoid. The possible accuracy generated by each kernel can be different depending on how effective the kernel is for the inputted dataset. In this twitter classification, the method that has been used is SVM by using scikit-learning as a library. There are

many types of SVM libraries with a variety of different kernels, including the SVC library, NuSVC and linear SVC. SVC and NuSVC have almost the same method, but have different sets of parameters and have different mathematical formulas. Meanwhile, the NuSVC library is a library for classifying using the SVM method, and using the regularization parameter that is Nu parameter.

IV. Result And Analysis

The experimental results for each method are as follows.

A. C-Support Vector Classification

C-support vector classification is one of the classification methods in Support Vector Machine. The implementation is based on libsvm. The time is according to the number of samples and may not practically exceed thousands of samples. the key to C-SVC is in parameter C[13]. The C parameter is in charge of controlling the tradeoff between margin and classification error which thus helps in increasing the accuracy of the output.

The greater the value of C, the greater the penalty imposed for each classification error. C values range from 0 to infinity. The modification for this case is the introduction of the parameter nu which has a range between 0 to 1 and represents the lower and upper limits of the number of examples that support vectors and is on the wrong side of the hyperplane.

B. Linear Support Vector Classification

Similar to SVC with the kernel = 'linear' parameter, but is applied in terms of liblinear rather than libsvm [13], so it has more flexibility in the choice of penalty and loss functions and must scale better for a large number of samples. The results can be seen in Table III.

TABLE III. Accuracy With C-Svc And Linear-SVC Method

Fold	Accuracy	
	C-SVC	*Linear-SVC*
1	0.815	0.790
2	0.803	0.788
3	0.777	0.720
4	0.788	0.766
5	0.777	0.788

C. Nu-Support Vector Classification

1) Kernel Performance Test on SVM Method

In this scenario, there are 600 tweets that are manually labeled. then divided into 2 parts, data train and data testing. using 5-fold cross validation. That way there are 120 test data and 480 random data train with a ratio of 2: 8. Accuracy values from 5 iterations will be averaged to get accuracy.After preprocessing, each tweet sentence is calculated using the GloVe method to get features in each sentence. Each feature will be evaluated by the SVM method 3 times with different kernels, that is Linear kernel, RBF and Poly. The results show that the highest average accuracy is in the Radial Base Function kernel with 73.26% accuracy. 72.77% for linear, 69.96% for Polynomial. the results of the accuracy of each kernel with the default nu-Support vector classification can be seen in Table IV. Why RBF kernel is better than linear and polynomial kernels is because the rbf kernel fits the twitter data feature extraction and is more flexible than linear and polynomial kernels in that you can model a whole lot more functions with its function space.

2) SVM Classification with Changes in Nu Parameters

Nu is a parameter with an upper limit for the margin of error and a lower limit for the number of support vectors. Nu values between 0 to 1. The greater the Nu value, the greater the error limit that may occur and the more vector support. In this scenario the same dataset is used as before and all tweets are classified using NuSVC 9 times by changing the value of the parameter nu ranging from 0.1 to 0.9 to see the amount of penalty due to errors in data classification. The highest NuSVC accuracy result is 85.2% at Nu value of 0.2. And the average accuracy results can be seen in Table V.

TABLE IV. Accuracy Each Kernel

Kernel	Accuracy (%)
Linear	72.774
RBF	73.259
Polynomial	69.962

TABLE V. Average Accuracy On Nu Parameter On Each Kernel

Nu Parameter	Avg Accuracy 5 cross validation		
	Linear	*RBF*	*Polynomial*
0.1	0.838	0.850	0.647
0.2	0.822	0.852	0.750
0.3	0.785	0.787	0.807
0.4	0.758	0.763	0.797
0.5	0.730	0.725	0.787
0.6	0.698	0.705	0.742
0.7	0.677	0.662	0.682
0.8	0.648	0.653	0.597
0.9	0.593	0.597	0.490

D. Confusion Matrix Analysis

In Table VI shows that there are 26 of the total 27 tweets that are correctly entered into the Festival class while there are 1 tweet from 27 tweets that are entered into another class. In the Seminar class there are 28 tweets out of 31 correct tweets, while each has 1 wrong tweet entered into the Other class. On the other hand the Competition class correctly predicted 31 tweets out of 37 tweets where 3 tweets entered the Seminar class, and 3 prediction to Other class. While the Other class successfully predicted 19 tweets from 25 tweets and 4 tweets entered the Seminar class, while each had 1 tweet at Festivals and Competitions.

In the Table VII there are 24 out of 30 correct data entered into the Festival class and 6 tweets classified into Other class. In the Seminar class there are 12 tweets out of 24 correct tweets,

while each has 2 tweets entered into the Festival, 3 tweets to the Competition and 7 tweets to the Others. On the other hand the Competition class correctly predicted 15 tweets out of 33 tweets where 5 tweets entered the Festival class, 6 Seminars and 8 Others. While the Other class successfully predicted 20 tweets from 32 tweets and 4 tweets entered the Seminar class, 6 tweets for the Festival and 2 tweets for the Competition.

TABLE VI. CONFUSION MATRIX NU 0.2 RBF HIGHEST SCENARIO

Confusion Matrix		Predicted Class			
		Festival	*Seminar*	*Competition*	*Others*
Actual Class	*Festival*	26	0	0	1
	Seminar	1	28	1	1
	Competition	0	3	31	3
	Others	1	4	1	19

TABLE VII. CONFUSION MATRIX NU 0.9 RBF LOWEST SCENARIO

Confusion Matrix		Predicted Class			
		Festival	*Seminar*	*Competition*	*Others*
Actual Class	*Festival*	24	0	3	3
	Seminar	2	12	3	7
	Competition	5	6	15	8
	Others	6	4	2	20

V. CONCLUSION

This research has produced a system that is able to classify event in Surabaya. In this study, we propose Support Vector Machine Classification to solve estimation problem. The proposed method is implemented using tweets on Twitter as a dataset. From various trials that have been carried out, there are conclusions from this study; The result shows that the best SVM method is 85.2% using Nu-Support Vector Classification using kernel RBF with parameter nu is 0.2. The second best result is Nu-Support Vector Classification with RBF kernel and 0.1 Nu with the accuracy is 85.0%. RBF and linear kernel accuracy values with Nu values above 0.3 produce accuracy below 76%, the greater the Nu value, the smaller the accuracy. Polynomial kernels are good at Nu parameter values between 0.3 to 0.5. In this study, some misclassifications occur because of the tendency of other classes to enter the other class, this problem can be minimized by increase the number of datasets that are used in the classification.

ACKNOWLEDGMENT

The Authors would like to thank the Department of Technology Management and Institut Teknologi Sepuluh Nopember, for supporting this research.

REFERENCES

[1] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, 2010, doi: 10.17148/ijarcce.2016.51274.

[2] A. Shelar and C. Y. Huang, "Sentiment analysis of twitter data," in *Proceedings - 2018 International Conference on Computational Science and Computational Intelligence, CSCI 2018*, 2018, doi: 10.1109/CSCI46756.2018.00252.

[3] Twitter Inc., "About Twitter," *Twitter About*. 2014.

[4] B. Y. Pratama and R. Sarno, "Personality classification based on Twitter text using Naive Bayes, KNN and SVM," *Proc. 2015 Int. Conf. Data Softw. Eng. ICODSE 2015*, pp. 170–174, 2016, doi: 10.1109/ICODSE.2015.7436992.

[5] P. W. Wang and C. J. Lin, "Support vector machines," in *Data Classification: Algorithms and Applications*, 2014.

[6] N. Guenther and M. Schonlau, "Support vector machines," *Stata J.*, 2016, doi: 10.4018/978-1-60960-557-5.ch007.

[7] Z. Wang and X. Xue, "Multi-class support vector machine," in *Support Vector Machines Applications*, 2013.

[8] "One-class svms for document classification," *J. Mach. Learn. Res. - JMLR*, 2002.

[9] C. Huang, L. S. Davis, and J. R. G. Townshend, "An assessment of support vector machines for land cover classification," *Int. J. Remote Sens.*, 2002, doi: 10.1080/01431160110040323.

[10] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *Annals of Statistics*. 2008, doi: 10.1214/009053607000000677.

[11] M. I. Jordan, "The Kernel Trick," *October*, 2004.

[12] L. Bertelli, T. Yu, D. Vu, and B. Gokturk, "Kernelized structural SVM learning for supervised object segmentation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011, doi: 10.1109/CVPR.2011.5995597.

[13] C. C. Chang and C. J. Lin, "LIBSVM: A Library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, 2011, doi: 10.1145/1961189.1961199.

[14] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu, "The entire regularization path for the support vector machine," *J. Mach. Learn. Res.*, 2004.

[15] and C.-J. L. Chih-Wei Hsu, Chih-Chung Chang, "A Practical Guide to Support Vector Classification," *BJU Int.*, 2008.

[16] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, 2002, doi: 10.1023/A:1012487302797.

[17] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2014, doi: 10.3115/v1/d14-1162.

[18] E. H. Huang, R. Socher, C. D. Manning, and A. Y. Ng, "Improving word representations via global context and multipleword prototypes," in *50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference*, 2012.

[19] T. Schnabel, I. Labutov, D. Mimno, and T. Joachims, "Evaluation methods for unsupervised word embeddings," in *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, 2015, doi: 10.18653/v1/d15-1036.

[20] W. Ling *et al.*, "Finding function in form: Compositional character models for open vocabulary word representation," in *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, 2015, doi: 10.18653/v1/d15-1176.

[21] M. Jupri and R. Sarno, "Taxpayer compliance classification using C4.5, SVM, KNN, Naive Bayes and MLP," *2018 Int. Conf. Inf. Commun. Technol. ICOIACT 2018*, vol. 2018-January, pp. 297–303, 2018, doi: 10.1109/ICOIACT.2018.8350710.

[22] J. Lever, M. Krzywinski, and N. Altman, "Classification evaluation," *Nat. Methods*, 2016, doi: 10.1038/nmeth.3945.

[23] E. Faisal, F. Nurifan, and R. Sarno, "Word Sense Disambiguation in Bahasa Indonesia Using SVM," *Proc. - 2018 Int. Semin. Appl. Technol. Inf. Commun. Creat. Technol. Hum. Life, iSemantic 2018*, pp. 239–243, 2018, doi: 10.1109/ISEMANTIC.2018.8549824.

[24] S. García, J. Luengo, and F. Herrera, "Data Preprocessing in Data Mining," *Intell. Syst. Ref. Libr.*, 2015.

[25] M. Fikri and R. Sarno, "A comparative study of sentiment analysis using SVM and Senti Word Net," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 13, no. 3, pp. 902–909, 2019, doi: 10.11591/ijeecs.v13.i3.pp902-909.