ASTEEC Conference Proceeding: Computer Science
3rd International Conference on Information Science and Technology Innovation (ICoSTEC)
July 27, 2024, Yogyakarta, Indonesia

Support Vector Machine With Feature Selection Chi-Square On Analysis Twitter Sentiment

Alvinur Alvinur[1], Hartono Hartono[2], Zakarias Situmorang[3]

[1,2]*Magister of Computer Science, Potensi Utama University*
JL. KL. Yos Sudarso Km. 6.5 No. 3-A, Medan, Indonesia
[3]*Departement of Computer Science, Katolik Santo Thomas University*
JL. Setia Budi, Kp. Tengah, Kec. Medan Tuntungan, Medan, Indonesia
[1]alvinurvinopandora@gmail.com
[2]hartonoibbi@gmail.com
[3]zakarias65@yahoo.com

Abstract— **The democracy party that occurred in Indonesia caused an increase in public comments on social media. Twitter is one of the most famous social media platforms in Indonesia. By utilizing a dataset of various public comments on the 2024 general election, we can do sentiment analysis. Sentiment analysis is carried out for the purpose of extracting positive or negative patterns of people's behavior in the implementation of the 2024 election. The algorithm used in analyzing sentiment is a support vector machine by substantiating chi square in the selection of dataset features. After testing 2809 data, the results of the classification accuracy of support vector machine by 73.06%, and support vector machine with chi square feature selection of 82.77% and F1-score 53.0764 against support vector machine and F1-score 70.3222 support vector machine with chi square feature selection.**

Keywords— **General Election, Sentiment Analysis, Support Vector Machine, Chi Square**

I. INTRODUCTION

Sentiment analysis is part of natural language processing and machine learning. All tweet data is classified into created categories. Sentiment analysis by analysing tweet data that includes ratings, reviews, attitudes and emotions towards entities such as products, services, organizations, people, topics, and events [1]. Of the various classification techniques used for data classification, the support vector machine method is used. The Support Vector Machine algorithm was chosen because it is suitable for separating two or more classes of data in the input space. The advantage of Support Vector Machine is that its method is simple, but provides accuracy in classifying text in a number of dimensions greater than the number of samples [2]. The development of the online social network Twitter supports the emergence of an unlimited amount of textual information, therefore it is necessary to assess the value of such information. Terms on Twitter are called Tweets and are messages or statuses created by its users. Tweets can express the Twitter user's feelings or situation as well as information [1]. The democracy party, which is an integral part of Indonesia's political culture, has sparked a surge in comments from the public on various social media platforms. One of the most widely used platforms is Twitter, which has become a digital public space for Indonesians to share their opinions and views. In the context of the 2024 general election, a lot of comments and discussions are happening on Twitter. By utilizing this dataset containing diverse comments, we can perform sentiment analysis. Sentiment analysis is a technique used to extract and understand emotional nuances from text data. It is used to determine whether a comment has positive, negative, or neutral connotations [3]. Feature selection is one of the most important factors in influencing classification accuracy, because when a data set contains many features, its spatial dimensions are large and thus reduce classification accuracy. Feature selection is the process of optimizing to reduce a large data set with large features from the original source to a relatively small subset of features that are important for improving classification accuracy quickly and effectively [4]. Common feature selection methods used in text classification are Term Frequency-Inverse Document Frequency, Information Gain, Mutual Information, Chi-square, Ambiguity Measure, Term Strength, Term Frequency-Relevance Frequency and Symbolic Feature Selection [5]. It is difficult to choose an ideal feature selection method that can be implemented and combined with all machine learning algorithms. Conduct research on the ideal feature selection method for text classification and conclude that Chi-Square provides excellent accuracy when the number of attributes or features is less than twenty [6]. The Chi-Square method is one of the text classification methods that is widely used by various researchers because it is easy to apply and verify sample data based on Chi-Square metrics measured based on True Positives, False Positives, True Negatives, False Negatives, Probability of Number of Positive Cases and Probability of Number of Negative Cases [7]. Some researchers have proposed the use of Feature Selection on SVM to limit the number of feature inputs into the classifier to achieve good accuracy and computational time [8]. The feature selection also provides parameter settings for the Support Vector Machine that significantly affect the classification accuracy results [4]. Therefore, this study will focus on categorizing tweet sentiment using Chi Square and Support Vector Machine in terms of analyzing Chi Square's accuracy in influencing feature selection on SVM to improve SVM's performance in categorizing tweet sentiment.

II. RESEARCH METHODS

This study focused on categorizing tweet sentiment using Chi Square and Support Vector Machine (SVM). Twitter sentiment data is collected, categorized, and then analysed using Chi Square to evaluate the significance of each feature to sentiment categories. The results of Chi Square analysis are used in the feature selection process on SVM to select the most relevant and informative features in tweet sentiment classification. The purpose of this study was to analyse how effective Chi Square was in improving SVM performance in categorizing tweet sentiment. The results of this study provide an understanding of the combined use of Chi Square and SVM in sentiment analysis, as well as encourage the development of more efficient and accurate methods for sentiment categorization in the context of rapidly evolving social media. Fig. 1. display the flow chart of the research methodology used for this study

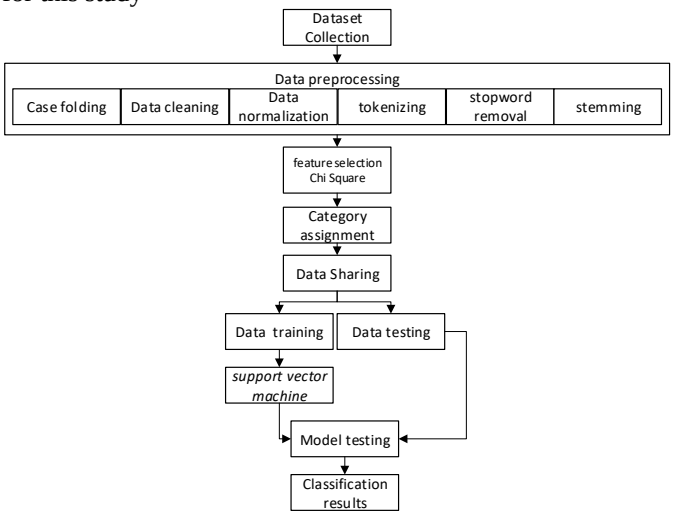

Fig. 1 Research flow chart

A. Data Collection

The first stage of the study is data collection. The data used in this study was collected using data crawling techniques to obtain information related to the 2024 general election. Data is taken from Kaggle.com website.

III. RESULTS AND DISCUSSIONS

A. Data Preprocessing

Before the classification stage is carried out using a support vector machine, the data that has been obtained needs to be organized and cleaned to get more accurate analysis results.

B. Data Cleaning

The process of data cleansing includes removing unnecessary emoticons and symbols.

C. Case Folding

At this stage the text is converted entirely to lowercase.

D. Tokenizing

At the tokenizing stage, the thing that is done is to separate sentences into stand-alone words.

E. Stopword Removal

There are 32 stopwards that are possible to delete on each word that has been separated from the sentence.

F. Stemming

Stemming is done to remove affixes. Affixes to words that have been generated from the tokenizing stage will be removed to get the base word.

G. Feature Selection

The data that has been formed in the previous stage is divided into two, namely training data and test data. The data is presented using a combination of TF-IDF and Chi Square to select further features. TF-IDF in this case is used to measure the extent to which a word is considered important in a document. This is a consideration because of the frequency with which the word appears in the document and how common the word is in the entire document. The chi square model is used to evaluate the statistical importance of each feature. The equation of the chi square formula is used in the selection of chi square features. Using TF-IDF and chi square, you can select features that have a significant relationship with the sentiment data to be predicted.

H. Modelling

The next stage is to do data modeling to get the results of sentiment analysis. The support vector machine is a model that has the function of separating data by searching the hyperplane to maximize the distance between different classes in the feature space. The support vector machine can be applied in sentiment analysis by changing the feature vector text using a feature selection model. After model creation, predictions will then continue using training data from each feature resulting from feature selection.

I. Testing

From the modeling that has been built, testing will be carried out using test data that has been formed from the dataset used.

J. Model Evaluation

The evaluation process is carried out by calculating precision, recal and F1-score for the applied model. Accuracy is the ratio between the number of correct predictions and the total number of samples.

K. Results

From the model built by substantiating feature selection and support vector machien algorithms will be evaluated using accuracy evaluation matrices, press, recall and f1-score to get performance from the model built.

Accuracy is used to measure the extent to which the classification model used is able to correctly predict sentiment on the entire dataset. The following table 1 is the accuracy generated by the built model :

TABLE I
ACCURACY RESULTS

Model	Accuracy
SMV	60.0000
SVM-CHI SQUERE	70.5556

Precision measures the extent to which the model is able to correctly identify positive sentiment from predictions made. Here are the precision values of the built model.

TABLE II
PRECISION RESULTS

Model	Precison
SMV	71.3767
SVM-CHI SQUERE	73.0338

Recal is used to determine how well the model is able to correctly find the sentiment of all sentiments in the database. The recall values on this model are :

TABLE III
RECALL RESULTS

Model	Recall
SMV	60.0000
SVM-CHI SQUERE	70.5556

In measuring harmonic mean from precision and recall f1-core provides a balanced measurement between the two matrices.

TABLE III
F1-SCORE RESULTS

Model	F1-Score
SMV	53.0764
SVM-CHI SQUERE	70.3222

Table 1 shows the accuracy value of the SVM model with the selection of Chi Squere features having an accuracy value of 70.5556 and an F1- Scrore value of 70.3222. on precision get at 73.0338 for SVM-Chi Sequere and 71.3767 for SVM.

IV. CONCLUSIONS

Evaluation matrix testing, it is known that using feature selection in the SVM model can produce higher accuracy, the accuracy number that appears reaches 70.5556. This shows that the feature selection model can perform well in sentiment analysis. In addition, the use of feature selection substance in SVM provides good results such as precision, recall and also f1-score. Therefore, this can show that models with feature selection in SVM can provide consistent and balanced results.

REFERENCES

[1] Y. Cahyono, "Analisis Sentiment pada Sosial Media Twitter Menggunakan Naïve Bayes Classifier dengan Feature Selection Particle Swarm Optimization dan Term Frequency," *J. Inform. Univ. Pamulang*, vol. 2, no. 1, hal. 14, 2017, doi: 10.32493/informatika.v2i1.1500.

[2] C. Chairunnisa, I. Ernawati, dan M. M. Santoni, "Klasifikasi Sentimen Ulasan Pengguna Aplikasi PeduliLindungi di Google Play Menggunakan Algoritma Support Vector Machine dengan Seleksi Fitur Chi-Square," *Inform. J. Ilmu Komput.*, vol. 18, no. 1, hal. 69–79, Agu 2022, doi: 10.52958/IFTK.V17I4.4594.

[3] T. D. Putra, E. Utami, dan M. P.Kurniawan, "Analisis Sentimen Pemilu 2024 Dengan Naive Bayes Berbasis Particle Swarm Optimization (Pso)," *Explore*, vol. 13, no. 1, hal. 1–5, Agu 2022, doi: 10.35200/EXPLORE.V13I1.617.

[4] M. N. Rusardi, B. Rahayudi, dan P. P. Adikara, "Analisis Sentimen Masyarakat terhadap Isu New Normal Scenario berdasarkan Opini dari Twitter menggunakan Algoritma Naive Bayes Classifier," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 3, hal. 1434–1440, Feb 2022, Diakses: 26 Juni 2023. [Daring]. Tersedia pada: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/10831

[5] B. S. Harish dan M. B. Revanasiddappa, "A Comprehensive Survey on various Feature Selection Methods to Categorize Text Documents," *Int. J. Comput. Appl.*, vol. 164, no. 8, hal. 1–7, 2017, doi: 10.5120/ijca2017913711.

[6] S. Bahassine, A. Madani, M. Al-Sarem, dan M. Kissi, "Feature selection using an improved Chi-square for Arabic text classification," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 32, no. 2, hal. 225–231, 2020, doi: 10.1016/j.jksuci.2018.05.010.

[7] I. Sumaiya Thaseen dan C. Aswani Kumar, "Intrusion detection model using fusion of chi-square feature selection and multi class SVM," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 29, no. 4, hal. 462–472, 2017, doi: 10.1016/j.jksuci.2015.12.004.

[8] Y. Niu, Y. Shang, dan Y. Tian, "Multi-view SVM Classification with Feature Selection," *Procedia Comput. Sci.*, vol. 162, hal. 405–412, 2019, doi: 10.1016/j.procs.2019.12.004.