J. Intell. Comput. Health Inform. Vol. 3, No. 2, 2022
ISSN: 2715-6923, pp.47~51 | DOI: 10.26714/jichi.v3i2.11179

Islamophobia Sentiment Classification Using Support Vector Machine

Aidil Halim Lubis*, Ali Darta, Silva Ukthi Filla

Department of Computer Science, Universitas Islam Negeri Sumatera Utara Medan, Lapangan Golf Street, Kp. Tengah, Pancur Batu Subdistrict, Deli Serdang Regency 20353, North Sumatra, Indonesia
*Corresponding author: aidilhalimlubis@uinsu.ac.id

Article histroy :
Received: 30 DES 2022
Accepted: 15 FEB 2023
Avalaible online: 25 FEB 2023

Research article

Abstract: Sentiment analysis is the process of understanding and classifying words into several categories. It is also known as opinion mining, which involves exploring opinions and emotions from text data. Sentiments can be classified into positive, negative, and neutral categories. Islam is a religion that has been in existence for centuries. Its teachings aim to foster peace and surrender to its creator, namely Allah SWT. The constructivist view of Islam has given rise to Islamophobia, which is the result of a long-standing construct that presents a negative image of Islam. Currently, Islamophobia is a growing issue that generates diverse views, especially on social media platforms. The analysis was conducted using the SVM algorithm and a dataset comprising 1000 tweets sourced from Twitter. The algorithm achieved an accuracy rate of 99.99% after testing, indicating its suitability for sentiment analysis. The error rate generated using MSE was 0.010, while the RMSE was 0.099.

Keywords: SENTIMENT ANALYSIS; ISLAMOPHOBIA; CLASSIFICATION; SUPPORT VECTOR MACHINE

1. Introduction

Islam is a religion that emerged centuries ago. This religion first appeared in the Arab region, specifically in 610 AD, which was marked by the receipt of the first revelation of the Qur'an in Mecca by Muhammad SAW. The teachings of Islam aim to build and create humans in peace and an attitude of surrender to their creator, Allah SWT. In general, Muslims must strive to save themselves and others to create peace on Earth.

There are many challenges to creating peace on earth, including the current issue of Islamophobia. According to constructivism, Islamophobia is the result of a long-standing construction that portrays Islam as cruel, vile, and inhumane. This is a frightening problem because it results in harassment and injustice towards Muslims. Furthermore, people who hate Islam often punish innocent individuals, despite there being no causal relationship between the problem and the punishment (Zulian, 2020).

In a recent case regarding the deportation of an Islamic religious figure by a country in Asia, the situation sparked debates and controversies on social media, leading to a polarization of opinions. Social media serves as a platform for quick and easy information sharing, allowing individuals to freely express themselves with the assumption that their message can be viewed and heard by everyone (Dewi and Delliana, 2020). One widely used social media platform is Twitter, which has unique features such as special symbols and rules. Twitter restricts the amount of text that can be posted, known as a tweet, to a maximum of 140 characters (Zhang et al., 2011).

Social media allows users to express their opinions and emotions in the form of positive, negative, or neutral expressions. These expressions are analyzed using an algorithm that calculates the percentage of positive, negative, and neutral tweets (Permatasari et al., 2021). The classification process in this study utilizes the Support Vector Machine (SVM) algorithm.

1.1 Text Mining

Text mining is a process that involves extracting information from a collection of documents using analytical tools. These tools can help with tasks such as processing, organizing, grouping, and analyzing large amounts of unstructured text. Text mining draws on techniques from various fields, including data mining, information retrieval, statistics, mathematics, machine learning, linguistics, natural language processing (NLP), and visualization (Permatasari et al., 2021).

The process of text mining is knowledge-intensive, as it involves users interacting with a collection of documents using analytical tools over time. Text mining is similar to data mining in that it seeks to extract useful information from data sources by identifying and exploring interesting patterns. Research in the field of text mining focuses on solving problems related to text

representation, classification, grouping, extraction, information retrieval, and pattern modeling.

There are many applications for text mining, one of which is text classification. Text classification involves grouping a document into predetermined categories or classes. The primary task of text classification is to determine the category of each document based on the unique characteristics of each class.

1.2 Machine Learning

Machine learning was first introduced in 1959 by Arthur Samuel, a computer expert from the United States, while he was still employed at IBM. Samuel defined machine learning as a system that can learn and predict data, and make decisions using various algorithms without requiring reprogramming. Machine learning is a branch of artificial intelligence that involves programming a system and teaching it to learn independently, reducing the need for repeated programming by humans. Data learning and testing are required to process data using machine learning. Learning data is used by the system to learn, while testing is used to evaluate the system's abilities. The main approaches to machine learning are supervised and unsupervised learning.

1.3 Sentiment Analysis

Sentiment analysis involves understanding and categorizing words into several classes. It is also known as opinion mining and is used to explore opinions and emotions from test data. The goal of sentiment analysis is to extract attributes and components from social media comments to determine positive, negative, and neutral classes (Permatasari et al., 2021). In (Liu, 2012) describes sentiment analysis or opinion mining as a field of natural language processing, computational linguistics, and text mining that aims to analyze a person's opinions, sentiments, evaluations, attitudes, judgments, and emotions regardless of whether the speaker or writer is pleased with a particular subject, topic, product, service, organization, individual, or activity. Sentiment analysis is very useful in processing comments, such as opinions, and turning them into something more meaningful, such as a rating. By analyzing sentiments, we can apply comments to a rating system. The trend in sentiment analysis research focuses on opinions that express positive or negative sentiments, as they can significantly influence a person's behavior. In a sentence, sentiment analysis describes the assessment of a particular entity or event (Pang, B., & Lee, 2008).

1.4 Sentiment Analysis on Twitter

According to its definition, Twitter sentiment analysis is a part of the opinions expressed on the Twitter social media platform. Messages, commonly known as tweets, are easier to analyze due to their concise writing as compared to discussions on other social media platforms. Discussion forums are more challenging to analyze since users can discuss anything and interact with one another. Usually, a sentence expresses a single opinion, although not every sentence contains a single opinion. Occasionally, a sentence may contain more than one opinion, but it's usually a small part (Liu, 2012). Sentiment analysis is generally a classification process. However, sentiment classification on Twitter (which is unstructured) is marginally more challenging than structured document classification. With Twitter sentiment analysis, which describes a sentence, the first step part is to classify whether a sentence expresses an opinion or not. The second step is to classify opinion sentences as positive or negative.

According to Barbosa in (Liu, 2012), there are several features that can be used, including retweets, hashtags, links, letter words, emoticons, exclamation marks, and question marks. There are several reasons why twitter is used as a sentiment analysis tool, like:

- Twitter is a microblogging service used by different people to express their opinions on different topics so it is a good resource for finding other people's opinions.
- Twitter contains a large number of text messages every day.
- Twitter users come from a variety of backgrounds. Therefore, it can find user messages from various social and interest groups.

1.5 Pre-Processing

Before undertaking the text mining process, it is necessary to first perform preprocessing. This involves tasks such as cleaning the data, removing duplicates, correcting typos, checking for inconsistent data, and more. The preprocessing stage is where data selection occurs, ensuring that the data to be used is structured and aligned with the existing format (Aditama et al., 2020). This preprocessing is particularly crucial in processing sentiment analysis data from social media, with the aim of achieving consistent word forms, eliminating noise, reducing the volume of words, and condensing words into a few essential tokens. There are several stages carried out in preprocessing, including:

1) Case folding refers to the process of converting all letters to lower case. This is done to standardize the word format in the text data so that it is easy to process at the next stage.
2) Filtering is the stage of removing noise, removing characters, emoticons, and punctuation marks such as "@,..." numbers, websites, url links, or so on to make the next step easier.
3) Tokenizing is the separation of a series of terms, where each row of words in a sentence, paragraph, or page is separated into tokens or single word pieces, also known as termed words.
4) Normalization is a re-examination stage to identify redundant words or letters in each token so that they are replaced according to the Big Indonesian Dictionary (KBBI).
5) Stopword Removal is the process of eliminating words that appear frequently in a document. These are words that are not so important because there are too many.
6) Stemming is the process of changing words into the basic words of all words without losing the meaning of each word. Every word that has an affix at the beginning and end of the word will be removed. For example, the word "design" becomes "design". The stemming algorithm is the same as that applied to the Nazief and Andriani algorithms, namely:

a. Look up the word in the dictionary list. If there is a word in the dictionary list, it is the root word.
b. If a word like the first step isn't in the dictionary, check the suffixes like "lah" and "kah". If there is, remove the suffix.
c. Check for possessive pronouns like "-ku", "-mu", and "nya" in the document. If there is such a word, it is removed.
d. Check the endings ("-i", "-an"), ("-lah", "-kah") and possessive pronouns ("-ku", "-mu", "-nya"). If any is found, then the suffix is omitted.
e. Check document prefixes such as "di-," "se-," "ke-," "te-," "be-," "pe-," and "me-." When found, the prefix is removed. It is possible that multi-prefix words require repeated examination of the document.
f. After all the steps have been completed and successfully, the basic words found will be returned by the algorithm.

2. Method

2.1 Support Vector Machine

The support vector machine (SVM) is a classification algorithm in the field of machine learning. Its main objective is to identify the best hyperplane that can separate two decision areas effectively. In general, SVM operates by following these steps:

1) The features of the data are mapped into a high dimensional space using a kernel function (linear, polynomial, sigmoid, or radial basis function). That is, by using kernel functions, new features will be generated that will be used to classify (no longer use old features). Each kernel is formulated as follows in Eq. (1), (2), and (3).
   a. Linear Kernel
   $$\boldsymbol{K(x_i, x_j) = x_i^T x_j} \quad (1)$$
   b. Polynomial Kernel
   $$\boldsymbol{K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0} \quad (2)$$
   c. Sigmoid Kernel
   $$\boldsymbol{K(x_i x_j) = tanh(x_i^T x_j + r)} \quad (3)$$
   d. Radial Basis Function Kernel
   $$\boldsymbol{K(x_i x_j) = exp(-\gamma||x_j^T - x_j||^2), \gamma > 0} \quad (4)$$
2) Search for the best hyperlane that separates the mapped data in a high-dimensional space. The best hyperlane can be obtained by maximizing the hyperlane distance to the nearest data point (Support Vectors), can be seen in Fig. 1.

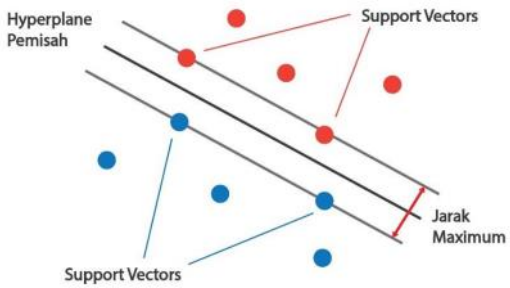

Fig 1. Hyperplane illustrate of support vector machine.

An example of using the kernel is as follows. For example, the Radial Basis Function kernel function is used with data $x = x_1, x_2, \ldots, x_n$. The features of the RBF kernel result are, for example $f = f_1, f_2, \ldots, f_m$, where these features are used to determine the classification results as follows (w is weight, f is feature).

If : $w_1f_1 + w_2f_2 + \ldots + w_mf_m > 0$, *then :class = 1, else:class = 0* $\boldsymbol{f = K(x_i x_j) = exp(-\gamma||x_j^T - x_j||^2)}$ yields the value of f. In the meantime, substitute xj with a random point l, so the equation will be temporarily changed to $\boldsymbol{f_1 = K(x_i x_j) = exp(-\gamma||x_j^T - x_j||^2)}$. By using the RBF kernel function and understanding that $\boldsymbol{||x_i^T - l||^2}$ is the distance between xi and point l, it can be concluded that when a data point xi is very close to point l, the value of f will approach 1. On the other hand, if xi is further away from point l, then the value of f will approach 0. There are parameters that can be changed to produce the best separating hyperlane, including parameter C and parameter. The regulation parameter C is used by SVM to find the best distance from the hyperlane. The best value of C must be sought so that SVM strikes a balance between the maximum distance and the minimum possible misclassification. A high C value means that the model will receive more penalties when it fails to classify the data properly (Pisner and Schnyer, 2019). On the other hand, a low C value means that the model will tolerate misclassified data. A low C value is usually good for data that has a lot of noise, whereas a high C value is usually good for data that has little noise.

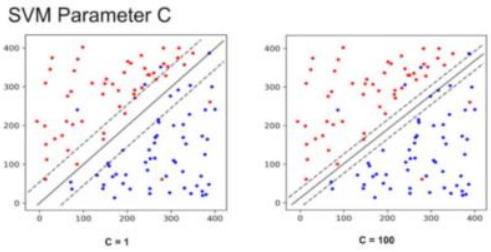

Fig 2. Ilustrate effect C parameter on support vector machine.

3. Result and Discussion

In this study, we will implement a support vector machine (SVM) algorithm to analyze a dataset of tweets sourced from the Twitter social media application. The dataset focuses on the issue of "Islamophobia" and was obtained in June 2022. The tweets were collected from Indonesia.

1) Pre-Processing.

At this stage, the dataset of tweets that has been collected will undergo text preprocessing. In this section, the entire dataset will be preprocessed by converting all words to lowercase. Then, the next stage of data cleaning will involve removing punctuation marks, URLs, and hashtags. The dataset comprises of 1000 tweets. Once the preprocessing is complete, the data will be ready for further analysis. The preprocessed dataset is presented in Table 1.

Table 1. Results of pre-processing stage.

INITIAL TEXT	PRE-PROCESSING RESULT	LABEL
#Islamophobiahttps://t.co/RTIFDvTd5W	Islamophobia	Netral
Mayjen (Purn) Deddy Budiman: Era Oligarki Neo-Komunisme Propaganda Islamophobia Terstruktur, Sistematis, dan Masif! https://t.co/U2PlaJAuEX	Mayjen purn deddy budiman era oligarki neo komunisme propaganda islamophobia terstruktur sistematis dan masif	Netral
@Dennysiregar7 Proporsional dalam menilai dong den. Islamophobia amat makhluk pemakan bangkai satu ini.	proporsional dalam menilai dong den. islamophobia amat makhluk pemakan bangkai satu ini	Netral
@OposisiCerdas ISLAMOPHOBIA LAGI NGARANG BEBAS...😂😂😂	oposisicerdas islamophobia lagi ngarang bebas	Netral
@geloraco Kok Kemensos yg selewengkan dana Bansos tidak dibubarkan, Islamophobia ❓	geloraco kok kemensos yg selewengkan dana bansos tidak dibubarkan islamophobia	Netral
Islamophobia https://t.co/tTg0a9aOka	Islamophobia	Netral
yg urgent itu UU ANTI ISLAMOPHOBIA & ANTI KORUPSI	yg urgent itu uu anti islamophobia anti korupsi	Negatif
@yaniarsim Dikit2 islamophobia, padahal sih cuma ini orang aja sebenernya males nyari fakta	yaniarsim dikit islamophobia padahal sih cuma ini orang aja sebenarnya males nyari fakta	Netral
@HisyamMochtar Mereka adalah Islamophobia group	hisyammochtar mereka adalah islamophobia group	Netral
islamophobia suka cari gara2. Ntar kalo dibunuh, Islam dibilang teroris.	Islamophobia suka cari gara ntar kalo dibunuh islam dibilang teroris caper banget sih insecure kah sama agamanya sendiri	Negatif

2) TF-IDF Word Weighting

The TF-IDF method assigns weights to each token in a document. The weight value is obtained by multiplying the frequency of occurrence of tokens in a document (term frequency) and the frequency of tokens in all documents (inverse document frequency). The process section generated by TF-IDF is shown in Figure 3.

3) Cross Validation

This is the most commonly used method for evaluating a model's predictive performance. The model generalizes independent data through the process of separating training and testing data using the K-Fold function. In this research, the value of K used is 5.

Word	TF-IDF
singapura	0.347
islamophobia	0.112
yg	0.099
@uas_abdulso...	0.071
konser	0.065
singapura,	0.057
indonesia	0.054
2022	0.051

Fig 3. TF-IDF weighting results.

4) Evaluation Results and Predictions

The following results were obtained from the evaluation using the SVM algorithm can be seen in Table 2.

Table 2. Model evaluation result.

MODEL	MSE	RMSE	MAE
SVM	0.010	0.100	0.098

In Table 3, displays the results of the model evaluation, which achieved an evaluation accuracy of 99.99%. Additionally, the accuracy of the model's predictions using the SVM algorithm was also around 99.99%. These results indicate that the model has performed well during testing.

Table 3. Model prediction result.

MODEL	MSE	RMSE	MAE
SVM	0.01	0.099	0.098

The results of the visualization of the distribution of datasets sourced from tweets that have been evaluated using the SVM algorithm can be shown in Fig. 4.

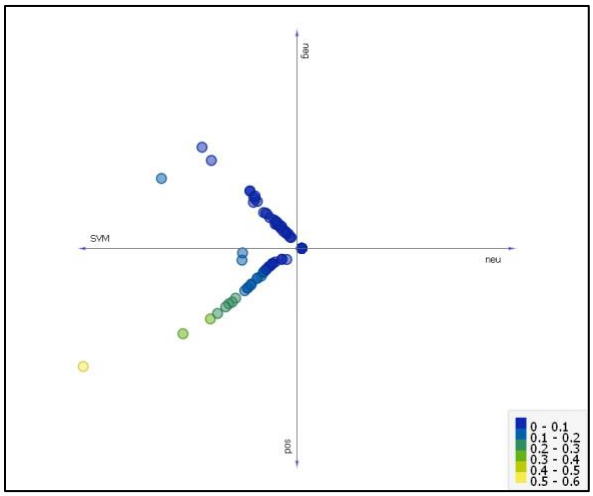

Fig 4. Data distribution visualization.

4. Conclusion

In this study, we have presented a classification of islamophobia sentiment using support vector machines. Based on the research and discussion that have been done, we have determined that the performance accuracy is

99.99%. According to the accuracy result, the SVM algorithm works optimally in analyzing the tweet dataset and is able to classify into several sentiments, including positive sentiment, negative sentiment, and neutral sentiment. It can be proven based on the mean square error (MSE) value of 0.010 and the root mean square error (RMSE) value of 0.099. The mean absolute error (MAE) value is 0.098. It can be concluded that the SVM algorithm can be used to measure people's response to issues that develop through social media such as Twitter.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Aditama, M.I., Pratama, R.I., Wiwaha, K.H., Rakhmawati, N.A., 2020. Analisis Klasifikasi Sentimen Pengguna Media Sosial Twitter Terhadap Pengadaan Vaksin COVID-19. Journal of Information Engineering and Educational Technology, 4(2), 90-92. https://doi.org/10.26740/jieet.v4n2.p90-92

Dewi, A.P., Delliana, S., 2020. Self Disclosure Generasi Z Di Twitter. Ekspresi Dan Persepsi J. Ilmu Komun. 3, 62. https://doi.org/10.33822/jep.v3i1.1526

Liu, B., 2012. Sentiment Analysis and Subjectivity. Handbook of natural language processing, 2(2010), 627-666.

Pang, B., & Lee, L., 2008. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval. 2(1–2), 1-135. http://dx.doi.org/10.1561/1500000011

Permatasari, P.A., Linawati, L., Jasa, L., 2021. Survei Tentang Analisis Sentimen Pada Media Sosial. Maj. Ilm. Teknol. Elektro 20, 177. https://doi.org/10.24843/mite.2021.v20i02.p01

Pisner, D.A., Schnyer, D.M., 2019. Support vector machine, in: Machine Learning: Methods and Applications to Brain Disorders. https://doi.org/10.1016/B978-0-12-815739-8.00006-7

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., Liu, B., 2011. Combining lexicon-based and learning-based methods for twitter sentiment analysis. HP Lab. Tech. Rep.

Zulian, I., 2020. Analisis Pengaruh Islamophobia Terhadap Kebijakan Luar Negeri Amerika Serikat Di Pemerintahan Donald Trump. J. PIR Power Int. Relations 3, 140. https://doi.org/10.22303/pir.3.2.2019.140-155