Journal of Soft Computing Exploration

Comparison of LSTM, SVM, and naive bayes for classifying sexual harassment tweets

Tiara Lailatul Nikmah[1], Muhammad Zhafran Ammar[2], Yusuf Ridwan Allatif[3], Rizki Mahjati Prie Husna[4], Putu Ayu Kurniasari [5], Andi Syamsul Bahri[6]
[1,2,3,4,5]Department of Computer Science, Universitas Negeri Semarang, Indonesia
[6]PPWNI Klang, Malaysia

Article Info

Article history:

Received Sep 15, 2022
Revised Sep 24, 2022
Accepted Sep 29, 2022

Keywords:

Tweet classification
Sentiment analysis
Sexual harassment tweets

ABSTRACT

Twitter is now a very open and extensive social media; anyone can freely express their opinion on any topic on social media. The content or discussion on Twitter is also quite diverse and unlimited. However, because it is unlimited, many misuse it for negative things. One of them is verbal sexual harassment through Twitter. This research aims to identify sexual harassment in an Indonesian tweet using sentiment analysis using the LSTM, SVM, and naive bayes methods with text normalization. In this study, 2990 tweets in the Indonesian language were tested from 4th to 6th in May 2022. The Twitter data shows that tweets included in sexual harassment are more than those not included in sexual harassment, totaling 2026 data. From the results of the evaluation of tweet data classification using text normalization with LSTM, the accuracy is 84.62%, SVM is 86.54%, and naive bayes is 85.45%. Using the SVM algorithm with text normalization gets the highest accuracy compared to LSTM and naive bayes in classifying Indonesian sexual harassment tweets.

Corresponding Author:

Tiara Lailatul Nikmah,
Department of Computer Science,
Universitas Negeri Semarang,
Sekaran, Gunungpati, Semarang 50229, Indonesia.
Email: tiaralaila21@gmail.com

1. INTRODUCTION

Today's technological advances place social media as a space for sharing information and communication. As a communication space, social media has reached all ages of society, both young and old. According to Carr and Hayes, social media is internet media that makes it easy for users to introduce themselves and interact with others, directly or indirectly, thereby encouraging the values and perceptions of others [1]. This has indicated that social media has become an essential part of today's society.

Social media's unrelenting expansion, particularly the information posted to social media networks, can impact community members' ability to communicate with one another. It can reflect how individuals communicate if it is conducted virtually rather than in person. Users express especially concern over this given that social media's openness and flexibility are expanding in variety and boundlessness. According to Zhong, Kebbell, and Webster [2], the accessibility of internet access has altered the structure of interpersonal interactions that influence the norms in acting and interacting. Under the guise of freedom of speech, conversation, and opinion on social media, this has led to the susceptibility to social media abuse.

Twitter is one of the social media platforms that has been abused. Verbal sexual harassment is the form of abuse used in the application. This can take the form of telling crude jokes, using sexually explicit language, making direct comments about the victim's organs, and so on. The prevalence of sexual harassment on social media will increase users' anxiety and insecurity when using the platform. Therefore, it is vital to establish a good and effective system to overcome this.

The purpose of this research was to develop a better and more efficient way to identify sexually harassing and deviant content. Text mining is a technique that can be used to do this. Text mining converts unstructured data words and phrases into numerical values, which can then be linked with structured data in a database and analyzed using traditional data mining techniques [3]. The distinctive patterns of text written in natural language are identified using this technique.

Support Vector Machine (SVM), Long Short-Term Memory (LSTM), and naive bayes are some frequently employed techniques in text mining. Pal, Ghosh, and Nag reported the relative outcomes of sentiment analysis research employing three types of LSTM, namely conventional LSTM with a validation accuracy of about 80.92%, deep LSTM with a validation accuracy of around 81.32%, and bidirectional deep LSTM with validation accuracy of around 83.83% [4]. This demonstrates that the LSTM technique has the potential to reach a high degree of accuracy. According to a study by Rahat, Kahir, and Masum, the SVM approach has an accuracy rate of about 83%, which is greater than the accuracy rate of the naive bayes algorithm, about 77% [5]. The SVM approach outperforms the naive bayes algorithm in the research of Le and Nguyen while evaluating 200,000 tweets from Stanford University because it thoroughly analyzes tweets containing public opinion, which may contain several words in them [6]. The naive bayes algorithm is a statistical categorization that can be used to estimate the likelihood that a class would contain specific members, in contrast to the SVM approach [7].

This study classified tweet data using three techniques: naive bayes, SVM, and LSTM. This is accomplished by evaluating the outcomes of classification computations made using various approaches in terms of accuracy, recall, and precision. Each method's computation results will be compared to determine which is best at analyzing and categorizing tweets to determine whether they contain aspects of sexual harassment. The findings of this study are anticipated to serve as a source for more in-depth learning about sentiment analysis, the foundation for choosing sentiment analysis approaches, or the basis for improving sentiment classification and analysis techniques. Furthermore, this article will likely help social media platform developers optimize sexual harassment detecting technologies.

2. METHOD

In this case, the sentiments contained on Twitter are further classified into sexual harassment or non-sexual harassment entities [8]. The steps taken in the study are described as shown in Figure 1.

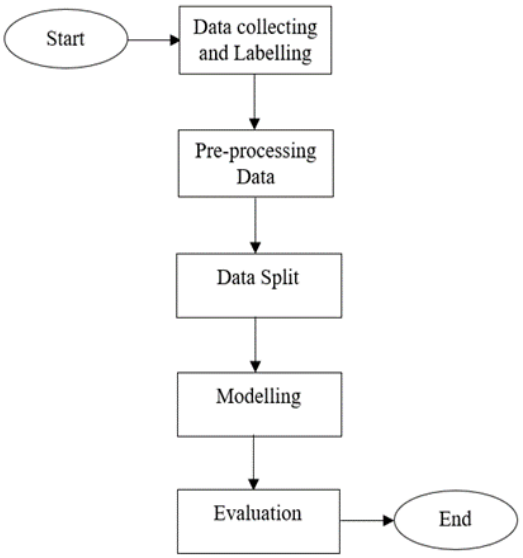

Figure 1. Diagram of research stages

2.1 Data Collection and Labeling

The technique scraping on Twitter was used to collect samples. They were scraping data using one of the libraries python scrapes. The data is taken based on the given keywords. In this case, the keywords given are words related to sexual harassment. There are 11 keywords entered. The tweet data obtained were 2990 data taken from tweets on 4 - 6 May 2022. At the data labeling stage, the researcher manually labeled whether a tweet was classified as sexual harassment with label 1 or not sexual harassment with label 0.

2.2 Pre-Processing Data

The preprocessing is carried out to make the data included in the model suitable for use. Clean data will produce good results, such as a high accuracy value. The preprocessing data stage consists of case folding to clean text data from symbols, numbers, URLs and mentions, text normalization, namely changing non-standard words or abbreviations into proper words, removing stopwords, namely removing common words such as conjunctions because they are considered unimportant, stemming i.e. removing prefixes and suffixes on words so that words are more straightforward so that the model is easier to process. This stage uses pandas, numpy, nltk, and academic libraries.

2.2.1 Text Normalization

Text normalization is changing irregular language or abbreviations into words with the correct spelling. It is an essential step in text preprocessing to reduce the noise produced by a single word with multiple forms. A corpus with a list of appropriate and inappropriate terms is used in this instance to normalize the use of words. The terms in the dataset are generally matched by the corpus, which was acquired from github.

2.2.2 Tokenization

Tokenization is breaking down a continuous stream of text into sentences and words. In essence, it is the task of dividing a text into tokens. This stage breaks the text into chunks of words and is given a number index [9]. These numbers become a representation of the word in the text. The tokenization used in this paper is the Term Frequency and Inverse Document Frequency (TF-IDF) method. TF-IDF is a numerical statistic that shows the relevance of keywords to specific documents, or it can be said that it provides those keywords that can be used to identify or categorize specific documents [10]. The Term Frequency (TF) counts the number of times a term appears in a document [11]. IDF is the frequency of documents that gives less weight to words that appear frequently and more weight to words that appear rarely in a document [10].

2.3 Data Split

The dataset is divided into training data and testing. Training data is used in the training or model learning. The model will study the training data and its labels so that it can classify the new data. At the same time, the testing is used for evaluating model performance. Data split is done with the help library of the sci-kit-learn.

2.4 Modelling

After the data preprocessing stage, such as the data grouping process, converting the data into the numeric form, and dividing the data between training data and test data [12], models are trained using training data. The model was then tested using data testing. The three algorithms used in this paper are LSTM, Naïve Bayes, and SVM.

The first algorithm is LSTM. Deep learning techniques with the LSTM algorithm are used in Twitter sentiment analysis. Sentiment analysis a view, behavior, or opinion into an entity [8]. LSTM is a particular type of RNN that has internal memory and a multiplication gate [13]. [14] LSTM uses three gates: an input gate that regulates the flow of input activations into the memory cell, an output gate that regulates the flow of cell activations into the rest of the network, and a forget gate that is used to forget or reset the cell's memory [15]. In LSTM, a single gate component is used to control the information entered memory which is tasked with solving the problem of gradient loss and division. Repeated connections add state or memory to the network, allowing it to take advantage of ordered observations. Internal memory means that the network output depends on the last context in the input queue instead of the input presented as a network [16].

The model consists of three layers: embedding, LSTM, and dense. Embedding layers are used to convert words into vectors. The LSTM layer uses 100 units to find relationships between words with labels. A dense layer with sigmoid activation function will give output in the form of probability. The model was created using the keras tensorflow the model consists of three layers: embedding, LSTM, and dense. The following is

a summary of the model created. At this stage, for the LSTM method, the number of epochs is 20, and the batch size is 64. The early stopping technique is used to minimize overfitting.

The next algorithm is naive bayes. The algorithm has been used several times for sentiment analysis for information retrieval systems for determining trend titles of Indonesian-language journals [17], classifying fake news on Twitter [18], analysis the public's perception of the new normal campaign through Twitter [19] and conducting sentiment analysis of review hotels [7]. Naïve bayes is also used by [20] to detect name spam on LinkedIn. The resulting naive bayes algorithm is derived by calculating the events in the data into the sample to make decisions on the problems at hand [21]. The resulting naive bayes algorithm is derived by calculating the events that occur in the data into the sample in order to make decisions on the problems at hand [22]. The naive bayes classifier algorithm minimizes variation within an attribute [23]. The naive bayes method will compute the likelihood of each case. The target attribute's value in each data sample. The naive bayes classifier will then classify the sample data into the class with the highest probability value [17]. At this stage, the naive bayes method used the sklearn model modules.

The third algorithm used in this paper is SVM. SVM proposed by Cortes and Vapnik in 1995, is one of the most successful classifications of learning techniques, where the SVM is enhanced and adapted to different application areas [24]. SVM is a supervised learning data mining model in which data is classified and analyzed linearly [25]. SVM is a method that finds the best hyperplane for separating two classes in input space using the structural risk minimization principle [26] [14]. SVM works by locating the hyperplane with the most significant margin between two classes. SVM is perfect for large data sets with many dimensions [27]. At this stage, the SVM method used the sklearn model modules.

2.5 Evaluation

Evaluation is the step to measure the performance of each of our proposed methods. In this step, we use the confusion matrix. The confusion matrix can present predictions and the actual state of the data generated by machine learning algorithms. By using the confusion matrix, accuracy, precision, and recall can be determined.

2.5.1. Accuracy

Accuracy is a representation of the ratio of correct predictions, True Positive (TP) and True Negative (TN), based on the overall data [26]. Equation (1) is a formula for calculating accuracy.

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \quad (1)$$

2.5.2. Precision

Precision represents the ratio of true positive predictions to the overall predicted positive results, True Positive (TP) and False Positive (FP). Equation (2) is a formula for calculating precision.

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

2.5.3. Recall

Recall or sensitivity is the representation of the ratio of True Positive (TP) predictions to the overall data that True Positive (TP), namely True Positive (TP), and False Negative (FN). Equation (3) is a formula for calculating recall.

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad (3)$$

3. RESULTS AND DISCUSSIONS

The samples of 3300 raw data in total were collected by scraping. The search for tweets is based on a set of keywords. Eleven keywords are utilized. The keywords used are frequently used in instances of sexual harassment. About 300 tweets are contained in each term. Then, the researcher did the data labeling manually with the help of microsoft excel software. Data cleaning removes tweets that are not in Indonesian after data cleaning. The remaining data is reduced to 2990 in total. Table 1 shows the details of the remaining data.

Table 1. Data total

Label	Amount of Data
1	2026
0	964

The data preprocessing stage is completed when the data from tweets in other languages have been cleaned. The data must go through various rounds of preprocessing. Cleaning up symbols, numbers, URLs, and mentions in a tweet comes first. A python code that uses Regex to remove symbols, numbers, URLs, and Twitter mentions is created to complete this stage. Regex is used in conjunction with data cleaning techniques to quickly identify unnecessary characters and remove them from the main data to raise the dataset quality [28]. Text normalization with a public corpus downloaded from GitHub is the second stage. The corpus includes slang or abbreviated Indonesian words. Slang terms and misspelled words will currently be replaced in tweets with the correct spelling. The next step is to remove stopwords with the help of NLTK libraries. After the preprocessing stage, the text data is reduced to words without numbers, symbols, emojis, URLs, or mentions. The result after the text normalization stage is shown in Table 2.

Table 2. Clean text

Tweet	Clean_text
@alter18base Bersihin kek kramiknya ajg malah foto kontol.	bersihin kek kramiknya ajg foto kontol
@FOODFESS2 la lonte	la lonte
Goblok https://t.co/E7BQCjqBRJ	goblok
……	…….
@tarochocochips @Unchonz Grepe grepe club	grepe grepe club
@Cintada16 Wan jembut cuma gunting pita...	wan jembut gunting pita

The normalized text enters the tokenization stage, breaking the text into words and giving a number index. This paper uses the TF-IDF method. The words will be converted into numbers representing the text's words. The greater or higher the occurrence of a word in documents (TF), the higher the term frequency, and the lesser the occurrence of a word in documents, the higher the importance (IDF) for that keyword search in that specific document [10]. Then the dataset is split into training and testing data in an 80:20 ratio, with a more significant amount of training data overall, as shown in Table 3.

Table 3. Training-testing data division

Data	Amount of Data
Training	2392
Testing	598

LSTM, SVM, and naive bayes were all used in the model training process. Twenty epochs and 64 data batches were used for the LSTM model training. The model came to an end in 8 epochs under the influence of the early stopping strategy. However, the model used with SVM, and naive bayes is the sckit-learn modul model with default hyperparameters. The results of the model evaluation using accuracy, recall, and precision for the LSTM, SVM, and naive bayes methods are shown in Table 4.

Table 4. Model evaluation result

Algorithms	Accuracy	Precision	Recall
LSTM	84.62%	85.61%	73.77%
SVM	86.54%	83.58%	76.01%
Naïve Bayes	85.45%	88.05%	73.75%

Based on the evaluation results shown in Table 4, it can be concluded that SVM produced more accurate predictions than naive bayes and LSTM in classifying tweets regarding sexual harassment written in the Indonesian language, with a model accuracy of 86.54%.

4. CONCLUSION

The classification model uses 2990 datasets taken from Twitter through scrapping based on 11 keywords. Classification using LSTM, SVM, and naive bayes methods show good results. The evaluation results of tweet data classification using LSTM, SVM, and naive bayes with text normalization, show the accuracy of LSTM 84.62%, SVM 86.54% and naive bayes 85.45%. Thus, the SVM algorithm performs better than LSTM and naive bayes in classifying sexual harassment tweets.

REFERENCES

[1] C. Carr and R. Hayes, "Social Media: Defining, Developing, and Divining," *Atl. J. Commun.*, vol. 23, pp. 46–65, 2015.

[2] L. R. Zhong, M. R. Kebbell, and J. L. Webster, "An exploratory study of technology-facilitated Sexual Violence in online romantic interactions: Can the Internet's toxic disinhibition exacerbate sexual aggression?," *Comput. Hum. Behav.*, vol. 108, p. 106314, 2020.

[3] R. Sagayam, S. Srinivasan, and S. Roshni, "A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques," *Int. J. Comput. Eng. Res.*, vol. 2, no. 5, pp. 2250–3005, 2012.

[4] S. Pal, S. Ghosh, and A. Nag, "Sentiment Analysis in the Light of LSTM Recurrent Neural Networks," *Int. J. Synth. Emot.*, vol. 9, pp. 33–39, 2018.

[5] A. M. Rahat, A. Kahir, and A. K. M. Masum, "Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset," in *8th Int. Conf. Syst. Model. Adv. Res. Trends*, 2019, pp. 266–270.

[6] B. Le and H. Nguyen, "Twitter Sentiment Analysis Using Machine Learning Techniques," in *Adv. Comput. Methods Knowl. Eng.*, 2015, pp. 279–289.

[7] I. E. Tiffani, "Optimization of Naïve Bayes Classifier By Implemented Unigram, Bigram, Trigram for Sentiment Analysis of Hotel Review," *J. Soft Comput. Explor.*, vol. 1, no. 1, pp. 1–7, 2020.

[8] A. Yadav and D. Kumar, "Sentiment analysis using deep learning architectures : a review," *Artif. Intell. Rev.*, vol. 53, no. 6, pp. 4335–4385, 2020.

[9] R. Adelia, S. Suyanto, and U. N. Wisesty, "Indonesian Indonesian Abstractive Abstractive Text Text Summarization Summarization Using Using Bidirectional Bidirectional Gated Recurrent Unit Gated Recurrent Unit," *Procedia Comput. Sci.*, vol. 157, pp. 581–588, 2019.

[10] S. Qaiser and R. Ali, "Text Mining : Use of TF-IDF to Examine the Relevance of Words to Documents Text Mining : Use of TF-IDF to Examine the Relevance of Words to Documents," *Int. J. Comput. Appl.*, vol. 181, no. 1, pp. 0975 – 8887, 2018.

[11] A. A. Hakim, A. Erwin, K. I. Eng, M. Galinium, and W. Muliady, "Automated Document Classification for News Article in Bahasa Indonesia based on Term Frequency Inverse Document Frequency (TF-IDF) Approach," in *6th Int. Conf. Inf. Technol. Electr. Eng.*, 2014, pp. 0–3.

[12] T. U. Haque, N. N. Saber, and F. M. Shah, "Sentiment analysis on large scale Amazon product reviews," in *IEEE Int. Conf. Innov. Res. Dev.*, 2018, pp. 1–6.

[13] K. Smagulova and A. P. James, "A survey on LSTM memristive neural network architectures and applications," *Eur. Phys. J. Spec. Top.*, vol. 228, no. 10, pp. 2313–2324, 2019.

[14] T. Saini, G. Tomar, D. Rana, S. Attri, P. Chaturvedi, and V. Dutt, "CloudIoT for pollution monitoring: A multivariate weighted ensemble forecasting approach for prediction of suspended particulate matter," in *CloudIoT: Concepts Paradig. Appl.*, CRC Press, 2020.

[15] A. Pulver and S. Lyu, "LSTM with working memory," in *Int. Jt. Conf. Neural Netw.*, 2017, pp. 845–851.

[16] I. Augenstein, T. Rocktäschel, A. Vlachos, and K. Bontcheva, "Stance Detection with Bidirectional Conditional Encoding," in *Proc. 2016 Conf. Empir. Methods Nat. Lang. Process.*, Nov. 2016, pp. 876–885.

[17] W. B. Trihanto, R. Arifudin, and M. A. Muslim, "Information Retrieval System for Determining The Title of Journal Trends in Indonesian Language Using TF-IDF and Naive Bayes Classifier," *Sci. J. Inform.*, vol. 4, no. 2, pp. 179–190, 2017.

[18] H. A. Santoso, E. H. Rachmawanto, and U. Hidayati, "Fake Twitter Account Classification of Fake News Spreading Using Naïve Bayes," *Sci. J. Inform.*, vol. 7, no. 2, pp. 228–237, 2020.

[19] R. L. Mustofa and B. Prasetiyo, "Sentiment analysis using lexicon-based method with naive bayes classifier algorithm on #newnormal hashtag in twitter," in *J. Phys.: Conf. Ser.*, 2021, vol. 1918, no. 4.

[20] D. M. Freeman, "Using Naive Bayes to detect spammy names in social networks," *Proc. ACM Conf. Comput. Commun. Secur.*, pp. 3–12, 2013.
[21] Walid and Alamsyah, "Naïve Bayesian classifier algorithm and neural network time series for identification of lecturer publications in realizing internationalization of Universitas Negeri Semarang," in *J. Phys.: Conf. Ser.*, 2019, vol. 1321, no. 3.
[22] L. Marlina, M. Muslim, and A. P. U. Siahaan, "Data Mining Classification Comparison (Naïve Bayes and C4.5 Algorithms)," *Int. J. Emerg. Trends Technol. Comput. Sci.*, vol. 38, pp. 380–383, 2016.
[23] Y. F. Safri, R. Arifudin, and M. A. Muslim, "K-Nearest Neighbor and Naive Bayes Classifier Algorithm in Determining The Classification of Healthy Card Indonesia Giving to The Poor," *Sci. J. Inform.*, vol. 5, no. 1, p. 18, 2018.
[24] E. Tuba and Z. Stanimirovic, "Elephant herding optimization algorithm for support vector machine parameters tuning," in *9th Int. Conf. Electron. Comput. Artif. Intell.*, Jun. 2017, pp. 1–4.
[25] S. Tyagi and S. Mittal, "Sampling Approaches for Imbalanced Data Classification Problem in Machine Learning," in *Proc. ICRIC*, 2020, pp. 209–221.
[26] Sulistiana and M. A. Muslim, "Support Vector Machine (SVM) Optimization Using Grid Search and Unigram to Improve E-Commerce Review Accuracy," *J. Soft Comput. Explor.*, vol. 1, no. 1, pp. 8–15, 2020.
[27] M. Sam'an and Y. N. Ifriza, "Performance comparison of support vector machine and gaussian naive bayes classifier for youtube spam comment detection," *J. Soft Comput. Explor.*, vol. 2, no. 2, pp. 93–98, 2021.
[28] T. Mustaqim, K. Umam, and M. A. Muslim, "Twitter text mining for sentiment analysis on government's response to forest fires with vader lexicon polarity detection and k-nearest neighbor algorithm," in *J. Phys.: Conf. Ser.*, 2020, vol. 1567, no. 3.