EKSAKTA
Journal of Sciences and Data Analysis

E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI : 10.20885/EKSAKTA.vol5.iss1.art7

Research Article

Sentiment Analysis and Topic Modelling of Bjorka Using Support Vector Machine and Latent Dirichlet Allocation

Muhammad Muhajir [1,2], **Dedi Rosadi** [1,*]

[1] Department of Mathematics, Universitas Gadjah Mada, Bulaksumur, Caturtunggal, Depok, Sleman Regency, Special Region of Yogyakarta 55281
[2] Department of Statistics, Universitas Islam Indonesia, Jln. Kaliurang KM 14.5, Sleman, Yogyakarta, Indonesia

[*] *Corresponding author: dedirosadi@ugm.ac.id*

Received: 30 October 2023; Accepted:14 Maret 2024; Published: 26 April 2024 (57-66)

Abstract: A wide range of data is now easily accessible via the microblogging service Twitter thanks to the rapid advancement of technology. The Bjorka controversy, one of the most talked-about topics right now, has generated numerous comments from the general public and thus has risen to the top. The Bjorka phenomenon is an obvious example of cybercrime, with a sharp uptick in incidents occurring in Indonesia during the COVID-19 pandemic. Sentiment analysis employing the Support Vector Machine technique allows for the statistical analysis of public opinion about Bjorka as it appears on the Twitter social network. Latent Dirichlet Allocation (LDA) will be used to analyze the sentiment analysis with SVM results, which have been separated into positive and negative sentiments. In this study, using LDA for sentiment analysis resulted in an accuracy of 89.5%. Dismantling government data, including personal data and government crimes, was the most positively predicted topic, with 75.2% of all predictions leaning in that direction. It is hoped that the government will be able to use the information gleaned from this study to better understand the public's perspective and the trust deficits that need to be addressed

Keywords: Bjorka, Cybercrime, Support Vector Machine, Latent Dirichlet Allocation (LDA), Sentiment analysis

Introduction

The rapid development of technology makes various information easily accessible through social networks. The variety of development topics is quite interesting to attract the attention of all strata of society and fulfill the interaction of digital space. Twitter is a social media that is quite popular in facilitating users to write and publish their activities and opinions. Through Twitter, users can spread information, promote the opinions or views of other users, discuss trending topic issues, and become part of the issue [1].

Recently, social networking in Indonesia has been shocked by a Twitter account with the username Bjorka, which claims the leak of 1.3 billion Indonesian citizens' SIM Card registration data and 26 million IndiHome user data on dark forum [2]. Moreover, Bjorka uploaded other controversial matters and involved some important figures by uploading documents that claimed belonged Bjorka also uploaded them to President Jokowi during the 2019-2021 period.

The leaked data includes personal information of Indonesian government officials including Johny G Plate from the Ministry of Communication and Information, Mahfud MD from the Coordinating Ministry for Political, Legal and Security Affairs, and Anies Baswedan, Governor of DKI Jakarta. Bjorka has increasingly drawn public attention by exposing the murder case of human rights activist Munir and mentioning that one of the murder masterminds was Muchdi Purwoprandjono currently serves as Chairman of the Berkarya Party [3].

Media analyst Ismail Fahmi said the Bjorka issue that able to grab much attention and occupied the first position involving many public responses [4]. Some people considered Bjorka's controversial action as a form of demonstration or protest by utilizing technological development. However, the statement was denied by a digital communication expert from the Faculty of Social and Political Science (FISIP) Airlangga University Prof. Dra. Rachmah Ida, M.Com., Ph.D. who said the Bjorka phenomenon is not a form of modern protest or demonstration but is a blatant form of digital crime or cybercrime [5].

Cybercrime in Indonesia has increased quite rapidly during the Covid-19 pandemic. The House of Representatives of the Republic of Indonesia (DPR) Akhmad Muqowam revealed that the level of cybercrime in Indonesia ranks second in the world after Ukraine [6]. The issue is a significant concern for the Indonesian government to follow up on violation cases in the digital world, especially with the appearance of Bjorka, who openly committed cybercrime and caused various public reactions regarding the motives. The Bjorka phenomenon is an interesting topic because a few people unintentionally support and give a positive response to Bjorka who committed a digital crime.

Public opinion about Bjorka's appearance on the social network Twitter can be analyzed using a statistical method, namely sentiment analysis. Sentiment analysis is the process of classifying information by extracting and classifying opinions according to the polarity of data, namely positive, negative, and neutral [7]. The sourced data is from Twitter by utilizing the keywords ”Bjorkanism” and ”Hacker Bjorka”. The applying process of sentiment analysis is used to find out whether the public's response to Bjork tends to be positive or negative.

One of the methods of classification text that has the best performance and can handle infinite dimensions is the Support Vector Machine [8]. This is proven by previous studies such as A Comparative Study of Support Vector Machine and Na¨ıve Bayes Classifier for Sentiment Analysis on Amazon Product Reviews which concluded that SVM has a higher accuracy value compared to Na¨ıve Bayes for polarizing Amazon product reviews [9]. In addition, another study on Sentiment Polarity Detection in Bengali Tweets Using Multinomial Na¨ıve Bayes and Support Vector Machine concluded that the unigram-trained SVM method is the best method for classifying tweet data in Bengali [10]. In the article “Dataset Indonesia untuk Analisis Sentimen”, it's found that the SVM method produces the best accuracy value compared to the K-Nearest Neighbour (KNN) and Stochastic Gradient Descent (SGD) methods. Based on these studies, the researchers concluded that the best sentiment analysis method for classifying public opinion about Bjorka was the Support Vector Machine (SVM) Method [11].

Sentiment analysis results with SVM are divided into positive and negative sentiments, there will be an advanced analysis of topic modeling using the Latent Dirichlet Allocation (LDA) Method. This algorithm is applied to extract essential topics such as cybercrime, especially controversial Bjorka hackers, into several topics discussed about attitudes and perceptions formed by society [12].

The LDA model produces a brief, clear, and coherent summary and can perform information retrieval, especially in information taking, specifically document classification and modeling of connection between topics [13][14]. LDA can also solve the overfitting problem experienced by the PLSA method. Over-fitting describes a condition where the model has too many parameters that lead to a high match rate for the sample, but the new sample makes the match rate low [15]. LDA is found to be an efficient computational method and can be interpreted by adopting uncreated Language rather than the Latent Semantic Index (LSA) method [16]. In sentiment analysis using Twitter data, it's produced that LDA provides better insights on the topic and better accuracy than LSA [17].

In this study, we aim to combine Support Vector Machine and Latent Dirichlet Allocation techniques to analyze sentiment and model discussion topics related to Bjorka, providing new insights into public perception and discussion dynamics regarding personal data security and privacy in the digital era. Through an in-depth exploration of text data from the Twitter platform, this research is expected to contribute to a more comprehensive understanding of the public's response to cybersecurity issues, while making a significant contribution to the development of more effective security strategies and innovative sentiment analysis and topic modeling methodologies.

Data and Methods

Data

This section describes the methodology used to construct sentiment analysis and Topic Modeling on analysis. Data were taken from comments of social media Twitter users related to issues or problems pro-con Hacker Bjorka with the keywords ”Bjorkanism” and ”Hacker Bjorka”, totaling 15206 tweets. Tweet data is data in the form of unstructured text, it happens because the data still contains a lot of noise, so the classification stage cannot be carried out, there needs to be a process so that the data becomes more structured or commonly known as data preprocessing. The preprocessing step is used to clean text data, there are several stages, including case folding, tokenizing, cleaning, filtering, and stemming.

The following step was labeling the sentiment class into three classes, namely the positive sentiment class of 2035 tweets, the negative sentiment class of 1992 tweets, and the remaining neutral sentiment classes. In the topic modeling process, only positive sentiment classes and negative sentiment classes are used from the SVM results, because we want to see what topics are discussed by the social media Twitter community using the Latent Dirichlet Allocation (LDA) method. The following are the steps of the research flow, presented in **Figure 1**.

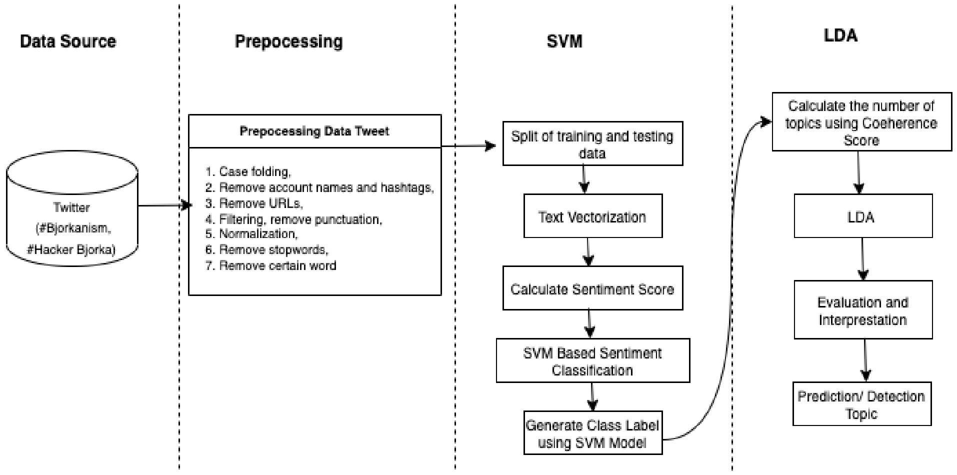

Data Source
Prepocessing
SVM
LDA
Twitter
(#Bjorkanism,
#Hacker Bjorka)
Prepocessing Data Tweet
1. Case folding,
2. Remove account names and hashtags,
3. Remove URLs,
4. Filtering, remove punctuation,
5. Normalization,
6. Remove stopwords,
7. Remove certain word
Split of training and testing data
Text Vectorization
Calculate Sentiment Score
SVM Based Sentiment Classification
Generate Class Label using SVM Model
Calculate the number of topics using Coeherence Score
LDA
Evaluation and Interpretation
Prediction/ Detection Topic

Figure 1. Research Flowchart

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a classification analysis that uses a field separator (hyperplane) in classifying. In a p-dimensional space, a hyperplane will have a p-1 dimension. SVM can produce various possible hyperplanes. The best separator field is the one with the maximum margin (maximal margin hyperplane). The margin is the closest distance observed from the training data to the Hyperplane. Observations that have the closest distance to the Hyperplane are called support vectors [18].

On non-linear text data, SVM is modified by including kernel functions. The kernel can be defined as a function that maps data features from initial (low) dimensions to higher features (even much higher). This approach differs from the typical classification method that reduces the initial dimension to simplify the computational process and provide better prediction accuracy [19]. For example, for n data samples $((\Phi(x_1), y_1);\ (\Phi(x_2), y_2);\ \ldots.;\ (\Phi(x_n), y_n))$, dot product of two vectors (x_i) and (x_j) are denoted as $\Phi(x_i)\ \Phi(x_t)$. The dot product value can be calculated without knowing the transformation function Φ by using the components of the two vectors in the origin dimension space, as follows [20]

$$\mathrm{K}(x_i, x_t) = \Phi(x_i)\ \Phi(x_t) \tag{1}$$

The value of $K(x_i, x_t)$ is a kernel function that shows a non-linear mapping in feature space. Dataset predictions with the newly formulated features are as follows:

$$\begin{aligned} \mathrm{f}(\Phi(\mathrm{x})) = \mathrm{sign}(\mathrm{w}.\Phi(x_t) + \mathrm{b}) &= \mathrm{sign}\left(\sum_{i=1}^{ns} \alpha_i y_i\ \Phi(x_i)\ \Phi(x_t) + b\right) \\ &= \mathrm{sign}\left(\sum_{i=1}^{ns} \alpha_i y_i\ \mathrm{K}(x_i, x_t) + b\right) \end{aligned} \tag{2}$$

With:

E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI : 10.20885/EKSAKTA.vol5.iss1.art7

ns = The amount of data that is a support vector
x_i = Support Vector
x_t = Predicted Testing data

Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is one of the topic modeling with generative probability models on document collections with the purpose of the document-making process in collections efficiently and providing explicit representations of documents. LDA assumes that the document consists of words that help define topics and map the document to a list of topics by assigning each word in the document to a different topic [21].

LDA processes two types of data, namely observed variables and hidden variables [15]. Observed variables are a set of forms in a document, while hidden variables are the structure of topics hidden in a document set. The generative process of LDA applied to observed and hidden variables can be formulated as follows:

$$p(\varphi, \theta, z, w|\alpha, \beta) = \prod_{k=1}^{K} p(\varphi_k|\beta) \prod_{m=1}^{M} p(\varphi_m|\alpha) \left(\prod_{n=1}^{N} p(z_{m,n}|\theta_m) p(w_{m,n}|\varphi_k, z_{m,n}) \right) \tag{3}$$

With:
M : Number of documents
N : Number of words in the document
K : Number of Topics
α : Parameter distribution of topics for each document
β : Parameters of word distribution for each topic
φ_k : Word distribution for topic k
θ_m : Topic distribution for document to $-m$
$z_{m,n}$: Topic determination for word to $-n$ in the document m
$w_{m,n}$:word to-n in the document m

Documents are represented as a mix of hidden topics, with each topic characterized by distribution over all words. LDA assumes a generative process for document collection consisting of M documents and each consisting of *N*-word lengths [21][22]:

1. Choose $\theta_m \sim Dir(\alpha)$, with $m \in \{1, \dots, M\}$ and $Dir(\alpha)$ is Dirichlet distribution with symmetric parameters α that are usually spread ($\alpha < 1$)
2. Choose $\varphi_k \sim Dir(\beta)$ with $k \in \{1, \dots, K\}$and β usually spread
3. For each position of words m, n with $m \in \{1, \dots, M\}$and $n \in \{1, \dots, N\}$
   a) Choose a topic $z_{m,n} \sim Multinomial\ (\theta_m)$
   b) Choose a word $w_{m,n} \sim Multinomial\ (\varphi_k\ , z_{m,n})$

Topic Coherence

A measure of coherence in topic modelling is used to evaluate the set of words that compiled the topic. It is based on the basic idea in classification that class members should be more similar to each other than other class members and measures the extent to which the top terms representing a topic are semantically related, relative to other terms in the documents. Coherence is considered more human-interpretable to evaluate the quality of the topic model than any other measure [23]. Evaluation of LDA uses the *Cv* (coherence value) method to find the optimum number of topics. *Cv* based on a sliding window, a set segmentation of top words and an indirect confirmation measure using normalized pointwise mutual information (NPMI) with the following formula [24].

$$C_v = \frac{2}{K.(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} NPMI\,(w_i, w_j) \qquad (4)$$

where

$$NPMI\,(w_i, w_j) = \left(\frac{log \frac{P(w_i, w_j) + \in}{P(w_i).P(w_j)}}{-\log(P(w_i, w_j) + \in)} \right)$$

Perplexity

The log likelihood of a held-out test set is the most popular approach to evaluate a probabilistic model. This is commonly accomplished by dividing the dataset into two parts: one for training and one for testing. A test set in LDA is a collection of unseen documents $\boldsymbol{w_d}$, and the model is characterized by the topic matrix $\boldsymbol{\Phi}$ and the hyperparameter $\boldsymbol{\alpha}$ for document topic distribution. The LDA parameters Θ are omitted since they indicate the topic-distributions for the documents in the training set and may thus be ignored when calculating the likelihood of unseen documents. As a result, we must assess the log-likelihood.

$$\mathcal{L}(w) = \log \mathrm{P}(w|\Phi, \alpha) = \sum_{d} \log \mathrm{P}(w_d|\Phi, \alpha) \qquad (5)$$

Given the topics Φ and the hyperparameter $\boldsymbol{\alpha}$ for topic-distribution $\boldsymbol{\theta_d}$ of documents, compute the $\boldsymbol{w_d}$ of a collection of unseen documents. Unseen document likelihood can be used to compare models, with higher likelihood signifying a better model. The confusion of held-out texts has typically been employed as a measure for topic models can be defined as

$$Perplexity\ (W) = exp\left\{-\frac{\mathcal{L}(w)}{count\ of\ tokens}\right\} \qquad (6)$$

This is a decreasing function of the unseen documents $\boldsymbol{w_d}$ and log-likelihood $\boldsymbol{\mathcal{L}(w)}$; the smaller the perplexity, the better the model. However, because the likelihood $\mathbf{P}(\boldsymbol{w_d}|\boldsymbol{\Phi}, \boldsymbol{\alpha})$ of one document is intractable, the assessment of $\boldsymbol{\mathcal{L}(w)}$, and hence the perplexity, is also intractable. develops several sampling approaches to estimate this probability [25][26][27].

Result and Discussion

Ovierview Data Tweet

The information used in this study was gathered by using the python application's text crawling feature to examine reviews and comments made in Indonesian by internet users about Hacker Bjorka under the hashtags "Bjorkanism" and "Hacker Bjorka" on the Twitter social networking site. There were a total of 4027 data reviews conducted, and the results of the data obtained during crawling are shown in **Table 1.**

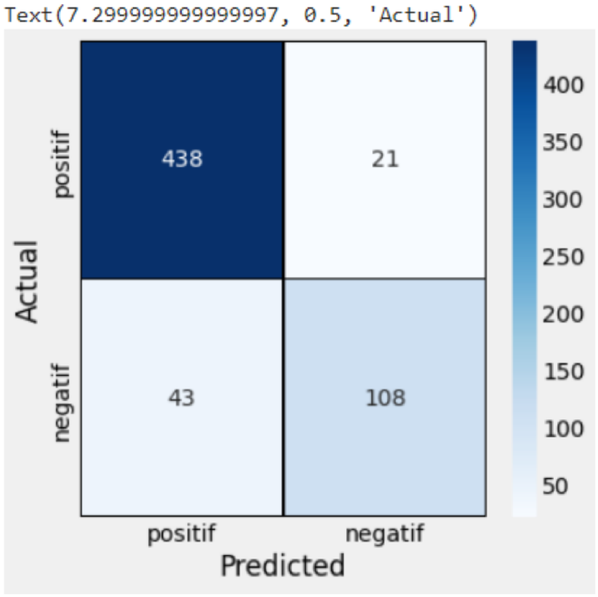
E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI : 10.20885/EKSAKTA.vol5.iss1.art7

Table 1. Data Tweet of Hacker Bjorka

Text
Jadi Tersangka Keterlibatan Hacker Bjorka, MAH Si Penjual Es Tak Ditahan Polisi, Ini Alasannya #bjorkanism #bjorkanesian #bjorka #hackerbjorka #hacking https://t.co/DLEfi0X7LK
…………………
Ramai kasus hacker Bjorka, inilah cara yang bisa diterapkan untuk mengamankan data pribadi. #Bjorka #bjorkanism #hackerbjorka https://t.co/cES6L0IBU9

Table 1 shows some of the tweet data, but it's possible that some of the characters, punctuation marks, and URL addresses listed there won't help with further analysis, will become an error factor, and will contribute to skewed results. Therefore, it is necessary to perform the Text Mining stage of preprocessing in order to generate the frequency of words that often appear in a set of documents.

Accuracy in training is 97.5% and accuracy in testing is 78.8% when the dataset is split 80:20. Researchers use a confusion matrix to evaluate the SVM model's predictive abilities beyond just accuracy results. The data is summarized in **Figure 2.**



```
Text(7.299999999999997, 0.5, 'Actual')
```

Figure 2. Confusion Matrix of SVM

Figure 2 shows that out of 438 positive labels, the SVM model correctly predicted 21 of them. There were 108 correct predictions made by negative labels, and 43 incorrect ones.

Topic Modelling using LDA

Furthermore, topic modeling analysis was carried out for each sentiment. In addition, a topic modeling analysis was performed on each sentiment. Data that has been labeled and cleaned up in the preprocessing stage is used in topic analysis modeling. A separate topic modeling exercise is conducted for each sentiment (positive and negative). Looking at the most frequently occurring words in each sentiment reveals what is most commonly discussed.

The optimal number of topics for each sentiment is determined based on the coherence value. There were 2 topics that emerged from the positive feedback, and 3 topics that emerged from the negative feedback. The following equation displays the outcomes of the formed topics.

Topic Positive of Sentiment

- **Topic 1**= 0.030(aja) + 0.020(mantap) + 0.018(banget) + 0.018(bang) + 0.017(asli) + 0.016(juga tidak) + 0.016(hacker) + 0.015(bjorkanism) + 0.014(hacker bjorka)
- **Topic 2**=0.061(aja) + 0.031(asli) + 0.026(hacker) + 0.022(presiden) + 0.021(akun) + 0.020(Indonesia) + 0.017(hacker bjorka) + 0.015(lu) + 0.013(udah) + 0.012(sabar)

Topic Negative of Sentiment

- **Topic 1**= 0.016(surat) + 0.016(selengkapnya) + 0.015(hacker) + 0.013(data) + 0.012(Jokowi) + 0.012(nih) + 0.011(hacker bjorka) + 0.011(tidak) + 0.011(ya) + 0.010(juga tidak)
- **Topic 2**= 0.022(hacker) + 0.021(munir) + 0.019(hacker bjorka) + 0.016(data) + 0.016(pembunuh) + 0.015(supersemar) + 0.012(istana) + 0.010(bongkar) + 0.010(pribadi) + 0.009(twitter)
- **Topic 3**= 0.024(hacker bjorka) + 0.024(hacker) + 0.018(kominfo) + 0.018(Indonesia) + 0.012(pemerintah) + 0.011(bodoh) + 0.011(data) + 0.009(terungkap) + 0.009(denny siregar) + 0.008*(tidak)

The model above shows the results of topic modeling positive sentiment. Topic modeling in positive sentiment is mostly concerned with the destruction of government crime data and state officials' personal information. The discussion focuses on society's positive perceptions of cybercrime crimes committed by Bjorka hackers against the Indonesian government, demonstrating that people's perceptions and trust in the government remain low.

Topic modeling in the context of negative sentiment Mostly discusses people's dissatisfaction with the attitudes of bjorka hackers, which were sparked by Kominfo and Denny Siregar, as well as the disclosure of past government crimes. The discussion is about the public's negative perceptions of cybercrime crimes committed by Bjorka hackers against entities that are unimportant in the eyes of netizens; this demonstrates that netizens want Bjorka hackers to use their abilities for more pressing interests, particularly those related to the current government's performance.

Evaluation Method

Perplexity is used as a metric in topic modeling to determine the number of topics. The lower the perplexity value, the better the resulting model, as it measures the likelihood of hidden text logs within the documents used in the testing of the topics' accuracy [28][29]. The perplexity value of the SVM-LDA method was compared with LSTM-LDA can be seen in **Figure 3.**

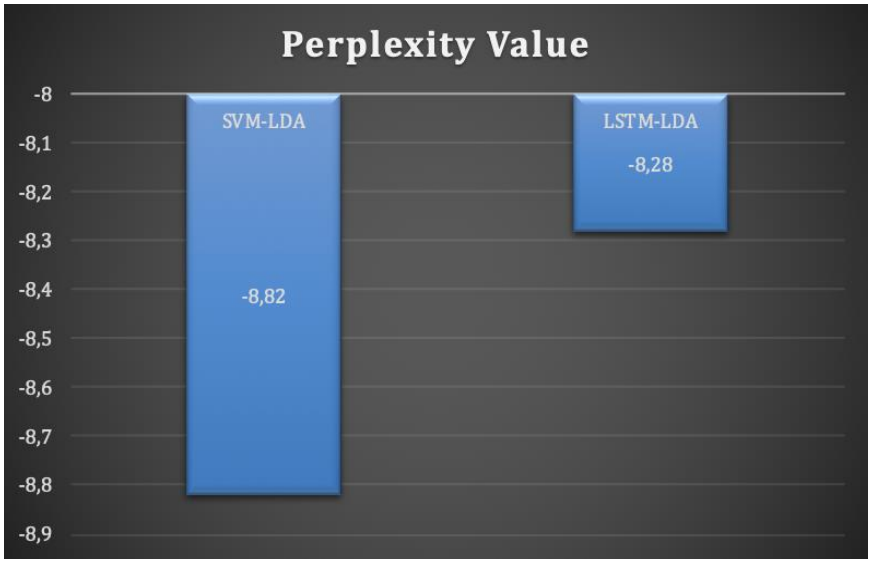


Figure 3. Validation Method of LDA

Figure 3 reveals that the LSTM-LDA approach has a lower perplexity value compared to the SVM-LDA method. This suggests that the LSTM-LDA approach performs better than the SVM-LDA approach.

Conclusion

In this study, the SVM method was used which showed high accuracy in classifying and predicting sentiment, with accuracy rates consistently above 85%. These results illustrate the effectiveness of SVM in sentiment analysis, in line with another study which found that the combination of RBF kernel function in SVM and TF-IDF for feature extraction can achieve an accuracy rate of up to 96.61% in sentiment analysis on Twitter data [30]. In addition, the dominance of positive sentiment in discussions related to Bjorka, which reached 75.2%, highlights the generally favorable public perception of Bjorka's actions in exposing government data. This signifies the importance of a deep understanding of public opinion in responding to cybercrime and data leaks.

In addition, the use of LDA in this study for topic modeling demonstrates its ability to identify key topics in discussions related to Bjorka. This finding gets support from the research of [31] which shows that sentiment analysis and topic modeling using LDA and SVM are effective in classifying hotel reviews into different sentiment categories. Another study by [32] also showed that SVM, when compared to Naïve Bayes, provided the highest accuracy in sentiment analysis of public opinion on Twitter. This demonstrates the adaptability and reliability of SVM and LDA in various text data analysis contexts and opens up opportunities for further research that can incorporate new methods or further elaborate existing approaches to gain deeper and more accurate insights into the dynamics of public opinion. further research needs to be done by elaborating classification methods with sequence-based methods other than LSTM, as it can provide optimal results on topic modeling.

References

[1] I. Anger and C. Kittl, "Measuring influence on Twitter," *ACM Int. Conf. Proceeding Ser.*, no. May, 2011, doi: 10.1145/2024288.2024326.

[2] Tempo Co, "Notorious Hacker Bjorka Denies Police Arrest Claim," 2022. https://en.tempo.co/read/1634887/notorious-hacker-bjorka-denies-police-arrest-claim

[3] DetikNews, "Memaknai Anomali Respons Publik terhadap Hacker Bjorka," 2022. https://news.detik.com/kolom/d-6306007/memaknai-anomali-respons-publik-terhadap-hacker-bjorka

[4] CNN Indonesia, "Drone Emprit: Bjorka Jadi Perbincangan Terpopuler Kalahkan Banjir," 2022. https://www.cnnindonesia.com/teknologi/20220911160859-192-846286/drone-emprit-bjorka-jadi-perbincangan-terpopuler-kalahkan-banjir

[5] Unair News, "Pakar Komunikasi: Aksi Peretasan Bjorka Bukan Bentuk Demonstrasi Modern," 2022. https://unair.ac.id/pakar-komunikasi-aksi-peretasan-bjorka-bukan-bentuk-demonstrasi-modern/

[6] Kominfo, "Indonesia Peringkat ke-2 Dunia Kasus Kejahatan Siber," 2015.

[7] S. Jardim and C. Mora, "Customer reviews sentiment-based analysis and clustering for market-oriented tourism services and products development or positioning," *Procedia Comput. Sci.*, vol. 196, no. 2021, pp. 199–206, 2021, doi: 10.1016/j.procs.2021.12.006.

[8] Y. Al Amrani, M. Lazaar, and K. E. El Kadirp, "Random forest and support vector machine based hybrid approach to sentiment analysis," *Procedia Comput. Sci.*, vol. 127, pp. 511–520, 2018, doi: 10.1016/j.procs.2018.01.150.

[9] S. Dey, S. Wasif, D. S. Tonmoy, S. Sultana, J. Sarkar, and M. Dey, "A Comparative Study of Support Vector Machine and Naive Bayes Classifier for Sentiment Analysis on Amazon Product Reviews," *2020 Int. Conf. Contemp. Comput. Appl. IC3A 2020*, pp. 217–220, 2020, doi:

10.1109/IC3A48958.2020.233300.

[10] K. Sarkar and M. Bhowmick, “Sentiment polarity detection in Bengali tweets using multinomial Naïve Bayes and support vector machines,” *2017 IEEE Calcutta Conf. CALCON 2017 - Proc.*, vol. 2018-Janua, pp. 31–35, 2017, doi: 10.1109/CALCON.2017.8280690.

[11] R. Ferdiana, F. Jatmiko, D. D. Purwanti, A. S. T. Ayu, and W. F. Dicka, “Dataset Indonesia untuk Analisis Sentimen,” *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 8, no. 4, p. 334, 2019, doi: 10.22146/jnteti.v8i4.533.

[12] M. B. Mutanga and A. Abayomi, “Tweeting on COVID-19 pandemic in South Africa: LDA-based topic modelling approach,” *African J. Sci. Technol. Innov. Dev.*, vol. 14, no. 1, pp. 163–172, 2022, doi: 10.1080/20421338.2020.1817262.

[13] R. Rani and D. K. Lobiyal, “An extractive text summarization approach using tagged-LDA based topic modeling,” *Multimed. Tools Appl.*, vol. 80, no. 3, pp. 3275–3305, 2021, doi: 10.1007/s11042-020-09549-3.

[14] S. Bellaouar, M. M. Bellaouar, and I. E. Ghada, “Topic modeling: Comparison of LSA and LDA on scientific publications,” *ACM Int. Conf. Proceeding Ser.*, pp. 59–64, 2021, doi: 10.1145/3456146.3456156.

[15] D. M. Blei, “Probabilistic Topic Models,” *Commun. ACM*, vol. 55, no. 4, pp. 77–84, 2012, doi: 10.1145/2133806.2133826.

[16] V. K. Garbhapu, “A comparative analysis of Latent Semantic analysis and Latent Dirichlet allocation topic modeling methods using Bible data,” *Indian J. Sci. Technol.*, vol. 13, no. 44, pp. 4474–4482, 2020, doi: 10.17485/ijst/v13i44.1479.

[17] H. P. Suresha and K. Kumar Tiwari, “Topic Modeling and Sentiment Analysis of Electric Vehicles of Twitter Data,” *Asian J. Res. Comput. Sci.*, vol. 12, no. 2, pp. 13–29, 2021, doi: 10.9734/ajrcos/2021/v12i230278.

[18] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machine*. Cambridge: Cambridge University Press, 2000.

[19] R. Nariswari and H. Pudjihastuti, “Support Vector Machine Method for Predicting Non-Linear Data,” *Procedia Comput. Sci.*, vol. 227, pp. 884–891, 2023, doi: 10.1016/j.procs.2023.10.595.

[20] B. Yekkehkhany, A. Safari, S. Homayouni, and M. Hasanlou, “A comparison study of different kernel functions for SVM-based classification of multi-temporal polarimetry SAR data,” *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. - ISPRS Arch.*, vol. 40, no. 2W3, pp. 281–285, 2014, doi: 10.5194/isprsarchives-XL-2-W3-281-2014.

[21] D. M. Blei, A. Y. Ng, and M. . Jordan, *Latent Dirichlet Allocation*, vol. 3. 2003. doi: 10.1016/B978-0-12-411519-4.00006-9.

[22] M. Kim, Y. Park, and J. Yoon, “Generating patent development maps for technology monitoring using semantic patent-topic analysis,” *Comput. Ind. Eng.*, vol. 98, pp. 289–299, 2016, doi: 10.1016/j.cie.2016.06.006.

[23] S. Lafia, W. Kuhn, K. Caylor, and L. Hemphill, “Mapping research topics at multiple levels of detail,” *Patterns*, vol. 2, no. 3, p. 100210, 2021, doi: 10.1016/j.patter.2021.100210.

[24] M. Röder, A. Both, and A. Hinneburg, “Exploring the space of topic coherence measures,” *WSDM 2015 - Proc. 8th ACM Int. Conf. Web Search Data Min.*, pp. 399–408, 2015, doi: 10.1145/2684822.2685324.

[25] L. Huang, J. Ma, and C. Chen, “Topic detection from microblogs using T-LDA and perplexity,” *Proc. - 2017 24th Asia-Pacific Softw. Eng. Conf. Work. APSECW 2017*, vol. 2018-Janua, pp. 71–77, 2018, doi: 10.1109/APSECW.2017.11.

[26] P. Tijare and P. J. Rani, “Exploring popular topic models,” *J. Phys. Conf. Ser.*, vol. 1706, no. 1, 2020, doi: 10.1088/1742-6596/1706/1/012171.

[27] B. X. Du and G. Y. Liu, “Topic Analysis in LDA Based on Keywords Selection,” *J. Comput.*, vol. 32, no. 4, pp. 001–012, 2021, doi: 10.53106/199115992021083204001.

[28] K. W. Lim, C. Chen, and W. Buntine, “Twitter-Network Topic Model: A Full Bayesian

Treatment for Social Network and Text Modeling," pp. 1–6, 2016, [Online]. Available: http://arxiv.org/abs/1609.06791

[29] J. Rieger, "ldaPrototype: A method in R to get a Prototype of multiple Latent Dirichlet Allocations," *J. Open Source Softw.*, vol. 5, no. 51, p. 2181, 2020, doi: 10.21105/joss.02181.

[30] P. H. Prastyo, I. Ardiyanto, and R. Hidayat, "Indonesian Sentiment Analysis: An Experimental Study of Four Kernel Functions on SVM Algorithm with TF-IDF," *2020 Int. Conf. Data Anal. Bus. Ind. W. Towar. a Sustain. Econ. ICDABI 2020*, pp. 18–23, 2020, doi: 10.1109/ICDABI51230.2020.9325685.

[31] E. Erniyati, P. Harsani, M. Mulyati, and L. D. Fahriza, "Topic Modeling LDA and SVM in Sentiment Analysis of Hotel Reviews," *Komputasi J. Ilm. Ilmu Komput. dan Mat.*, vol. 20, no. 2, pp. 93–100, 2023, doi: 10.33751/komputasi.v20i2.7604.

[32] G. R. Gustisa Wisnu, Ahmadi, A. R. Muttaqi, A. B. Santoso, P. K. Putra, and I. Budi, "Sentiment analysis and topic modelling of 2018 central java gubernatorial election using twitter data," *2020 Int. Work. Big Data Inf. Secur. IWBIS 2020*, pp. 35–40, 2020, doi: 10.1109/IWBIS50925.2020.9255583.