ISSN 2614-0047

Topic modeling using latent dirichlet allocation (LDA) on twitter data with Indonesia keyword

Edi Surya Negara [1,*], Dendi Triadi [2]

[a] Data Science Interdisciplinary Research Center, Computer Science, Universitas Bina Darma, Palembang, Indonesia
[1] e.s.negara@binadarma.ac.id*; [2] dendi.triadi@binadarma.ac.id
* corresponding author

ARTICLE INFO

Article history
Received August 4, 2021
Revised August 20, 2021
Accepted September 2, 2021

Keywords
Classification
Text Mining
News Documents
Natural Language Processing
Latent Dirichlet Allocation (LDA)

ABSTRACT

Digital transformation causes an increase in the volume of information in the form of text such as news. On social media, a lot of news is uploaded in such a fast time and one of them is Twitter. Twitter is a social media service that has served many users, making it one of the social media that has very large data. From this very large data, it can be used as a news source for online news web. However, with the many topics extracted from Twitter data, the incoming data has a variety of topics which causes difficulties in identifying the topics from the data set taken and will require a lot of time if it has to be done manually by humans. Meanwhile, the data is potentially needed to provide information as quickly as possible. This study aims to classify topics on data taken from Twitter automatically so that it can make a classification on the news taken, can be more effective and efficient and does not take as much time as done manually by humans. The research was conducted using the Latent Dirichlet Allocation (LDA) method. News documents that will be classified are Indonesian news documents and will be classified into topics to be determined. The results of the research using topic modeling using the LDA method concluded that the number of topics formed from 9094 tweet data was 10 topics.

1. Introduction

Social media is one of the factors that causes changes in people's social interactions. From social media, humans get a lot of information because the scope of social media is unlimited. Social media exists in many different forms, including social networks, internet forums, weblogs, social blogs, micro blogging, wikis, podcasts, images, videos, ratings and social bookmarking [1]. Of the many social media Twitter is one of the most popular.

Twitter exists as a means of communication to exchange information about events in the real world, short messages on twitter in general Influence various events experienced by users in real-time [2]. The huge Twitter data can be used as a data source for online news webs. However, with the many topics extracted from Twitter data, the incoming data has a variety of topics which causes difficulties in identifying the topics from the data set taken and will require a lot of time if it has to be done manually by humans. Meanwhile, the data is potentially needed to provide information as quickly as possible [3].

Summarization is a frequent biomedical text mining activity that heavily relies on information extraction. Summarization is the process of automatically finding the most salient features of one or more papers and representing them in a logical manner. It has recently attracted significant interest as a result of the massive growth of unstructured data in the biomedical arena, such as scholarly papers and clinical data. [4]. A large volume of news stories presents a potential difficulty in the work of automatic classification. The debates over how to classify English news stories have been extensively

examined. This differs from the automatic categorisation of Indonesian news articles. The implemented classification method is confined to classical methods such as Naive Bayes and Support Vector Machine. Both techniques are rigorous in their classification of documents into a single topic. As a result, we employ one of the Topic Modeling techniques, in which a document is represented as a distribution of topics, each of which is represented by a set of words. Latent Dirichlet Allocation is the technique used. The experimental investigation is conducted using a 10-fold cross validation technique and numerous parameters, including the number of subjects (5, 10, and 15) and both LDA hyperparameters (0.001, 0.01, and 0.1). The result indicates that the best overall accuracy is approximately 70% when categorizing documents of Indonesian news stories into five categories, including economic, tourist, criminal, sport, and politics. [5].

Classification using independent variables has been widely implemented using the Naive Bayes method and the results depend on the features used, and do not apply if the conditional probability is zero. while classification using the SVM method there are many texts that cannot be classified correctly due to high dimensional characteristics, still rigid and performance depending on the selection of kernel functions causing data sparseness problems [6].

This article presents text classification with a topic modeling approach with the aim of classifying news topics for online news webs using the Latent Dirichlet Allocation (LDA) method. The news texts that will be used in this study include the results of crawling Twitter data using the Application Programming Interface (API) provided by Twitter to produce text data sets based on updates uploaded by Twitter users. The crawling process is successful on Twitter data using the Application Programming Interface and has produced informative data through the Crawling process [7].

Research topic modeling using the Latent Dirichlet Allocation (LDA) method has been carried out by several previous researchers. But this article presents topic modeling using the keyword "Indonesia" by crawling the data create_at, id, id_str, text, entity, metadata, source, user, geo, coordinates, retweet which will be used for online news web development with the topic of information modeling, sentiment analysis and geolocation graph mapping.

2. Literature Review

2.1. Social Media

Social media is a type of media that is comprised of three components: an information infrastructure and tools for producing and disseminating media content; and Individuals, organizations, and industries make and consume media information in digital form in the form of personal communications, news, ideas, and cultural items. [8]. Social media and other online media are a place for information and a place for disseminating information that can be accessed for anyone who wants to find information easily through the internet, either through smartphones or with computers/laptops. Social media is defined as an online information technology tool that enables users to connect easily via the internet in the form of text messages, audio, video, and photographs [9].

2.2. Twitter

Twitter is an online social networking and microblogging website that enables users to send and read text-based messages of up to 140 characters, which was raised to 280 characters on November 7, 2017 and is referred to as tweets. Jack Dorsey founded Twitter in March 2006, and the social networking platform launched in July. Since its inception, Twitter has grown to become one of the top ten most visited websites on the Internet and has been termed the "internet's short message service." Since its inception in 2006, Twitter has grown to be one of the most popular social media platforms on the planet. Indonesia has been a Twitter user since the service's launch and is one of the most productive users. Indonesia has approximately 29 million Twitter users. In 2014, Indonesia was rated sixth in terms of the number of tweets. Numerous pieces of information can be gleaned from the social media platform Twitter, which is becoming increasingly popular. It is used for a variety of purposes, including public, government, and business purposes. Twitter users share a range of different types of tweets. Users can express their opinions on shared tweets. The term "topic" refers to the substance of numerous tweets that cover the same subject. By running topic analysis on these tweets, you may determine the primary themes being discussed by people at the moment [10].

2.3. Topic Modelling

The notion of topic modeling is made up of three entities: "word," "document," and "corpora." The term "word" refers to the fundamental unit of discrete data in a text, which is defined as a piece of vocabulary that is indexed for each unique word in the document. The term "Document" refers to an array of N words. Corpora is the plural version of corpus, which is a collection of M documents. While "subject" refers to the distribution of a specific vocabulary word. To put it simply, each document in the corpus contains a unique fraction of topics mentioned based on the words included within [11]. The basic idea of Topic Modeling is a topic consisting of words certain words that make up the topic, in a document may be composed of several topics and probabilities. While human thought documents are something that can be observed but topics, topic distribution per document, and labeling on each topic are hidden structures, therefore topic modeling helps to find topics and words contained in the document [12]. The concept of topic modeling is aimed at right in Fig. 1.

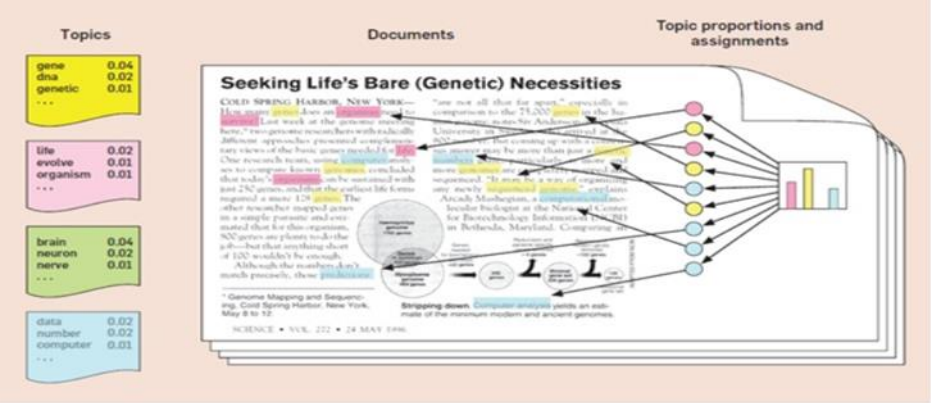

Fig. 1. Topic Modeling Concept

2.4. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a topic modeling and analysis method that is currently in great demand. In analyzing large documents, LDA is present as one of the methods that can be used. One of the functions of LDA is that it can be used to summarize, cluster, and connect or process large data, this is because LDA can produce a list of topics that are weighted in each document [13]. LDA is an analysis method on very large documents. LDA can be used to summarize, cluster, connect or process very large data because LDA produces a list of topics that are weighted for each document [13]. The Dirichlet distribution is used to obtain the topic distribution for each text, and the LDA generating process uses the Dirichlet results as a proposal for allocating words in the document to different subjects. However, from a human perspective, documents are visible elements, whereas topics, their distribution within documents, and the classification of words inside topics are hidden structures; thus, this approach is named Latent Dirichlet Allocation (LDA) [11]. Latent Dirichlet Allocation (LDA) is a method in topic modeling that provides flexibility in organizing, understanding, searching and summarizing electronic archives that have proven to be successfully implemented quite well in the text field and information retrieval. Blei represents the LDA method as a probabilistic model visually as shown in Fig. 2.

3. Method

The method used in this study is to use a Text Mining approach with the main objective of classifying text based on modeling topics. Iteratively, the research steps are arranged based on the framework shown in the sub-chapter of the research stages.

In summary, the development of the topic modeling method began with the creation of the Latent Semantic Indexing (LSI) algorithm, which attempted to address issues with the tf-idf scheme's dimension reduction. Then, the LSI method was developed utilizing the maximum likelihood or Bayesian approach, dubbed the Probabilistic LSI (PLSI) method, which incorporates the concept of probability [14]. Although the PLSI method can be useful for probabilistic modeling of topics, it does not fully produce probabilistic models at the document level. To solve this problem, the Latent Dirichlet Allocation method emerged [11]. While the LDA method can perform effectively at the

document level for documents that contain a large number of subjects, it cannot be used optimally for topic modeling in datasets that contain compact language, such as the twitter dataset. To accomplish this, we require the appropriate method for modeling topics on a small dataset, which is why LDA was developed into twitter-LDA for the purpose of modeling topics on the twitter dataset [15].

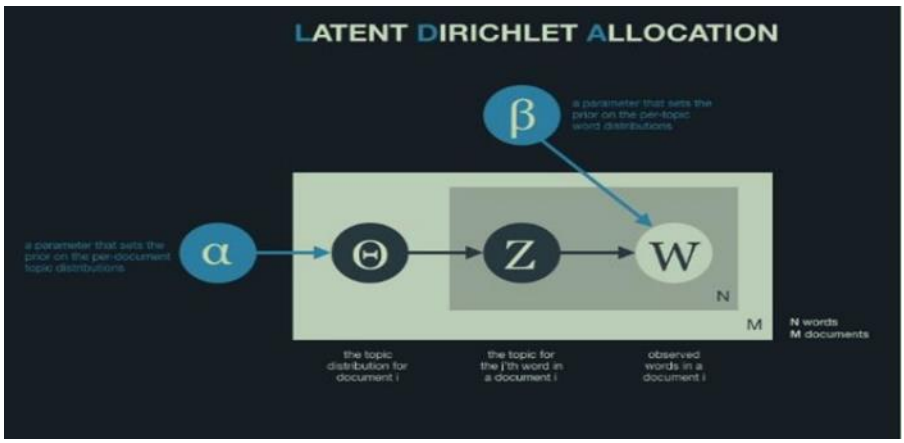

Fig. 2. Topic modeling visualization with LDA method [11].

3.1. Research Stages

This research was conducted by following the main framework of Text Mining which is divided into four stages. The framework is shown in Fig. 3.

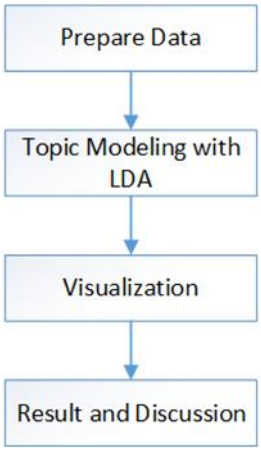

Fig. 3. Research stages

In Fig. 3 it is explained that the research was carried out by preparing the data in advance, if the data has been prepared it will proceed to the topic modeling stage with the Latent Dirichlet Allocation method, after the data is processed with this method, it will proceed to visualization to display the results of the data that has been processed. If the visualization has been completed, the data obtained will enter the results and discussion stage [16].

3.2. Modeling Topics with Latent Dirichlet Allocation

At the Stages, the modeling topic with Latent Dirichlet Allocation (LDA) consists of several stages, namely: Document, Pre-Processing, Modeling Topic, Latent Dirichlet Allocation, Output Document, these stages are shown in Fig. 4.

In general, the classification of text based on modeling topics using the Latent Dirichlet Allocation algorithm is carried out using five main stages. The first stage is to prepare a document or dataset that has followed the format in data processing. The second stage is pre-processing, where at this stage the data is cleaned. The third stage is modeling with the aim of classifying text by preparing training data as the main model for topic classification. The fourth stage is to run the Latent Dirichlet Allocation algorithm to calculate and execute the test data that has been prepared. The fifth stage is the result or output of the resulting modeling topic [17].

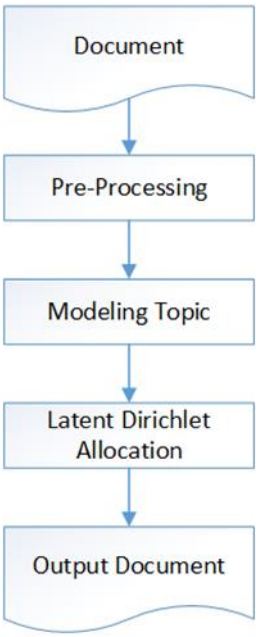

Fig. 4. Modeling Topic Stage with Latent Dirichlet Allocation (LDA)

3.3. Corpus Pre-Processing

In conducting modeling topics with LDA, steps are needed to prepare the data so that it can be processed at the next stage, this stage is called the corpus pre-processing stage [18]. The sub-activities of the corpus preprocessing stage are shown in the Fig. 5.

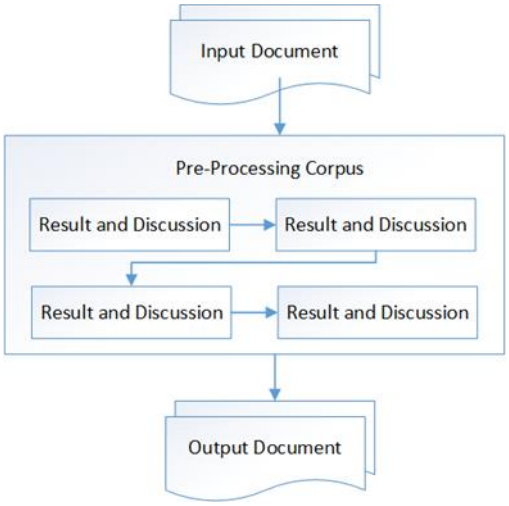

Fig. 5. Sub-activity of the corpus pre-processing stage [19]

The inputted document will be formed into a lovercase with the intention that the same word but different in capitals and not, is considered the same word. Furthermore, the Takenization process is the process of separating the data in the sentence into single word pieces or termed words. Stopwords are common words that have no meaning and usually occur in large numbers [20]. Given the high frequency of occurrence of stopwords in a document, making the step of eliminating stopwords is very important, because it will make the topic not interpreted properly. In the stemming process it is used to change the form of a word into the root word of the word. The stemming process will eliminate all affixes, which consist of prefixes, insertions, suffixes and combinations of prefixes and suffixes on derived words, because text data needs to be formed into basic words so that there are no words that are the same but different because of affixes [21].

3.4. Modeling Topics

The Latent Dirichlet Allocation stage aims to ensure that the topic model generated from the topic modeling results carried out on the document is correct, both in the form of topics and words in the topic. In the Topic Model Validation stage, the topic's level of truth is adjusted according to the Perplexit method and based on the level of coherence [22]. Here is how the LDA algorithm works, which will be shown in the (1).

$$P(W,Z,\emptyset,\varphi;\ \alpha,\beta) = \prod_{I=1}^{K} p(\varphi_i;\beta) \prod_{j=1}^{M} P(\emptyset_j;\ \alpha) \prod_{t=1}^{N} P\ (Z_{j,t}|\emptyset_j) P(W_{j,t}|\varphi Z_{j,t}) \quad (1)$$

Note:

M : Shows the number of documents

N : The number of words in a particular document (document i N_words)

A : Dirichlet prior parameter on the topic distribution per document

B : Dirichlet prior parameter in the word-by-topic distribution

∅i : Topic distribution for document i

ϕ k : Word distribution for the topic k

z_{ij} : Topic for the jth word in document i

w_{ij} : Specific word.

4. Results and Discussion

In this section, we will discuss the results of the Twitter text mining trial for Topic classification on the online news web with a text mining approach using Latent Dirichlet Allocation (LDA). The trial was carried out with the aim of facilitating decision making in choosing topics to be made for news on the online news web [23].

4.1. Implementation

In conducting this research, it has been carried out in several stages, the first stage is Crawling Twitter data. At this stage the crawling of twitter data is carried out with the keyword "Indonesia" as the hashtag. Then the results of the crawl are changed from json to csv format before going to the next stage. The next stage is topic modeling with the Latent Dirichlet Allocation method, at this stage it will be managed through the Text preprocessing process, where there will be two actions to be taken, namely Lowercase and Tokenizing before finally arriving at the Latent Dirichlet Allocation (LDA) stage. From the results of the implementation of the Latent Dirichlet Allocation (LDA) stage, it will be visualized in the form of a word cloud to make it easier to understand. With visualization, it is hoped that it will make it easier for decision makers what topic to take in making news on the online news web [24].

4.2. Crawling Implementation

This twitter data crawl uses the twitter search API as the processed source text. In the process of crawling data using the python programming language. How to use it is to run the source code that has been prepared. In the management it takes a Consumer key, consumer secret as authentication to access steam twitter that has been provided by the search API [25]. In downloading tweets, the keyword "Indonesia" is written in the source code. After running it will download tweet data and then save it in json form and change it into csv form to continue to the next process. From the process carried out when the source code is finished, the data obtained is complete information containing information about the text of the tweet itself in the form of created_at, id, id_str, text, entities, metadata, source, user, geo, coordinates, retweeted, and others concerned. with full profile of wteet author or twitter user and saved in json form. Fig. 6 shows the small description of 9094 twitter data in json form obtained.

Fig. 6. Crawling results in json form

4.3. Implementation of Latent Dirichlet Allocation

At this stage the topic modeling with Latent Dirichlet Allocation is the stage where the Latent Dirichlet Allocaion process is called. This process is unsupervised learning where machine learning looks for previously undetected patterns in a data set without pre-existing labels and with minimal human supervision. So that the results of this method will be taken by the researcher through the results of the LDA and it is concluded that the topic belongs to a certain category. Topics that will be formed from the processed documents are ten topics. To get maximum results, the document will be trained with vector data that has been previously input. Fig. 7 presents the result of the topic call source code.

```
[['indonesia', 'rakt', 'perintah', 'jakaa', 'country', 'lobster', 'cinta'],
 ['indonesia', 'china', 'india', 'tau', 'normal', 'menteri', 'promo', 'kerja'],
 ['indonesia',
  'philippines',
  'korea',
  'sinpore',
  'corona',
  'gw',
  'virus',
  'nih'],
 ['indonesia', 'nera', 'ui', 'coronavirus', 'collar', 'malaysia', 'kasih'],
 ['indonesia', 'udah', 'ka', 'video', 'bank', 'happy', 'love', 'bihday'],
 ['indonesia', 'pasuk', 'org', 'selamat', 'hasil', 'jawa', 'ovehinking'],
 ['indonesia', 'covid', 'anak', 'masuk', 'ra', 'bangsa', 'telkom'],
 ['indonesia', 'bahasa', 'orang', 'nct', 'baca', 'pake', 'inggris'],
 ['indonesia', 'uibergerak', 'diambukansolusi', 'ukt', 'dr', 'hukum', 'turun'],
 ['indonesia', 'bal', 'orang', 'papua', 'pt', 'thread', 'maju']]
```

Fig. 7. Results of topic calling

4.4. Visualization

At the visualization stage, it will be done using wordcloud. Wordcloud are words that exist, depicted in a visualization where the size of the letters depends on the frequency with which the word appears in the text. From the results of the previous process will be directly displayed by means of wordcloud. The results of the visualization can be seen in Fig. 8.

Fig. 8. Display of wordcloud

5. Conclusion

After doing the analysis and testing that was done previously, the resulting data shows that Latent Dirichlet Allocation can be used for text mining. The data generated from the twitter API crawling process is 9094 documents. From the 9094 document data produced, it was cleaned using text processing and generated tweet data from 9094 to 2909 and resulted in 10 main topics. From the process of this medeling topic, it can be concluded that Latent Dirichlet Allocation can be used for text mining.

Acknowledgment

We thank the Data Science Interdisciplinary Research Center Universitas Bina Darma for providing laboratory facilities in conducting this research.

References

[1] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Bus. Horiz.*, vol. 53, no. 1, pp. 59–68, 2010, doi: 10.1016/j.bushor.2009.09.003.

[2] H. Becker, M. Naaman, and L. Gravano, "Beyond Trending Topics: Real-World Event Identification on Twitter," 2011.

[3] S. Hong, "Online news on Twitter: Newspapers' social media adoption and their online readership," *Inf. Econ. Policy*, vol. 24, no. 1, pp. 69–74, Mar. 2012, doi: 10.1016/j.infoecopol.2012.01.004.

[4] M. Allahyari *et al.*, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," *arXiv Prepr. arXiv1707.02919*, Jul. 2017, [Online]. Available: http://arxiv.org/abs/1707.02919.

[5] R. Kusumaningrum, M. I. A. Wiedjayanto, and S. Adhy, "Classification of Indonesian news articles based on Latent Dirichlet Allocation," in *2016 International Conference on Data and Software Engineering (ICoDSE)*, 2016, pp. 1–5.

[6] P. Haffner, G. Tur, and J. H. Wright, "Optimizing SVMs for complex call classification," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, 2003, vol. 1, pp. I-632-I–635, doi: 10.1109/ICASSP.2003.1198860.

[7] E. S. Negara, D. Triadi, and R. Andryani, "Topic Modelling Twitter Data with Latent Dirichlet Allocation Method," in *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*, Oct. 2019, pp. 386–390, doi: 10.1109/ICECOS47637.2019.8984523.

[8] P. N. Howard and M. R. Parks, "Social Media and Political Change: Capacity, Constraint, and Consequence," *J. Commun.*, vol. 62, no. 2, pp. 359–362, Apr. 2012, doi: 10.1111/j.1460-2466.2012.01626.x.

[9] M. O. Odewole, "The Role of a Librarian in Using Social Media Tools to Promote the Research Output of HIS/ HER Clienteles," *J. Educ. Pract.*, vol. 8, no. 27, 2017.

[10] A. Ju, S. H. Jeong, and H. I. Chyi, "Will Social Media Save Newspapers?," *Journal. Pract.*, vol. 8, no. 1, pp. 1–17, Jan. 2014, doi: 10.1080/17512786.2013.794022.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003, doi: doi/10.5555/944919.944937.

[12] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012, doi: 10.1145/2133806.2133826.

[13] J. C. Campbell, A. Hindle, and E. Stroulia, "Latent Dirichlet Allocation," in *The Art and Science of Analyzing Software Data*, Elsevier, 2015, pp. 139–159.

[14] L. Chen, N. Tokuda, and A. Nagai, "A new differential LSI space-based probabilistic document classifier," *Inf. Process. Lett.*, vol. 88, no. 5, pp. 203–212, Dec. 2003, doi: 10.1016/j.ipl.2003.09.002.

[15] W. X. Zhao *et al.*, "Comparing Twitter and Traditional Media Using Topic Models," in *Advances in Information Retrieval*, Berlin, Heidelberg: Springer, 2011, pp. 338–349.

[16] K. L.Sumathy and M. Chidambaram, "Text Mining: Concepts, Applications, Tools and Issues An Overview," *Int. J. Comput. Appl.*, vol. 80, no. 4, pp. 29–32, Oct. 2013, doi: 10.5120/13851-1685.

[17] Z. A. Guven, B. Diri, and T. Cakaloglu, "Classification of TurkishTweet emotions by n- stage Latent Dirichlet Allocation," in *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, Apr. 2018, pp. 1–4, doi: 10.1109/EBBT.2018.8391454.

[18] Y. Lu, Q. Mei, and C. Zhai, "Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA," *Inf. Retr. Boston.*, vol. 14, no. 2, pp. 178–203, Apr. 2011, doi: 10.1007/s10791-010-9141-9.

[19] J. Ucherek *et al.*, "Auto-Suggestive Real-Time Classification of Driller Memos into Activity Codes Using Natural Language Processing," Feb. 2020, doi: 10.2118/199593-MS.

[20] V. Jha, N. Manjunath, P. D. Shenoy, and K. R. Venugopal, "HSRA: Hindi stopword removal algorithm," in *2016 International Conference on Microelectronics, Computing and Communications (MicroCom)*, Jan. 2016, pp. 1–5, doi: 10.1109/MicroCom.2016.7522593.

[21] J. R. Méndez, E. L. Iglesias, F. Fdez-Riverola, F. Díaz, and J. M. Corchado, "Analyzing the impact of corpus preprocessing on anti-spam filtering software," *Res. Comput. Sci.*, vol. 17, pp. 129–138, 2005.

[22] D. Maier *et al.*, "Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology," *Commun. Methods Meas.*, vol. 12, no. 2–3, pp. 93–118, Apr. 2018, doi: 10.1080/19312458.2018.1430754.

[23] N. S. Purohit, A. B. Angadi, M. Bhat, and K. C. Gull, "Crawling through web to extract the data from Social networking site - Twitter," in *2015 National Conference on Parallel Computing Technologies (PARCOMPTECH)*, Feb. 2015, pp. 1–6, doi: 10.1109/PARCOMPTECH.2015.7084522.

[24] R. Kusumaningrum, S. Adhy, and S. Suryono, "WCLOUDVIZ: Word Cloud Visualization of Indonesian News Articles Classification Based on Latent Dirichlet Allocation," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 16, no. 4, p. 1752, Aug. 2018, doi: 10.12928/telkomnika.v16i4.8194.

[25] J.-H. Lee, "Building an SNS Crawling System Using Python," *J. Korea Ind. Inf. Syst. Res.*, vol. 23, no. 5, pp. 61–76, 2018, doi: https://doi.org/10.9723/jksiis.2018.23.5.061.