Topic prediction modelling on social media content using machine learning

Izmi Dewi Aisha, Lili Ayu Wulandhari
Department of Computer Science, BINUS Graduate Program-Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Article Info

Article history:

Received Jun 28, 2023
Revised Sep 29, 2023
Accepted Oct 8, 2023

Keywords:

Clustering
Corporate social media
Latent dirichlet allocation
Topic modelling
Topic prediction

ABSTRACT

The simplicity to deliver an opinion about companies or institutions via social media has resulted in both positive and negative judgments. Through social media all positive and negative information will be easily found and spread. It is concerned that negative information will lead to negative public opinion. If this occurs, the company will suffer from a lack of trust, which will harm the company's reputation. Thus, to monitor uncontrolled issues, a company wants to know what topics or opinions are developing in the community. Therefore, the topic modelling using latent dirichlet allocation (LDA) is proposed to identify topics that are being discussed on social media. The findings of this study got the coherence score of 0.558 and based on the direct human judgment, the model got an average 80% correctly. The findings of this study reveal 4 topics groups that represent the corporate social media content. These findings offer information to companies about the latest topics or opinions that are currently developing in society which could provide recommendations related to decision-making on current issues thus increasing the trust and reliability towards the company.

Corresponding Author:

Izmi Dewi Aisha
Department of Computer Science, BINUS Graduate Program-Master of Computer Science
Bina Nusantara University
Jakarta 11480, Indonesia
Email: izmi.aisha@binus.ac.id

1. INTRODUCTION

Recently, the rise of social media has been a huge phenomenon, transforming the way individuals and organizations communicate, exchange information, and engage online. Social media users sometimes express themselves by sharing their problems or complaints about various topics of conversation [1]. Moreover, online social media can be used for commercial purposes, which are to share personal opinions with others about a particular product or business [2]. The widespread use of social media platforms has created an environment where rumors can quickly spread, as people can easily access online news and update information, thus providing many opportunities to disseminate misinformation [3]. The simplicity to deliver an opinion about companies or institutions through social media has resulted in both positive and negative judgments. It is concerned that negative information will lead to negative public opinion. Obtaining opinions and feedback from the public is important for companies, especially if the services are used by the public. This can help companies understand how their products and services are perceived by consumers, identify areas for improvement, and build trust with customers. One of the entities included is a construction company. The construction company has projects that are scattered and used by the community. Positive feedback from the community can indicate public support for a project, thereby increasing the chances of its successful completion, while negative opinions and experiences can raise concerns such as about environmental impacts, safety, or community disruption, which can potentially reduce their trust towards the company. Therefore, negative issues greatly affect how the facilities and infrastructure built by the company will be used in the community. If there are negative issues, it will influence public trust in the results of this project. Thus, construction companies need to know how opinions or topics are developing on social media about the company. However, companies find it extremely difficult to extract useful data and discover trends that are developing due to the overwhelming volume and unstructured nature of social media content. Companies are frequently struggling to extract insights from this information. Understanding the topics or themes that are being discussed in the community is very important for companies to stay relevant and make the right business decisions on a current issue. But manually analyzing and categorizing huge amounts of data is time-consuming, error-prone, and may not capture the essence of the problem. Thus, companies need an automated and efficient approach to identify and extract topics from social media.

Twitter is one of the most widely used social media to communicate with customers [4]. Twitter looks to have a twice greater probability than other social media platforms of improving customer engagement through satisfaction and positive emotions [5]. This information obtained from social media platforms has a significant influence on customer engagement, which can encourage active participation, strengthen brand evaluation, build customer loyalty and trust, increase purchase intentions, and contribute to overall customer satisfaction [6]. In order to better understand the contents of social media, a topic analysis or topic modeling is needed to understand the topics discussed on social media.

Topic modeling is a machine learning technique that organizes and understands a set of data with categories according to the topic or theme of each text [7], and is part of summarization, classification, segmentation, categorization of documents and others [8]. This modeling will find important topics in documents related to the information discussed in the text and find correlations between words, topics and documents. In topic modeling text analysis, it can determine which events or concepts are referred in document [9]. There are several methods in conducting topic modelling, one of them is latent dirichlet allocation (LDA) and latent semantic analysis (LSA). As in this study which compares LSA and LDA on the classification of unstructured scientific text documents (e-books) which shows that LDA has better results compared to LSA which is 0.54846 while LSA is 0.4047 [10]. Ahn *et al.* [11] implemented topic modelling using LDA to find similar and different topics between certain organisations regarding Ridgecrest earthquake tweets to predict public engagement and encourage more effective communication during natural disasters. Other than that, in this study, LDA got a resulting accuracy rate of 71.4% for documents correctly classified, which involves developing topic modeling for medical documents with constructing a document term matrix to capture word occurrences in each document [12].

Mangsor *et al.* [13] implemented in LDA on corporate social responsibility based on annual reports of companies using document clustering k-means and topic modelling LDA, in this study k-means is used to assign k clusters and assign each document to the cluster which later LDA was used in naming the labels that can provide a global picture of themes and their differences, while enabling a comprehensive analysis of the most related keywords to each topic. It was found that 87% of the topics generated by LDA made sense to human judges and could identify multi-faceted themes, extract latent themes that might be missed, and reflect those themes in their raw state without bias in e-petition data [14]. Therefore, one of the most widely used probabilistic text modeling methods in machine learning is LDA [15].

Based on the method of LDA, each document is a collection of a mixture of various topics, and LDA extracts topics from a collection of documents based on the words that are contained in the document. In this case, for a construction company, there are specific terms that are commonly used, they are '*proyek', 'bangunan', 'anggaran', 'warga'* and other specific terms, where we collected the up-to-date data to cover the newest term that will be included in the modeling. These specific terms will be extracted from the content of social media by utilizing the capabilities of LDA. Therefore, based on this background, research is conducted that focuses on the implementation of topic modeling using LDA on social media content, especially on Twitter, related to the construction company. We compare coherence score of Bag of Words and the term frequency-inverse document frequency (TF-IDF) vectorization approach to be used in the LDA. This method is a solution to finding out the topics that are developing in the community about the company, and the results of these findings offer information to companies that can provide recommendations related to decision-making on current issues so as to increase trust and reliability towards the company. This paper is arranged in 5 sections where section 1 provides an overview of the problem as well as an overview of the existing literature and studies in the current knowledge for topic modeling using LDA. Section 2 presents an approach developed by the researchers. Section 3 introduce the theoretical framework and research approach used in the study. Sections 4 and 5 presents the experiments results and conclusion from this research.

2. PROPOSED METHOD

The proposed method in this study is illustrated in Figure 1, which describes how the implementation process will be conducted. Based on Figure 1, the method in this study used the main steps of data collection, data cleansing, feature engineering or vectorization, modeling, and topic analysis. Each process in this topic modeling will be modeled using Python. In the context of this study, the data used for analysis was from Indonesia. By utilizing Indonesian data, this study ensures that the findings and conclusions are directly applicable and relevant to the case study and thus provide a more comprehensive understanding of the topics discussed and the public conditions related to construction company. Sample of data from data collection are shown in Table 1, this experiment used data scrapping from twitter using snscrape library. Scrapping data is collected by using keywords which contained construction company name. From the scrapping, 2,000 tweets are obtained on July 2022 to January 2023. This data scrapping is accomplished by pulling the relevant information or data from online pages and storing it in a processable format, such as a spreadsheet. After obtaining the data, it will be stored in excel format which can be used in further processing.

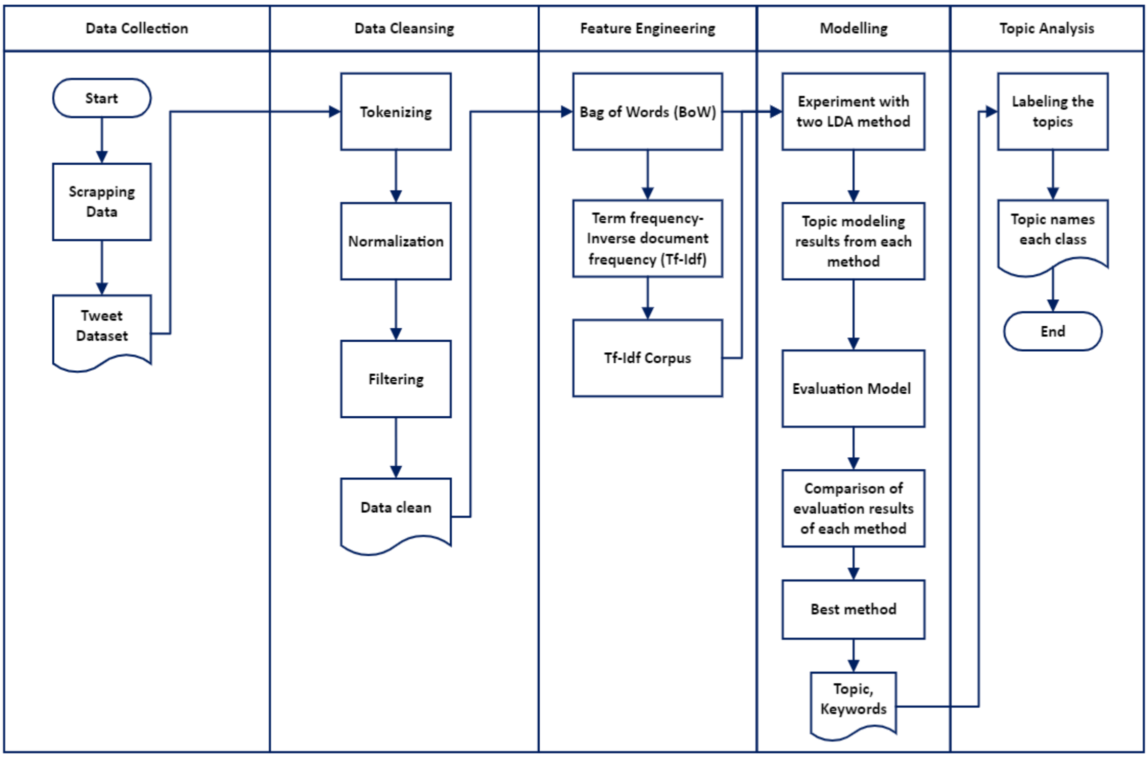


Figure 1. Proposed method

Table 1. Sample scrapping data

No	Text
1	PT xyz *Siap Memulai Pembangunan* Fly Over Aloha, *Untuk Urai Kepadatan Kendaraan Arah* Sidoarjo &amp;ndash; Surabaya.
2	*Dalam pembangunannya, pelaksanaan pembangunan Temef dibagi menjadi 2 paket yaitu untuk Paket 3 dilaksanakan oleh* PT xyz *yang bertugas menyelesaikan bangunan spillway dan bangunan penunjang lainnya*.

In the data cleansing, which consists of the same steps as in section 3.1: tokenizing, normalization, and filtering. Before these processes, hashtags, mentions, and punctuation are removed from tweet data. All or any natural language processing (NLP) jobs require tokenizing, which is the process of breaking a string into its intended component parts [16]. In this tokenization, the researcher uses the phrase tokenizer from the nlp_id library. Phrase tokenizer from nlp_id will identify and tokenize Indonesian phrases or multi-word expressions that commonly occur in the language. As shown in Table 2, which is a sample of the phrase tokenizer. In this phrase tokenizer, capital characters in word prefixes will affect token split, so converting them to lowercase will be done after the phrase tokenizer. This tokenizer considers the unique grammar and syntax of Indonesian to correctly tokenize phrases, which may consist of multiple words.

Table 2. Sample tokenizer

Text	Tokens
Kendaraan Sidoarjo Surabaya	‘*Kendaraan’*,’Sidoarjo Surabaya’
Pelaksanaannya materi budaya organisasi interaktif	*'Pelaksanaan', 'nya', 'materi', 'budaya organisasi', 'interaktif'*

In normalization, where the abbreviated words found in the data will be converted into standard form. Text normalization is essential for increasing the language parser's ability to understanding the lexical meaning of the text [17]. In this study, researchers used the list of public slang words as well as adding words that are often encountered in the datasets. Normalization for out-of-vocabulary (OOV) words could increase the performance of language processing [18]. Therefore, the word which is written in abbreviation should be translated into the original form [15]. Some of these words are written as abbreviations or slang words, as in the words below:

ga → *tidak*
smwa → *semua*
ngamau → *tidak*

In the filtering stage, it used stopwords from the nlp_id library. The use of stopword removal could reduce time in the training process since the size of the dataset was reduced, and it could improve performance since the space was narrowed down [19]. This will allow the process to concentrate on important words. Words like *'ada'*, *'adalah'*, *'adanya'*, *'akhir'* and other common words that frequently appear but have no significant meaning to the overall sentence will be removed. Table 3 shows an example of data cleansing that has been completed.

Table 3. Sample data cleansing

Tweet	Data cleansing
PT xyz *siap memulai pembangunan* fly over aloha*, untuk urai kepadatan kendaraan arah* Sidoarjo &amp;ndash; Surabaya	['*pembangunan*', 'fly over aloha *urai kepadatan', 'kendaraan'*, 'Sidoarjo Surabaya']
Untuk pasok lahan irigasi, bendungan rukoh di Aceh *yang dikerjakan* PT xyz *ditargetkan rampung* 2023	[*'pasok lahan irigasi bendungan rukoh* Aceh*', 'ditargetkan', 'rampung'*]

After the data has been cleaned, the next step is vectorization. In this vectorization, it will be using bag of words (BoW) and TF-IDF. These two vectors will be implemented in topic modeling and will be compared with the coherence score results to get better performance. The identifications (IDs) and frequency of a collection of words are represented by BoW in a document [7]. Therefore, this BoW will calculate how many times the word appears in the document. The creation of this BoW will be done by gensim using doc2bow from the dictionary that has been generated. This dictionary represents a unique word that appears in the entire document. The output of this corpus is the word ID and the frequency of the word in the document as shown in Table 4. Meanwhile, the TF-IDF corpus was created using the gensim model, TF-IDF model. The input for this TF-IDF model was the bag of words that had been created before. The output of this corpus is the word ID and TF-IDF results, which are based on their frequency in the document and rarity throughout the collection. As shown in Table 4 which are sample output using TF-IDF corpus.

The following step is topic modelling using LDA. The LDA algorithm resolves this issue by inferring topics from recurrent patterns of word occurrence in documents because topics are concealed in the first place, there is no information about them that can be directly observed in the data [20]. The first step is to select the number of topics to be used. Researchers frequently test numerous candidate models with various numbers of topics in order to select an appropriate number [20]. Therefore, it will be iterated to select the number of topics by looking at the best coherence score. To determine the number of topics used in the first model, which is using the baseline LDA, testing is conducted with the number of topics between 2 up to 10. After obtaining the number of topics from the baseline LDA which is using BoW and the coherence score as well as their visualization, the same iteration is performed to find the number of topics using TF-IDF corpus. Consequently, in this way, without even affecting the fundamentals of LDA, it tries to purify the vocabulary, which immediately affects the results [19]. When using the TF-IDF corpus, the selection of the number of topics will be used by the initial number of topics being the number selected in the baseline model up to 10. When it comes to how it truly cleanses the vocabulary that is used to refer to the dictionary in this work, the TF-IDF score will eventually make things more clear [19]. After getting the number of topics and the coherence score of the two models, compare the coherence score results to get the best model.

Table 4. Sample output BoW and TF-IDF

Tokens	Output BoW	Output TF-IDF
['*pembangunan*', 'fly over aloha *urai kepadatan', 'kendaraan'*, 'Sidoarjo Surabaya']	[(0, 1), (1, 1), (2, 1), (3, 1)]	[(0, 0.5869309633804809), (1, 0.501911573313011), (2, 0.24312313963294205), (3, 0.5869309633804809)]
['*pasok lahan irigasi bendungan rukoh* Aceh', *'ditargetkan', 'rampung'*]	[(210, 1), (211, 1), (212, 1)]	[(210, 0.5402125780821485), (211, 0.6610414293915309), (212, 0.5207634771274194)]

After getting the best method, the following step is topic analysis to label the clusters generated from the LDA. The LDA will give the topic distribution for each document. Clustering documents can be obtained from the highest topic distribution on each document because each document will have a number of topic distributions for each topic, and the highest topic distribution is the topic cluster of the document. To make topics easier for humans to understand, they may need to be given names with significance [21]. Therefore, in this paper, the name will be manually summarized by observing each keyword and sample sentence taken from the cluster.

3. THEORY AND METHOD

This section presents the concepts and methods utilized in the research, which includes the data cleansing procedure, which includes removing noise and irregularities from the dataset to ensure its quality and reliability, as explained in section 3.1. Section 3.2 goes into detail about the LDA algorithm used for topic modeling. Section 3.3 describes the evaluation criteria used to assess the performance and efficacy of the proposed approaches.

3.1. Data cleansing

Data cleansing refers to the process of correcting incorrect, incomplete, duplicate, or otherwise erroneous data in a dataset. Data cleansing enhances data quality and allows an organization to give more accurate, consistent, and reliable information for decision-making. This step will convert raw data into a form that is easier to understand. The steps are described as:

3.1.1. Removing hashtag, mention, link and punctuations

At this step, tweets which contained mentions, tagging, hashtags, links, retweets, punctuation, digits will be removed. Some emoticons, which are often found in the datasheet, will be translated into strings. This emoticon replacement was used to improve the readability of the data and the meaning expressed in the tweet.

3.1.2. Tokenizer

In this experiment, the Indonesian phrase tokenizer will be used. The phrase tokenizer splits the text into tokens, with the word tokens representing phrases (single or multi-word tokens). It was designed to break down the text into meaningful tokens.

3.1.3. Normalization

Transforming text data into a format that is consistent with the source data. For example, if there are abbreviations or slang words such as '*kzl*, '*ga*, '*smwa'* then normalization will be carried out to bring it more closely to a predetermined standard. This change can help reduce the variety of information that must be managed by the computer; therefore, it can increase efficiency.

3.1.4. Filtering

This step will select important words from the tokenization. Words that are in the stop word will be removed from the document. Removing these words helps the model to consider only key features.

3.2. Latent dirichlet allocation

Topic modelling is an unsupervised learning approach. The algorithm that will be used is LDA. LDA extends probabilistic latent semantic analysis (PLSA) by using dirichlet priors for document-specific subject mixtures, resulting in the discovery of texts that would have usually passed unnoticed and Gibbs sampling which used by LDA to apply the model [13]. Two initial outputs are provided by an LDA model that is an estimated probability that a topic will produce a document (referred to as an affinity score) and the probability that a word will be used to describe a topic [14]. To generate topics, LDA adopts a probability distribution model [22]–[25]. The LDA topic model considerably reduces the dimensionality of the represented text and offers significant benefits when processing large amounts of text [20].

LDA which used in this experiment is from the gensim library. In gensim, the LDA model is generated by defining the corpus dictionary mapping, and number of topics to be used in the model [19]. This modeling will provide the value of topic distribution for each document and keywords for each topic. For the topic frequency of each document it will be calculated by the number of frequencies calculated during initialization and the Dirichlet generated multinominal distribution for topics in each document while for keywords on each topic its calculated by the number of frequencies calculated during initialization and generated Dirichlet-multinominal distribution for each word on each of these topics [13]. Finally, several iterations are performed until convergence is achieved and calculate the topic probability distribution and the probability of each word that best represents of the topic.

$$p(\beta_K, \theta_D, z_D, w_D|a, \eta) = \prod_{k=1}^{K} p(\beta_K|\eta) \prod_{d=1}^{D} p(\theta_d|a) \prod_{n=1}^{N} p\left(z_{d,n}\middle|\theta_d\right) p(w_{d,n}|z_{d,n}, \beta_{d,k}) \tag{1}$$

According to (1) represents the joint distribution of all hidden variables β_K as topics, θ_D is proportion of topics per document, z_D is word topic assignments, w_D is words in documents, every topic β_K is multinominal distribution over the vocabulary V which is comes from a Dirichlet distribution $\beta_K \sim Dir(\eta)$, each document is also represented as a distribution over K topics and which is come from a Dirichlet distribution $\theta_D \sim Dir(a)$, the smoothing of topics within documents is indicated by Dirichlet parameter a and the smoothing of words within topic (η) [26]. This topic modelling use baseline LDA model with BoW as the vectorization approach and will be compared to LDA with TF-IDF. BoW is a corpus consisting of the word id and how often it appears in each document. Therefore, each document will consist of a word id and then the frequency of words in the document such as (word_id, word_count). This will create a corpus using doc2bow which will be inputted in the LDA model.

When using TF-IDF in LDA, at the time of initialization a training dataset with integer values is required like the BoW model [19]. Term frequency-inverse document frequency is the result of two elements form of TF-IDF. The TF-IDF of a word t in a document d is written as (2), Where tf is the term frequency, which means how often a word or term appears in the document as (3), and *idf* stands for inverse document frequency, which refers to a measurement of a term's importance as (4).

$$tf\_idf_{t,d} = tf \times idf \tag{2}$$

$$tf_{t,d} = \frac{n}{T} \tag{3}$$

$$idf = \log\left(\frac{TD}{D_t}\right) \tag{4}$$

In a document d, n is the number of times a word or term appears. T stands for the total number of terms in the document d. The total number of documents in the corpus is *TD*, and the number of documents containing a word or term t is D_t. TF-IDF is used to evaluate how important a word t in document d in a collection corpus. A collection corpus's word importance is determined using the TF-IDF method. This TF-IDF corpus will be the input to the LDA model.

3.3. Evaluation

For the evaluation of model, coherence score and human judgement will be used. In the topic modelling, the average similarity between the top words in a topic with the highest weight i.e the relative distance between top words is measured by the coherence score [19]. In this experiment, gensim's default coherence score was used which is c_v measure. The c_v measure is based on a sliding window, a one-set segmentation of the most important words, and an indirect confirmation measure that makes use of normalized pointwise mutual information (NPMI) and the cosine similarity [27]. This measure consists of 4 parts which are segmentation of the data into word pairs, probability estimation which is calculating word or word pair probabilities, confirmation measure which calculating a confirmation measure that indicated how strongly one word set support another and lastly the aggregation individual confirmation measures into a final coherence score [26].

3.3.1. Segmentation

In c_v measure uses *S-one-set* which is compare individual words to the entire word set W [28]. Let W be the set of topic's top-N most probable words $W = \{W_1, \ldots, W_N\}$. S_i a segmented pair of each word W^l $\in W$ paired with all other words $W^* \in W$ therefore S is the set of all pairs [26].

$$S_{set}^{one} = \{(W^l, W^*)|W^l = \{w_i\}; w_i \in W; W^* = W\} \tag{5}$$

3.3.2. Probability calculation

In c_v measure uses the Boolean sliding window $(P_{sw})(110)$ [28]. To determine whether two words co-occur, a boolean sliding window of size 110 is used. The probabilities will be calculated over a sliding window of size 110 that moves across the texts.

3.3.3. Confirmation measure

For each $S_i = (W^l, W^*)$ will calculate using a confirmation measure $\emptyset$ that determines the degree to which W^* support W^l based on how similar of W^l and W^* in relation to all the words in W and to calculate the similarity, the context vector is used to gather the semantic context for each word within the set W as shown in (7) [26]. The cosine similarity between the two measure vectors determines the final score. The agreement between individual words w_i and w_j using NPMI as demonstrated in (6) [26]. Probabilities of single words $P(w_i)$ and joint probabilities of two words $P(w_i, w_j)$, $\in$ is used for algorithm of zero and γ to give greater weight to higher NPMI values and the vector $\vec{v}(W^l)$ and $\vec{v}(W^*)$ are generated by associating each of them with every word in the set W as shown in (7) [26].

$$NPMI(w_i,w_j)^{\gamma} = \left(\frac{log\frac{P(w_i,w_j)+\in}{P(w_i).P(w_j)}}{-\log(P(w_i,w_j)+\in} \right)^{\gamma} \tag{6}$$

$$\vec{v}(W^l) = \left\{ \sum_{w_i \in W^l} NPMI(w_i,w_j)^{\gamma} \right\}_{j=1,\dots,|W|} \tag{7}$$

$$\emptyset S_i(\vec{v},\vec{w}) = \frac{\sum_{i=1} |W| v_i.w_i}{\|\vec{v}\|_2.\|\vec{w}\|_2} \tag{8}$$

3.3.4. Calculating

The cosine vector similarity of each context vector $\emptyset\, S_i(\vec{v},\vec{w})$ in a pair S_i with $\vec{v}(W^l)\,\epsilon\,\vec{u}$ and $\vec{v}(W^*)\,\epsilon\,\vec{w}$ will generated the confirmation measure $\emptyset$ as shown in (8) [26]. The arithmetic mean of all confirmation measures is used to calculate the final coherence score. It offers a representative measure of the overall coherence in the data. This measure combines the indirect cosine measure with NPMI and a Boolean sliding window. NPMI is a measure relationship between two words that takes into account the frequency of their occurance. Thus, the text will be segmented first into a sliding window with size 110. The probability is calculated through this, NPMI will measure the relationship between the words in the sliding window. Then, the indirect cosine measure is used to measure the similarity between the probability distributions in each word of the window. Based on a thorough empirical comparison with other commonly used topic coherence measures, this measure has been found to be particularly suitable to evaluate the quality of topics and produces scores that are most similar to those of human evaluation [28]. Documents that have the same content should be on topics with the same cluster. Therefore, apart from the coherence score, human evaluation is also required. One of the other evaluation methods proposed is direct human reading and judgment [14]. This agreement will provide feedback whether documents that have the same content are indeed on the same topic. In this review, it is performed by comparing each document in the topic and keywords in the topic that have similarities and are really in the same topic.

4. RESULT AND DISCUSSION

In topic modeling, it used the baseline LDA model which is using BoW and LDA using the TF-IDF corpus. The experiment results show that LDA+BoW gives better performance around 9.1% of the coherence score than LDA+TF-IDF as shown in Table 5. Bag of words as vectorization gives the overview of appearance words in document, where this is the one that LDA needs when it extracts the keywords for each topic. LDA does not care how essential a word is in the document, as demonstrated by TF-IDF, therefore BoW yields better performance when extracting the topic.

Table 5. Coherence score

Topic modelling	Number of topics	Coherence score
LDA+BoW	4	0.558
LDA+TF-IDF	4	0.467

The topic modeling provides the topic distribution for each document in each topic and the top keywords that contribute to the topic as shown in Figure 2. It shows the topic number 3 with 30 keywords. In Figure 2, there are words that are not related, such as '*dr*', '*hingga*', and '*sdh*'. These words are stopwords that should have been removed in the data cleansing section. Due to an oversight during the input of the slang words, unfortunately, these words were not included in the list of slang words and thus were not transformed into their regular forms. For example, the words ‘*dr’*, ‘*sdh’* which should be normalized into ‘*dari’*, ‘*sudah’*. If it is not modified, then when performing the stopwords, this word will not be deleted. That is why there are some unrelated words found to be included in the top keywords. This is one example of a challenge when working with text, there are possibilities that we miss preprocessing the data before modeling. In data cleansing, researchers have performed filtering by stopwords twice: the first stopword is performed during cleansing before punctuation removal, and the second stopword is performed after normalization has been carried out. This consideration is made because the stopwords list contains words that use punctuation, such as ‘*bersama-sama*’, ‘*menanti-nanti*’ and thus it should be done at the beginning before the removal of punctuation. Meanwhile, the second use of stopwords is when normalization has been completed, which is after translating words such as abbreviations or slang. This normalization generally looks at public slang words and adds words that are often encountered in the dataset. However, there is a miss in inputting the word, thus potentially including it in the top keywords.

Therefore, data cleansing is an important and essential step in data preprocessing. In addition, language is constantly evolving, dynamic, and ever-changing. New words and phrases emerge over time, while others may fall out of use. In modern communication, slang, abbreviations, and internet jargon have become commonplace. As language evolves, data cleaning and preprocessing techniques must adapt to effectively handle new words, abbreviations, and language variations. Therefore, the normalization stage is important because it will convert abbreviations and slang to have the same meaning as sentences in other documents. Thus, the list of slang words must be able to keep up with the development of language every time. From the visualization in Figure 2, a summary of keywords that are related to the case in this study is made as in Table 6, which is reflected in the keywords.

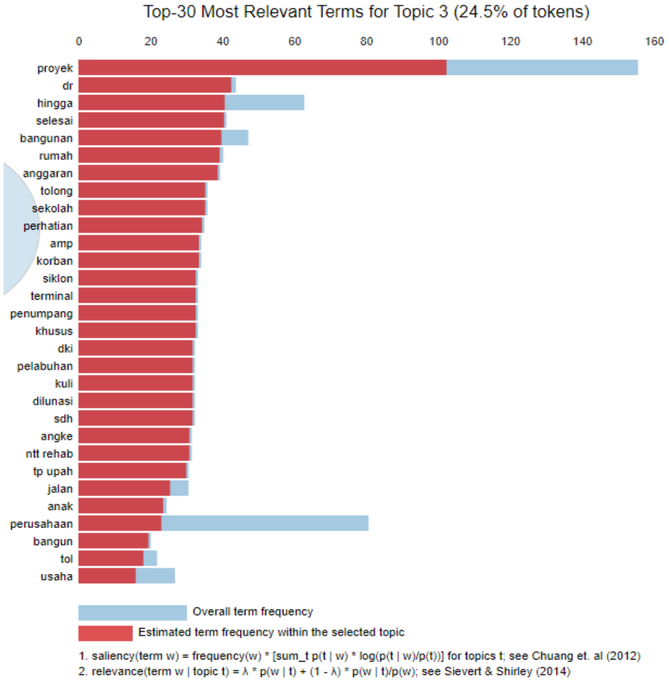
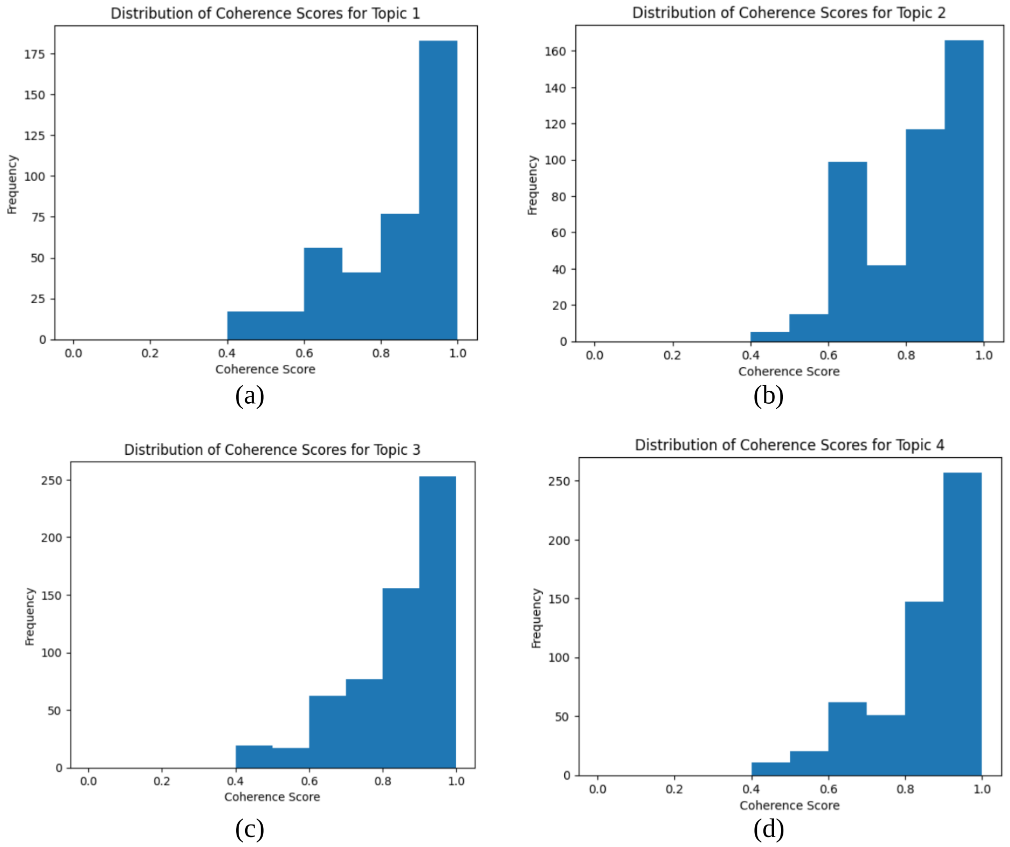


Figure 2. Top 30 most relevant term for topic 3

LDA will display each document's topic distribution values. After obtaining the topic distribution, to determine the coherence score of each document, the highest topic distribution is chosen which is reflected in the dominant topic. Figure 3 shows the distribution of the coherence score, which is used for taking samples that will be used in labeling the topics. This labeling of topics is conducted manually by reading keywords and sample sentences on each topic, which are taken randomly from the topic cluster by observing the dominant coherence score distribution on the topic. Based on Figure 3(a), Samples that will be collected for each topic range from a coherence score of 0.6 to almost 1. The second topic will have sample sentences from documents that have coherence scores between 0.6 and almost 1, as shown in Figure 3(b). The third topic shows a dominant distribution of coherence scores between 0.6 and almost 1, as shown in Figure 3(c). As well as the 4th topic, as shown in Figure 3(d), sample sentences will be taken for review in topic labeling, which are documents that have coherence scores between 0.6 and almost 1. The total number of samples that are collected is 22 for each topic. The labeling of the topic is conducted by reviewing the keywords of the topic and the sample sentences collected. In Table 6, The topic name of each cluster is a review of the sample of each topic that was taken and the contribution of keywords that are present in each topic, after that a summary will be made to name the topic. This topic name was chosen by several reviewers to get topic suggestions and validate each topic. Based on this, the company can find out the dominant topics discussed on social media, especially on Twitter. Topic modeling is expected to provide information about topic discussed in the public. Therefore, it can provide recommendations for feedback to the organization.

Table 6. Keywords related each topic

Topic No	Keywords	Keywords translation in English	Topic name
1	*Wilayah, menggandeng, penanaman, pohon, mangrove, sobat, korupsi, lowongan, perusahaan, pembangunan, kontribusi, nelayan, merawat, masyarakat, air, rangka, saham, infrastruktur, anggota*	Region, cooperate, planting, tree, mangrove, partner, corruption, vacancy, company, development, contribution, fisherman, care, community, water, order, stock, infrastructure, member	Corporate social responsibility and career development
2	*Jembatan, dihelat, peresmian, meresmikan, proyek, lowongan, pemerintah, program, bumn, konstruksi, pembangunan, kontrak, acara, tanggung, social, perusahaan, lingkungan, dibangun, modal*	Bridge, held, opening, inauguration, project, vacancy, government, program, bumn, construction, development, contract, event, responsibility, social, company, environment, developed, asset	Infrastructure development and government programs
3	*Proyek, selesai, bangunan, rumah, anggaran, tolong, sekolah, perhatian, korban, terminal, penumpang, khusus, pelabuhan, kuli, dilunasi, jalan, anak, perusahaan, bangun, tol, usaha*	Project, finished, building, residential, budget, favor, school, concern, victim, terminal, passenger, special, harbor, laborer, repaid, road, subsidiary, company, development, toll, venture	Effective project management and stakeholder satisfaction in building projects
4	*Kegiatan, upaya, perekonomian, bentuk, kawasan, pembangunan, warga, dampak, BUMN, perubahan, antisipasi, pertumbuhan, adaptasi, iklim, perusahaan, anggota, kerja, budaya, program, pelatihan, penanaman, mangrove, membangun, Indonesia*	Activity, effort, economy, form, region, development, community, impact, bumn, change, anticipation, growth, adaptation, climate, company, member, work, culture, program, training, planting, mangrove, building, Indonesia	Human resource management and accountability for sustainable growth



(a) (b)

(c) (d)

Figure 3. Distribution of coherence score; (a) topic 1, (b) topic 2, (c) topic 3, and (d) topic 4

For this evaluation, coherence scores were used, as shown in Table 5. A coherence score is used to measure how semantically coherent the main words in a topic are. The c_v coherence score is used in this research. The findings of this study got a coherence score of 0.558, and to see how well the model has been created, a direct human judgment of each sample topic based on the coherence score distribution as shown in Figure 3 will be conducted to determine whether each document that has similar content is indeed on the same topic. The result is that the model got an average of 80% correctly. Sample sentences taken in each cluster were reviewed by the reviewers, and it was concluded that some incorrect predictions almost occurred when the coherence score was 0.7. For example, a sentence with a coherence score in topic cluster 1, which is about facility development, is more appropriately in cluster 2, which is about infrastructure development and government programs. Therefore, it would be better if each sentence had a coherence score of 0.8 or higher.

5. CONCLUSION

This study conducted topic modeling based on corporate social media content with the keyword construction company name on Twitter. The application of LDA in this case study has provided valuable insights into the underlying topics and themes present in the dataset on the social media content. This finding can categorize the content into 4 topics, which are Corporate social responsibility and career development, Infrastructure development and government programs, Effective project management and stakeholder satisfaction in building projects, Human resource management and accountability for sustainable growth. The finding provides information about specific terms regarding this case, such as *’proyek’*, ‘*bangunan’*, ‘*anggaran’*, ‘*lowongan’* and other terms. Thus, from these terms it was possible to draw conclusions about the topic. These findings provide information to companies about the topics or opinions that are developing in the community, which can provide recommendations regarding decision-making on current issues in order to increase trust towards the company. However, as technology and customer demands evolve, in this study the dataset was obtained only from Twitter, which may not fully represent the entire spectrum of customer opinions and sentiments in this construction domain. The exclusion of data from other platforms or sources may have led to a partial understanding of the broader landscape. In addition, data inaccuracy was another challenge that the researcher faced during the analysis. Due to a mistake in manually inputting slang words, there were some top keywords that were conjunctions and thus would affect the topic labeling. Further improvements in retrieval from various social media sources, the normalization process for inputting slang words, and integrating LDA with real-time analytics can enable rapid responses to emerging topics and issues in the construction landscape.

REFERENCES

[1] N. A. Vidya, M. I. Fanany, and I. Budi, “Twitter sentiment to analyze net brand reputation of mobile phone providers,” *Procedia Computer Science*, vol. 72, pp. 519–526, 2015, doi: 10.1016/j.procs.2015.12.159.

[2] Z. Drus and H. Khalid, “Sentiment analysis in social media and its application: systematic literature review,” *Procedia Computer Science*, vol. 161, pp. 707–714, 2019, doi: 10.1016/j.procs.2019.11.174.

[3] Z. Li, Q. Zhang, X. Du, Y. Ma, and S. Wang, “Social media rumor refutation effectiveness: evaluation, modelling and enhancement,” *Information Processing and Management*, vol. 58, no. 1, p. 102420, Jan. 2021, doi: 10.1016/j.ipm.2020.102420.

[4] A. Ioanid and C. Scarlat, “Factors influencing social networks use for business: Twitter and YouTube analysis,” *Procedia Engineering*, vol. 181, pp. 977–983, 2017, doi: 10.1016/j.proeng.2017.02.496.

[5] F. de O. Santini, W. J. Ladeira, D. C. Pinto, M. M. Herter, C. H. Sampaio, and B. J. Babin, “Customer engagement in social media: a framework and meta-analysis,” *Journal of the Academy of Marketing Science*, vol. 48, no. 6, pp. 1211–1228, Nov. 2020, doi: 10.1007/s11747-020-00731-5.

[6] W. M. Lim and T. Rasul, “Customer engagement and social media: Revisiting the past to inform the future,” *Journal of Business Research*, vol. 148, pp. 325–342, Sep. 2022, doi: 10.1016/j.jbusres.2022.04.068.

[7] A. Meddeb and L. B. Romdhane, “Using topic modeling and word embedding for topic extraction in Twitter,” *Procedia Computer Science*, vol. 207, pp. 790–799, 2022, doi: 10.1016/j.procs.2022.09.134.

[8] S. Rani and M. Kumar, “Topic modeling and its applications in materials science and engineering,” *Materials Today: Proceedings*, vol. 45, pp. 5591–5596, 2021, doi: 10.1016/j.matpr.2021.02.313.

[9] I. Vayansky and S. A. P. Kumar, “A review of topic modeling methods,” *Information Systems*, vol. 94, p. 101582, Dec. 2020, doi: 10.1016/j.is.2020.101582.

[10] S. H. Mohammed and S. Al-Augby, “LSA & LDA topic modeling classification: comparison study on e-books,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 19, no. 1, p. 353, Jul. 2020, doi: 10.11591/ijeecs.v19.i1.pp353-362.

[11] J. Ahn, H. Son, and A. D. Chung, “Understanding public engagement on twitter using topic modeling: the 2019 ridgecrest earthquake case,” *International Journal of Information Management Data Insights*, vol. 1, no. 2, p. 100033, 2021, doi: 10.1016/j.jjimei.2021.100033.

[12] M. Nuser and E. Al-Horani, “Medical documents classification using topic modeling,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 17, no. 3, p. 1524, Mar. 2020, doi: 10.11591/ijeecs.v17.i3.pp1524-1530.

[13] N. S. M. N. Mangsor, S. A. M. Nasir, W. F. W. Yaacob, Z. Ismail, and S. A. Rahman, “Analysing corporate social responsibility reports using document clustering and topic modeling techniques,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 26, no. 3, pp. 1546–1555, Jun. 2022, doi: 10.11591/ijeecs.v26.i3.pp1546-1555.

[14] L. Hagen, “Content analysis of e-petitions with topic modeling: how to train and evaluate LDA models?,” *Information Processing & Management*, vol. 54, no. 6, pp. 1292–1307, Nov. 2018, doi: 10.1016/j.ipm.2018.05.006.

[15] Z. Tong and H. Zhang, “A text mining research based on LDA topic modelling,” in *Computer Science & Information Technology (CS & IT)*, May 2016, pp. 201–210, doi: 10.5121/csit.2016.60616.

[16] G. S. N. Murthy, S. R. Allu, B. Andhavarapu, and M. B. M. Bagadi, “Text based sentiment analysis using LSTM,” *International Journal of Engineering Research and*, vol. V9, no. 05, 2020, doi: 10.17577/ijertv9is050290.

[17] N. A. Salsabila, Y. A. Winatmoko, A. A. Septiandri, and A. Jamal, “Colloquial Indonesian Lexicon,” in *Proceedings of the 2018 International Conference on Asian Language Processing, IALP 2018*, Nov. 2019, pp. 226–229, doi: 10.1109/IALP.2018.8629151.

[18] A. T. Aw, M. Zhang, J. Xiao, and J. Su, “A phrase-based statistical model for SMS text normalization,” in *COLING/ACL 2006 - 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Main Conference Poster Sessions*, 2006, no. July, pp. 33–40, doi: 10.3115/1273073.1273078.

[19] R. K. Gupta, R. Agarwalla, B. H. Naik, J. R. Evuri, A. Thapa, and T. D. Singh, “Prediction of research trends using LDA based topic modeling,” *Global Transitions Proceedings*, vol. 3, no. 1, pp. 298–304, Jun. 2022, doi: 10.1016/j.gltp.2022.03.015.

[20] D. Maier *et al.*, “Applying LDA topic modeling in communication research: toward a valid and reliable methodology,” *Communication Methods and Measures*, vol. 12, no. 2–3, pp. 93–118, 2018, doi: 10.1080/19312458.2018.1430754.

[21] C. C. Silva, M. Galster, and F. Gilson, “Topic modeling in software engineering research,” *Empirical Software Engineering*, vol. 26, no. 6, 2021, doi: 10.1007/s10664-021-10026-0.

[22] D. M. Blei, A. Y. Ng, and M. T. Jordan, “Latent dirichlet allocation,” *Advances in Neural Information Processing Systems*, vol. 3, no. 3/1/2003, pp. 993–1022, 2002, doi: 10.5555/944919.944937.

[23] S. Koltcov and V. Ignatenko, “Renormalization analysis of topic models,” *Entropy*, vol. 22, no. 5, p. 556, 2020, doi: 10.3390/E22050556.

[24] D. L. M. Owa, “Identification of topics from scientific papers through topic modeling,” *Open Journal of Applied Sciences*, vol. 10, no. 04, pp. 541–548, 2021, doi: 10.4236/ojapps.2021.104038.

[25] Y. Kalepalli, S. Tasneem, P. D. P. Teja, and S. Manne, “Effective comparison of LDA with LSA for topic modelling,” in *Proceedings of the International Conference on Intelligent Computing and Control Systems, ICICCS 2020*, May 2020, pp. 1245–1250, doi: 10.1109/ICICCS48265.2020.9120888.

[26] S. Syed and M. Spruit, “Full-text or abstract? examining topic coherence scores using latent dirichlet allocation,” *Proceedings - 2017 International Conference on Data Science and Advanced Analytics, DSAA 2017*, vol. 2018-January, pp. 165–174, 2017, doi: 10.1109/DSAA.2017.61.

[27] S. Mifrah, “Topic modeling coherence: a comparative study between LDA and NMF models using COVID’19 corpus,” *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 4, pp. 5756–5761, Aug. 2020, doi: 10.30534/ijatcse/2020/231942020.

[28] M. Röder, A. Both, and A. Hinneburg, “Exploring the space of topic coherence measures,” in *WSDM 2015-Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, Feb. 2015, pp. 399–408, doi: 10.1145/2684822.2685324.

BIOGRAPHIES OF AUTHORS

Izmi Dewi Aisha is a Master’s student at BINUS Graduate Program-Master of Computer Science, Bina Nusantara University with a focus on data science. She holds a bachelor's degree in informatics engineering from Universitas Islam Negeri Sunan Gunung Djati, Bandung, Indonesia. She can be contacted at email: izmi.aisha@binus.ac.id.

Lili Ayu Wulandhari is a Lecturer at BINUS Graduate Program, Master of Computer Science, Bina Nusantara University. Her research area is machine learning for computer vision and natural language processing. She can be contacted at email: lili.wulandhari@binus.ac.id.