Journal of Applied Intelligent System (e-ISSN : 2502-9401 | p-ISSN : 2503-0493)
Vol. 6 No. 1, 2021, pp. 23 – 31

Keyphrase Extraction on Covid-19 Tweets Based on Doc2Vec and YAKE

Fahri Firdausillah*[1], Erika Devi Udayanti[2]
Universitas Dian Nuswantoro, Jl. Imam Bonjol No 207, Semarang, (024) 3517261
E-mail : fahri@dsn.dinus.ac.id[1], erika.devi.udayanti@dsn.dinus.ac.id[2]*

Abstract - Keyword and keyphrase extraction are one of the initial foundations for performing several text processing operations such as summarization and document clustering. YAKE is one of the techniques used for unsupervised and independent keyphrase extraction, it does not require a corpus for linguistic tools such as NER and POS-tag. However, the use of YAKE in microblogging documents such as Twitter often results in a keyphrase that is less representative because of the lack of words used for ranking. This paper offers a solution to this problem by looking for similar tweets in the keyphrase extraction process using Doc2Vec so that the number of words used in the YAKE ranking process can be greater. Covid-19 tweets related are used as dataset as the topic is currently widely discussed on social media to prove that the proposed approach could improve keyphrase extraction performance.

Keywords – YAKE, Doc2Vec, Covid-19, Keyphrase Extraction, Twitter

1. INTRODUCTION

Keyword or keyphrase extraction (KE) is an important process to obtain main information contained in one or several documents. Many studies on keyword extraction were conducted either using supervised approach (based on machine learning) such as [1] or unsupervised approach (based on statistics and ranking) as in [2]. One of the KE methods that uses an unsupervised approach is YAKE [3] which has the advantage of being able to perform KE on a document independently, which means it does not require a corpus to form language tools (such as NER and POS-tag). YAKE's ability to do KE on documents such as papers or internet news has proven to be quite good, as stated by [4].

Social media posts are one kind of online document that is widely available for free on the internet and have a quite large quantity, especially during the Covid-19 pandemic, the use of social media increased quite significantly [5]. The very large quantity of social media documents that are free to use provides a reason for researchers to process these documents as data sources so that the information and knowledge contained in them can be extracted. However, mining for insights on social media is not as easy as mining into news documents.

KE techniques that can be used well in news documents and papers, often have different results when implemented in short documents such as posts on microblogging (one of social media form). KE on microblogging documents such as Twitter has its own challenges due to its special characteristics, namely the text size which is much smaller than regular documents such as news or articles. Some main keyphrases were often not produced in keyphrase extraction result by using YAKE on several tweet documents due to lack of word quantity, despite only keyphrase that is mentioned several times were produced. Further discussion in this issue will be carried out in following sub section.

The proposed solution to this problem is to combine several tweets that are similar to the tweet that is being searched for the key phrase, before KE process is carried out using YAKE. A similar search for tweets is done using Doc2Vec which can convert a document into a vector collection in vector space so that the distance with other documents can be measured based on similarity.

The data used in this study is tweet data on the topic of Covid-19 and its relationship with government policies for handling the pandemic. The topic is chosen since it is one of the topics that are currently the most discussed in the meantime. During 2019-2020 timeframe, many papers are published that discuss social media analysis on Covid-19 and several topics related to pandemic handling, as such discussed in [6] and [7].

This paper will also show the evaluation results of the proposed solution using 3 assessment matrices, namely accuracy, recall, and F-Score. This paper uses YAKE as the baseline for the assessment which then compares the extraction results with the proposed approach. The KE process using these two methods will be executed on several different tweets which then an average value is taken to determine whether there is an increase in the performance of the proposed method.

1.1. Covid-19 Twitter Analysis

Social media is one of the largest corpus available in the internet, every day social media users around the world post something related to their activities and life, one of which is about Covid-19. During the pandemic (one of the effects of the lock down), the use of social media has increased rapidly, one of the topics that is always discussed is Covid-19, its impact and how the government handle it. A lot of research has been done to mine social media posts in connection with the Covid-19 pandemic, with one of the most discussed topics is sentiment analysis to find out people's perceptions regarding policies around Covid-19 in each country. The research as was done by [6], uses a support vector machine to find out what Twitter users think about the government's ability to deal with a pandemic. In addition, there are also studies related to the detection of misinformation around covid-19 as conducted by [8].

Several studies studied for this paper use KE in their processing pipelines. This shows that KE is one of the key processes in social media mining process for various purposes including topic modeling, document summarization, and document indexing [9].

1.2. Keyword and Keyphrase Extraction

Keyphrase and keyword extraction are the same process with a slight difference in the amount of n-grams to be extracted. Although the purpose of both is the same, that is to get certain words or phrases that represent a text data, the results of both will be different. Keyword extraction is only used to extract unigrams in a text data, while keyphrase is used to get more than one word (bigram, trigram, etc.) [10]. The difference in the extraction results on keywords and keyphrases can be seen in Table 1.

Table 1. The difference between keyword and keyphrase extraction

Sentence	Keyword Extraction	Keyphrase Extraction
Indonesia kemungkinan membutuhkan waktu lebih dari 10 tahun untuk memvaksinasi populasi mereka jika mereka melanjutkan vaksinasi dengan kecepatan saat ini	"Indonesia", "membutuhkan", "waktu", "populasi", "melanjutkan", "vaksinasi", "kecepatan"	"Indonesia", "membutuhkan waktu", "melanjutkan vaksinasi", "kecepatan saat ini"
Jika hanya dilakukan di akhir pekan, ia mempertanyakan peningkatan mobilitas di hari biasa yang bisa terus mencatatkan penambahan kasus COVID-19	"dilakukan", "akhir", "pekan", "mempertanyakan", "mobilitas", "hari", "biasa", "mencatatkan", "penambahan", "kasus", "COVID-19"	"hanya dilakukan", "akhir pekan", "mempertanyakan", "peningkatan mobilitas", "hari biasa", "penambahan kasus", "COVID-19"

Keyword extraction is easier to perform than keyphrase extraction, but keyphrase extraction represents the meaning of a text data better than keyword extraction. KE can be used for several purposes such as categorizing documents into specific groups to make them easier to find, summarizing one or several documents into fewer sentences without losing the core idea of the document, as well as searching for semantic similarity of content in multiple documents. [2]. This can also imply that KE is one of the most important initial phases in processing text data.

In general, there are three approaches used to conduct KE that are supervised learning, unsupervised learning, and deep learning [9]. Supervised learning and deep learning for KE requires training data to study the pattern first, then the pattern is implemented in the text data that you want to process. Some of the algorithms that can be used for KE with this approach are Support Vector Machine (SVM), Bayesian Network, Maximum Entrotopy, and Linear Regression.

The unsupervised learning approach in KE uses a ranking method with statistical calculations or a graph algorithm to determine the weight of the importance of a phrase in a document. Unsupervised algorithms commonly used for KE include Graph based ranking, Topic-based, clustering, Language models, and probabilistic models. The advantage of this approach is that there is no need for training data to use it.

1.3. Keyword Extraction for Tweets

KE on microblogging such as Twitter has new challenges in the form of using unstructured sentences, not following grammatical rules, and containing many words that are out of vocabulary (OOV). Several studies such as [11] and [12] have tried to normalize the OOV word for sentiment analysis purposes. Although the initial assumption of OOV in the sentiment analysis will produce noise at the feature extraction stage, the results obtained after the normalization are not significant for the classification of sentiment analysis. As for the unsupervised KE process, OOV in many documents also does not cause significant problems because it will automatically be eliminated during the feature extraction and ranking stages. However the KE in single document (single tweet) will make OOV appear in the KE results because of the few number of words in the document (tweet). This difference is shown in table 2, where KE in several tweet documents using YAKE produces some OOV keywords.

Table 2. OOV on tweets detected as a key phrase

Tweet	Detected Keyphrase
Ini pandemi global..negara2 maju sja kwalahan mnghadapi pandemi ini...Pemerintah tdk salah ..yg salah itu rakyat yg masih membangkang...Klw kita mau slmat dri virus ini bukan berhrap dri kebijakan pemerintah tpi hrus mulai sadar utk jalani protokol kesehatan penanganan Covid	penanganan covid, maju **sja**, pandemi global, pemerintah **tdk**, kebijakan pemerintah
Prediksi gue hari ini : Abis lebaran orang - orang masa bodo sama covid-19 dan pemerintah juga sudah "ya udahlah deh"	**abis** lebaran, prediksi **gue**, **udahlah deh**, **gue** hari, masa **bodo**
Wah gila. Ini sih keren bgt pemerintahnya. Tp pemerintah indonesia jg keren, dgn segala keterbatasan bisa mengendalikan penderita dan penyebaran covid, sehingga jumlah korban jiwa ga seekstrim negara2 lain.Semua negara hebat dgn kapasitasnya masing2.	gila, **bgt** pemerintahnya, keren, **dgn**, pemerintahnya
Thread tentang corona Denger karena Lockdown dan PSBB ngga berhasil dilakukan di Indonesia maka bakalan kearah *Herd Imunity*.Setelah Idul Fitri Pemerintah tidak akan update jumlah yang tekena Covid.	**herd imunity**, idul fitri, fitri pemerintah, tekena covid, corona **denger**

Some of OOVs detected in tweets are commonly, used and are still eligible to be candidates as key phrases such as the words "prediksi gue", "masa bodo", and "abis lebaran". However, some are not suitable as a keyword because the positioning is not quite right but not

widely used, such as "maju sja" and "corona denger". The detection of the commonly used OOVs depends on how much OOV is used in several tweets.

1.4. Yet Another Keyword Extractor (YAKE)

YAKE is an unsupervised keyword extraction algorithm, meaning that this algorithm does not require many documents to be used as a reference corpus. The advantage of YAKE is that it can calculate statistical relation to words in a document without requiring supporting documents such as wordnet or wikipedia [13]. In addition, YAKE is language independent, so it is very suitable for use in languages that do not yet have many linguistic tools (NER and POS tager) such as Indonesian. YAKE uses 5 text processing steps:

1. Text Preprocessing: to break the text into several sentences based on the punctuation found, then each sentence is broken down into several tokens which will then be tagged based on the position of each word.
2. Feature Extraction: to calculate statistics for each word using TCase (statistics based on the letter case of a word and their position), TPos (calculating the level of importance of words based on their position in a sentence), TFNorm (calculating the frequency of word usage in a sentence), TRel (counting the significance of a word according to its context), and TSent (calculating the frequency of occurrence of terms in several existing sentences).
3. Computing Term Score: next step is computing the value for features using formula depicted in (1) where the value of TRel, TPos, TCase, TFNorm, and TSentence are gotten from feature extraction correspondingly.

$$S(t) = \frac{T_{Rel} * T_{Pos}}{T_{Case} + \frac{TF_{Norm}}{T_{Rel}} + \frac{T_{Sentence}}{T_{Rel}}} \quad (1)$$

4. N-Gram generation: the desired number of N-grams can be determined so that the results can be adjusted according to specific needs. If N is more than 1, then the longer phrase has a higher weight than the lower phrase, so the phrase "government regulation" has a higher weight than the individual words "regulation" and "government" separately. The equation for N-Gram generation is shown in (2).

$$S(kw) = \frac{\Pi_{t\varepsilon kw} S(t)}{KF(kw) * (1 + \sum_{t\varepsilon kw} S(t))} \quad (2)$$

5. Data deduplication and ranking: in the final stage all the keywords that have been obtained are recalculated and compared to eliminate redundancy and also ranked as needed.

In several comparisons that have been made, YAKE is one of algorithms that has high accuracy compared to other unsupervised keyword extraction algorithms [4]. However, using YAKE for keyword extraction on microblogging such as Twitter will experience difficulties because the number of words per document is very limited. As a result, many irrelevant words, OOV words, are included as keywords.

1.5. Doc2Vec for Specifying Document Similarity

Although YAKE provides good keyphrase results for single document extraction, due to the limited number of words in a tweet, YAKE often provides keywords that are less relevant to the tweet. as an example of a tweet:

> *"Seharusnya saat ini Pemerintah Lampung bisa memberikan bantuan kepada mahasiswa tersebut. Mengingat begitu derasnya donasi yang masuk ke Pemprov Lampung melalui tim gugus tugas COVID-19"*

When KE was carried out using YAKE with a maximum of 2-grams, we managed to get the key phrase "pemerintah lampung", "pemprov lampung", and "memberikan bantuan" because of the repetition and positioning of the phrase, but failed to get the key phrase "mahasiswa" , "gugus tugas", and "donasi" which is only mentioned once in the sentence but should be one of the core phrases that can be used for document grouping.

The limited number of characters in each tweet causes the key phrases generated from YAKE to often include phrases that are less representative contextually. To reduce this possibility in this study a search for similar tweets was added to increase the number of sentences to be processed. Based on research [14], one technique that can be used to find similar documents with high accuracy is Doc2Vec. In addition to having a fairly good accuracy in obtaining document similarities, Doc2Vec can also be used on non-English documents without having to use an external corpus as was done by [15] to find similar Filipino documents.

Doc2Vec is a document representation algorithm which is an extension of the numeric representation of Word2Vec words by adding an ID representation to each document, so that each document will have a vector representation for each word contained as well as a unique document representation. The vectorization results of these documents enable us to get similar documents by using cosine similarity to each vector [16].

2. RESEARCH METHOD

2.1. Dataset Covid-19 Tweets

The dataset used for this study was taken from Kaggle [17]. There were 52959 tweets in Indonesian that were taken between April - July 2020 using the keywords “Covid-19”, “Corona”, and the “Pemerintah”. Before it can be used, data cleaning is done by removing URLs, mentions, and hashtags. As a result of data cleaning, it is known that there are some tweets that only contain URLs and mentions, so that after this process the text size of the tweets becomes 0. In addition, there are also tweets containing a small number of words as shown in table 3.

Table 3. Examples of short Tweets that were omitted

Original Tweet	After cleaning and stopword removal
https://www.antaranews.com/berita/1549664/pemerintah-diminta-perketat-perbatasan-cegah-gelombang-kedua-covid\xa0…\npemerintah-diminta-perketat-perbatasan-cegah-gelombang-kedua-covid\n#DisiplinCegahGelombangKe2'	[empty]
mungkin, pemerintah +62 to covid: https://twitter.com/PoemHeaven/status/1265598800690843650\xa0…	mungkin pemerintah to covid
Alasan Pemerintah Kasus Baru Covid-19 Terus Tinggi https://medcom.id/s/GKdOyV4k\xa0	alasan pemerintah kasus covid-19
Dukung Upaya Pemerintah Memutus Penyebaran Covid-19\n#DukungPemerintah #IndonesiaMaju #BersamaLawanCorona pic.twitter.com/XBqbv8RRcl	dukung upaya pemerintah memutus penyebaran covid-19

Tweets that have few words potentially causing noise during KE process since they have no apparent meaning, so it was decided to delete data that had tweets less than 20 words. After filtering these short tweets, finally we got 28592 tweet data used in the process of making models with Doc2Vec. For the evaluation purpose, only 100 tweets were checked for the keyphrase extraction results and the confusion matrix was calculated.

After obtaining the dataset, the keyphrase extraction process is then carried out as shown in Figure 1. In general there are two major processes that must be carried out, first Doc2Vec modeling for document vectorization to find closest documents, and second keyword extraction with YAKE. Tweets that are included in the vectorization process with Doc2Vec must be cleaned, stopword removal, and lemmatize first to reduce their possible features, meanwhile in KE process, original tweets that have been cleaned is used, without stopword removal and lemmatization.

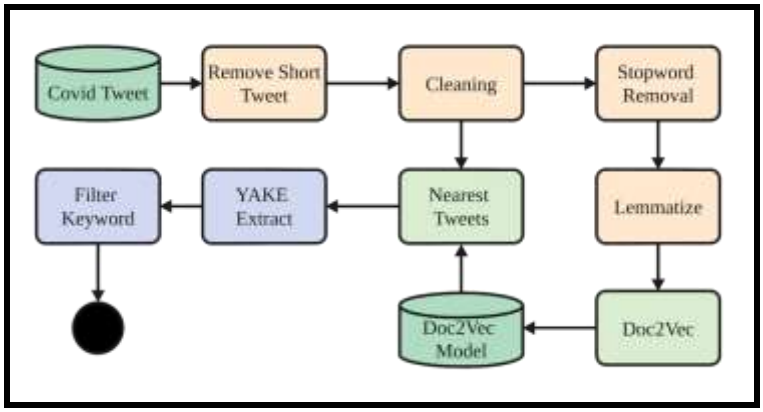

Figure 1. KE process flow with Doc2Vec and YAKE.

2.2. Preprocessing tweet data

Dataset that has been cleaned from URLs, mentions, hashtags, and short tweets are processed using the sequence shown in Figure 1.

Preprocessing used in Doc2Vec process includes cleaning tweets from URLs, mentions, and hashtags, equalizing case letters, eliminating stopwords, lemmatizing to get word base form, and tokenization. This is done so that the document grouping is more semantically representative for each tweet, not just a sequence of words. On the other hand preprocessing used at the keyword extraction stage is only cleaning tweets from URLs, mentions, and hashtags, since the punctuation sequence of words and even stopwords is needed in YAKE. It need to be noted that all particular tweet that need to be extracted and other similar tweet that will be grouped in process is not processed using stopword removal and lemmatization.

2.3. Keyphrase Extraction

The KE process starts from searching for tweets that are similar to the tweet to be extracted, the goal is to remove the word OOV and improve ranking performance for each phrase. The optimal number of similar tweets for this purpose ranges from 3 - 10 tweets, by manual observation its known that every tweets may has their own optimal number of similar tweets to associate. To simplify the result in this paper we try three possibilities of similar tweet number which are 3, 5, and 7 similar tweet as it's the most frequent optimal number in our observation this far. Next step, the similar tweets will be combined into one document for extraction using YAKE with a maximum parameter of 2 N-Gram.

The results from the KE in grouped tweets showed that there were many phrases that were not found in the tweets that were searched, so that after the KE process it is necessary to filter only the phrases found in the particular document that is being extracted are used.

2.4. Evaluation of Kephrase Extraction Results

The results of KE is usually evaluated using a three evaluation matrix [9] that are Accuracy (P) as shown in (3), Recall (R) as shown in (4), and F-Score (F) as shown in (5). In this study, 100 tweets have been carried out manually by the annotator, then KE from the manual extraction results is compared with KE obtained by automatic processing using YAKE as the baseline and YAKE with Doc2Vec for grouping similar tweets. The equation used to obtain the three measurement matrices is shown below.

$$P = \frac{The number of keywords extracted correctly}{The number of all keywords extracted} \quad (3)$$

$$R = \frac{The number of keywords extracted correctly}{The number of standard keywords} \quad (4)$$

$$F - measure = \frac{2 * P * R}{P + R} \quad (5)$$

3. RESULTS AND DISCUSSION

In this study, several trials were carried out by applying different parameters to the number of similar tweet retrievals, which are 3, 5, and 7 similar tweets to be grouped. Comparison sample of the results obtained from each test result are shown in table 4.

Table 4. Comparison sample of KE results using YAKE and Doc2Vec

clean	yake	yake_doc_3	yake_doc_5	yake_doc_7
Menteri Sosial (Mensos) Juliari Batubara menemukan kejanggalan pembagian bantuan sosial (bansos) yang dilakukan Pemerintah Provinsi DKI Jakarta selama pandemi virus corona (Covid-19).	juliari batubara, pemerintah provinsi, provinsi dki, dki jakarta, menteri sosial	bantuan sosial, virus corona, sosial, menemukan kejanggalan, kejanggalan pembagian,	bantuan sosial, pandemi, sosial, pandemi virus, virus corona	pandemi, bantuan sosial, sosial, pandemi virus, virus corona
Anggota Polsek Tempeh, AIPDA DHIMAS ADJI WALUYO, S.H. melaksanakan Patroli Kewilayahan dan bertatap muka dengan Anggota SKD Gesang dalam rangka berikan himbauan agar selalu patuhi protokol kesehatan dari Pemerintah untuk cegah penyebaran Virus Covid-19	polsek tempeh, aipda dhimas, adji waluyo, dhimas adji, patroli kewilayahan	protokol kesehatan, patuhi protokol, cegah penyebaran, berikan himbauan, bertatap muka	protokol kesehatan, patuhi protokol, cegah penyebaran, berikan himbauan, untuk cegah	protokol kesehatan, patuhi protokol, cegah penyebaran, berikan himbauan, protokol
…	…	…	…	…
Update angka akumulatif perkembangan data kasus Covid-19 dari seluruh Indonesia sampai dengan Jumat, Juli , yang disampaikan oleh Juru Bicara Pemerintah Penanganan Covid-19, dr. Achmad Yuriantodi Graha BNPB.	pemerintah penanganan, juru bicara, bicara pemerintah, update angka, data kasus	data kasus, angka akumulatif, akumulatif perkembangan, perkembangan data, kasus	data kasus, angka akumulatif, akumulatif perkembangan, perkembangan data, kasus	data kasus, angka akumulatif, akumulatif perkembangan, perkembangan data, kasus

Then the results of the experiment were compared with the extraction results using YAKE without Doc2Vec as a baseline, and then manual extraction from is also used as the gold standard of KE. After obtaining the value of Accuracy, Recall, and F-score for each tweet, the

average value is then obtained. The results of average performance calculations are shown in table 5.

Table 5. Average KE performance calculation of 100 tweets

	Acuracy	Recall	F-Score
YAKE	0.415	0.305	0.349
3 similar tweet + YAKE	0.594	0.446	0.507
5 similar tweet + YAKE	**0.676**	**0.510**	**0.578**
7 similar tweet + YAKE	0.615	0.471	0.531

The results shown in the table show that using Doc2Vec to group 5 similar tweets before doing KE with YAKE gave the best results on the three matrices used. These table also shows that combining 3, 5, and 7 similar tweets before the extraction process generally gives better results than using only the YAKE algorithm.

4. CONCLUSION

YAKE is language independent unsupervised algorithm to extract keyword or keyphrase from a single document. YAKE can be used for KE in Indonesian language without the help of a language tool such as NER and POS-tag. However, it cannot be used optimally in documents with a small number of words such as in tweet documents, there are often main keyphrases that do not appear in the extraction results because they are only mentioned once in a tweet. As a solution, in this research, Doc2Vec is used to search for similar tweets so that they can be joined and extracted as a single document.

Searching for similar tweets with Doc2Vec on YAKE has succeeded in improving the F-Score performance by 15% to 23% compared to using only YAKE. However, the selection of the number of similar tweets based on the vector distance is quite influencing the results of the KE itself, the addition of similar tweets does not always increase accuracy, as it can be seen that the use of 7 similar tweets has a lower level of accuracy than 5 similar tweets. In that case, further research is needed to determine the optimal number of similar documents automatically for documents with different characteristics.

REFERENCES

[1] Haddoud M, Mokhtari A, Lecroq T, Abdeddaïm S. Accurate Keyphrase Extraction from Scientific Papers by Mining Linguistic Information. CLBib@ ISSI. researchgate.net; 2015. pp. 12–17.

[2] Trisna INP, Nurwidyantoro A. Single document keywords extraction in Bahasa Indonesia using phrase chunking. TELKOMNIKA. 2020;18: 1917.

[3] Campos R, Mangaravite V, Pasquali A, Jorge AM, Nunes C, Jatowt A. YAKE! Collection-Independent Automatic Keyword Extractor. Advances in Information Retrieval. Springer International Publishing; 2018. pp. 806–810.

[4] Škrlj B, Repar A, Pollak S. RaKUn: Rank-based Keyword Extraction via Unsupervised Learning and Meta Vertex Aggregation. Statistical Language and Speech Processing. Springer International Publishing; 2019. pp. 311–323.

[5] Lamsal R. Design and analysis of a large-scale COVID-19 tweets dataset. Applied Intelligence. 2020. doi:10.1007/s10489-020-02029-z
[6] Prastyo PH, Sumi AS, Dian AW, Permanasari AE. Tweets Responding to the Indonesian Government's Handling of COVID-19: Sentiment Analysis Using SVM with Normalized Poly Kernel. Journal of Information Systems Engineering and Business Intelligence. 2020;6: 112–122.
[7] Boon-Itt S, Skunkan Y. Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study. JMIR Public Health Surveill. 2020;6: e21978.
[8] Sharma K, Seo S, Meng C, Rambhatla S, Dua A. Coronavirus on social media: Analyzing misinformation in Twitter conversations. arXiv preprint arXiv. 2020. Available: https://arxiv.org/abs/2003.12309
[9] Alami Merrouni Z, Frikh B, Ouhbi B. Automatic keyphrase extraction: a survey and trends. J Intell Inf Syst. 2020;54: 391–424.
[10] Siddiqi S, Sharan A. Keyword and keyphrase extraction techniques: a literature review. Int J Comput Appl Technol. 2015;109. Available: https://www.academia.edu/download/54323945/pxc3900607.pdf
[11] Hanafiah N, Kevin A, Sutanto C, Fiona, Arifin Y, Hartanto J. Text Normalization Algorithm on Twitter in Complaint Category. Procedia Comput Sci. 2017;116: 20–26.
[12] Arora M, Kansal V. Character level embedding with deep convolutional neural network for text normalization of unstructured data for Twitter sentiment analysis. Social Network Analysis and Mining. 2019;9: 12.
[13] Campos R, Mangaravite V, Pasquali A, Jorge A, Nunes C, Jatowt A. YAKE! Keyword extraction from single documents using multiple local features. Inf Sci . 2020;509: 257–289.
[14] Mandal A, Chaki R, Saha S, Ghosh K, Pal A, Ghosh S. Measuring Similarity among Legal Court Case Documents. Proceedings of the 10th Annual ACM India Compute Conference. New York, NY, USA: Association for Computing Machinery; 2017. pp. 1–9.
[15] Barco Ranera LT, Solano GA, Oco N. Retrieval of Semantically Similar Philippine Supreme Court Case Decisions using Doc2Vec. 2019 International Symposium on Multimedia and Communication Technology (ISMAC). ieeexplore.ieee.org; 2019. pp. 1–6.
[16] Li J, Huang G, Fan C, Sun Z, Zhu H. Key word extraction for short text via word2vec, doc2vec, and textrank. TURK J OF ELECTR ENG COMPUT SCI. 2019;27: 1794–1805.
[17] Hermansyah DD. COVID-19 Indonesian Tweets. 2020. Available: https://www.kaggle.com/dionisiusdh/covid19-indonesian-twitter-sentiment