**A Text Similarity Study : Big Data Analysis for Burmese Online News Media**

Ye Naung, Ye Myat Han, Naing Linn Aung

Department of computer science

**ABSTRACT**

Nowadays, the same news events are usually covered by the online news media. Finding out how similar these events' descriptions are would be interesting. It examines how an obvious news event is described by across different times and the news media. On the other hand, exploring a large amount of news is the primary challenge. So, in this paper, aims to investigate similarities in Burmese online news content. To do so, news content’s text is tokenized, extracted, saved, and analyzed using our system. In this paper, experiments were done on Burmese local news content which was stored in our database system and extracted by using TF-IDF feature extraction techniques extracted data is saved as a pickle. As a result, we explore the degree of similarity between many social media platform news articles that cover the same event at different times, and groups of news media share the same articles with little or no modifications at all.

**Keywords**— text similarity, feature extraction, TF-IDF, big data analysis

1. **INTRODUCTION**

One major issue is getting the quality of news and real information that is being shared on social media in recent an era. According to the use of various social media platforms, more than once fake news spreads misinformation to the public. In the digital age, understanding similar content and fake news uses content because of the difficulty of controlling online media. A phenomenon that is observed daily is the circulation of duplicate or nearly identical articles across online media platforms. Although it's straightforward to handle this issue of duplication through automated techniques, it's challenging for readers to discern the level of similarity between specific news articles. Frequently, minor modifications are made to an original article: information may be added or subtracted, the title might be changed, a different source acknowledged, the arrangement reorganized, and alternate images used [1].

The proposed system for checking the similarity of Burmese online news is illustrated in Figure 1. This paper focuses on analyzing news content in the Burmese language, which is considered a low-resource language in terms of datasets and processing tools. The system primarily researches local news articles rather than using an English dataset. In our proposed system, the first step involves collecting public news daily from social media platforms such as Facebook, YouTube, Telegram, and Twitter using our crawling system with a scraping method. This approach was presented at CCET2023 "Implementation of a Social Media Crawler Based on Microservices Architecture", the 6th International Conference on Computer and Communication Engineering Technology. In the second step, collected news is tokenized into words, and Burmese stop words and unusable characters are removed in the preprocessing stage. After this stage, processed texts are extracted using Term Frequency and Inverse Document Frequency (TF/IDF) and saved as features in pickle files. Simultaneously, the content that we want to check is also extracted as features in the same way. In the final stage, the similarity between contents is checked using the cosine similarity method. As a result, we explore the degree of similarity between contents that cover the same event at different times, and groups of news media share the same content with little or no modifications at all.

Save features as pickle file

Check similarity

Feature of new content

Similarity result

Preprocessing

Online media

Collected contents

New content

Tokenization

Removing stop words

Extracting features

Preprocessing

Extracting features

**Collecting Data**

**Preprocessing Data**

**Extracting features**

**Checking similarity**

Fig. 1 System for checking similarity of Burmese online contents

1. **COLLECTING OF BURMESE ONLINE NEWS**

This section explores the methodologies for collecting data from social media platforms, with a focus on Facebook, X, YouTube, and Telegram, as utilized in Myanmar. It aims to provide a systematic approach for research interested in analyzing social media trends, user behavior, and content dissemination within the context of Myanmar’s unique digital landscape.

Data extraction is a critical step in social media data collection. It involves retrieving relevant information from various platforms to understand user interactions, content popularity, and overall platform engagement. Here, we expand on the methods for extracting data from social media platforms, focusing on the use of platform-specific APIs and web scraping techniques.

1. **Utilizing Platform-Specific APIs**

Most social media stages offer APIs (Application Programming Interfacing) that permit to get to particular sets of information in an organized and mechanized way. These APIs are outlined to give a standardized way to ask and get information, guaranteeing consistency and unwavering quality in the information collection handle. APIs can collect real-time information, track particular measurements, and indeed mechanize activities such as posting substance to make a more data-driven encounter. This makes a difference spare time and assets whereas picking up more exact and up-to-date insights.

1. **Web Scraping Techniques**

When APIs are not available or do not provide the required data, web scraping can be employed as an alternative method. Web scraping involves programmatically navigating web pages and extracting the needed information. Data collection from social media is a multidisciplinary endeavor that requires technical expertise, ethical considerations, and a robust tech stack. By combining Python, Selenium, PostgreSQL, Kubernetes, and RabbitMQ, organizations can unlock valuable insights from the vast social media landscape.

In table.1 can see the data sample of different social media platforms by using APIs and web scraping techniques. The collected data is publicly available and legally acceptable.

Table.1. Collected data sample from different platforms

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Facebook** | Content text | Reaction count | Share count | Comment count | Comment text | View | Date Time |
| **YouTube** | Description | Like count | Comment count | Video Duration | Comment text | View | Date Time |
| **Telegram** | Content text | Reaction count | Comment count | Video Duration | Comment text | View | Date Time |
| **X (Twitter)** | Tweet | Like count | Retweet count | Reply count | Reply | View | Date Time |

In this research we used 4 platforms (Facebook, YouTube, Telegram, X), the duration from 2021 to 2024 April. The data volume is nearly 2 million content text.

1. **PREPROCESSING TEXTS**
2. **Tokenization**

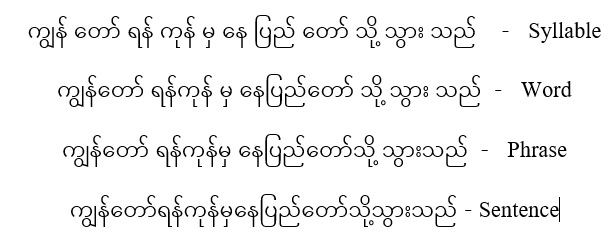
****The main difficulty encountered is so-called tokenization which segments text from sentence to word. When writing the Burmese language, can be written with white space and also written without white space. So, segmentation is not as easy as in the English language. If you build a sentence like “ကျွန်တော်ရန်ကုန်မှနေပြည်တော်သို့သွားသည်” which is translated to the English language like “I go from Yangon to Naypyidaw”, you have to start with the first syllable and a sentence will be formed through words and phrases. Figure (2) will be shown as an example of the step-by-step form of a sentence[2].

Fig. 2. The step by step form of a sentence

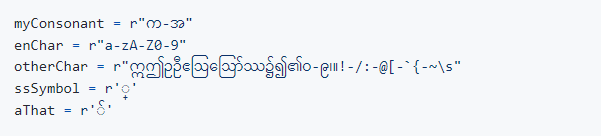
In our proposed system, we successfully tokenize a sentence to word such as “ကျွန်တော် ရန်ကုန် မှ နေပြည်တော် သို့ သွားသည်”. As shown in Figure (3), the consonants (က-အ), English letters, and English numbers (a-z, A-Z, 0-9); Other characters (vowels like “ဣ”, “ဤ”, “ဥ”, “ဦ”, Burmese numbers, some symbols); A total of five variables will be defined, including Unicode symbols and punctuation. The rule to break is if it's not a consonant followed by the symbol of subscripted letters in Pali. And if it's not a consonant that goes along with the syllable, make a breakpoint in front of that consonant letter. Or, if the characters are specified as other Char or English characters, it is specified that a break is added in front of those characters [3].

Fig. 3. Variables for segmentation rule

1. **Removing Stop words**

Another thing that should be done in the preprocessing section is removing stop words. Many languages have stop words. Stop words are common words, and the presence of these letters can increase the number of features and have more similarities when calculating as a value when generating features, which can lead to errors when using machine models. For example, “သည်”, “သို့”, and “မှ” are Burmese stop words.  In NLP (Natural Language Processing) technology, data cleaning or text cleaning is also important. If you don't do text cleaning, the data will be large and contain unnecessary interference (noise), and the accuracy of extraction features may decrease. To work properly, data processing must be done to reduce errors. Text cleaning can be considered as the process of removing unnecessary noise and organizing the language to make it easier to understand. It doesn't mean that always have to do text cleaning. If it should not be done according to the field. Unnecessary data (noise) is as follows:

• Having too many typos or misspellings

• Having many numbers and punctuations

• Having many emojis, usernames and links

• Containing lots of contractions and hyphens (-).

• Repetition of letters.

In the preprocessing stage, contents are tokenized by using the tokenization method and removed stop words by using Burmese stop words. Text preprocessing typical example is shown in figure (4).

Word segmentation

ကျွန်တော်ရန်ကုန်မှနေပြည်တော်သို့သွားသည်

Sentence

ကျွန်တော် ရန်ကုန် မှ နေပြည်တော် သို့ သွား သည်

Word

ကျွန်တော် ရန်ကုန် နေပြည်တော် သွား

Stop words removing

Processed Text

I Yangon Naypyidaw go

Fig. 4. Text preprocessing typical example

1. **EXTRACTING FEATURES**

After the text preprocessing is completed, the method known as TF-IDF (Term Frequency-Inverse Document Frequency) is widely used to extract features from text data. The process involves converting texts into numerical vectors that indicate the value of each word in a document about the corpus as an entire unit. To determine a word's frequency within a given document, the term frequency (TF) of each word in the document is calculated. A word's rarity over the full corpus is also calculated using the inverse document frequency (IDF). In this process, words that are used frequently in a text but rarely within the corpus are given higher TF-IDF scores to indicate their importance within that particular document. On the other hand, words that are often used with low TF-IDF scores are given less importance [4]. In this paper, prepared texts are extracted with features by using Term Frequency and Inverse Document Frequency (TF/IDF), and the maximum number of features is 3000 by using Bi-gram for the proposed dataset. Below (1), (2), and (3) equations denote the TF-IDF vectorization[4].

TF-IDF (w,d,D)=TF(w,d)\*IDF(w,D) (1)

Tf(w,d)=((occurences of w in document d))/((total number of words in document d) ) (2)

IDF (w,D)=((ln (Total number of documents (N)in corpus D))/((number of documents containing w) ) (3)

In this section, we use scikit-learn library for TF-IDF feature extraction and maximum features is 3000.

1. **IMPLEMENTATION OF THE PROPOSED SYSTEM AND RESULTS**

Finding similarity between words is a crucial portion of content similarity which is at that point utilized as an essential arrange for sentence, section, and archive similarity. Words can be comparative in two ways lexically and semantically. Words are comparable lexically if they have a comparative character grouping. Words are comparable semantically if they have the same thing, are inverse of each other, are utilized in the same way, utilized in the same setting and one is a sort of another. In our proposed system, String-Based measures are utilized for lexical similarity. String-based measures work on string groupings and character composition. A string metric is a metric that measures similarity between two content strings for surmised string coordinating or comparison. String-based measures can be divided into two sorts character-based and term-based measures. There are 7 different term-based similarity measures such as block distance, cosine similarity, Dice’s coefficient, Euclidean similarity, Jaccard similarity, Matching coefficient, and overlap coefficient. In this case, we utilize cosine similarity which is a degree of similarity between two vectors of an internal item space that measures the cosine of the point between them and is reasonable for TF-IDF values.

To verify the effectiveness of the proposed system in solving the text similarity measurement problem, this paper collected Burmese content data from social media platforms such as Facebook, YouTube, Telegram, and X. This collected data is preprocessed by using tokenization and stop words removing techniques. Processed data is extracted by using TF-IDF feature extraction and saved as a pickle file. Content, which we want to check for similarity is preprocessed with the same techniques and extracted feature. Then check the similarity value between collected data and new content using the cosine similarity method. In this paper, implemented the process of text similarity for Burmese language contents and it is shown in figure (5) and content, which we want to check similarity is shown in Table 1. This content is about “Chaunk City entered the list of the hottest cities in the world” and is published at April 10, 2024. We checked similarity for this content between collected data from 2021 to 2024. This section presents the results obtained from the implementation of the proposed text similarity system for Burmese online news content. The results of these steps are illustrated in Figures 6, 7, and 8.

Start

Load features

Loaded features on Memory

Contents as pickle files

Check similarity

>60%

Output Results

End

Content

Fig. 5. Process of text similarity for Burmese language contents

Table 2. Example of the content

“ကမ္ဘာ့အပူဆုံးမြို့စာရင်းဝင်ခဲ့တဲ့ ချောက်မြို့”

NVS 003- April 10, 2024.

ဧပြီလ ၉ ရက်အတွက် ကမ္ဘာ့အပူချိန်အမြင့်ဆုံးမြို့တွေထဲမှာ မြန်မာနိုင်ငံမှ ချောက်မြို့က အပူချိန် ၄၅ ဒသမ ၂ ဒီဂရီဆဲလ်စီးယပ်နဲ့ ထိပ်ဆုံးမှာရှိနေတယ်လို့ ကမ္ဘာ့မိုးလေဝသအချက်အလက်တွေကို ဖော်ပြပေးနေတဲ့ Eldorado ဝက်ဘ်ဆိုက်မှာ ဖော်ပြထားချက်အရ သိရပါတယ်။

နောက်ထပ်မြို့တစ်မြို့ဖြစ်တဲ့ မကွေးမြို့ကတော့ အပူချိန် ၄၃ ဒသမ ၇ ဒီဂရီဆဲလ်စီးယပ်နဲ့ ဧပြီလ ၉ ရက်အတွက် ကမ္ဘာ့အပူချိန်အမြင့်ဆုံး အဆင့် ၁၃ မှာရှိနေတယ်လို့ ဆိုပါတယ်။ ပြီးခဲ့တဲ့ဧပြီလ ၈ ရက်က ချောက်မြို့မှာ အပူချိန် ၄၆ ဒသမ ၄ ဒီဂရီဆဲလ်စီးယပ်၊ မကွေးမြို့မှာ အပူချိန် ၄၄ ဒသမ ၅ ဒီဂရီဆဲလ်စီးယပ် ရှိခဲ့တာကြောင့် ဧပြီလ ၉ ရက်မှာ အရင်နေ့ကထက်အပူချိန် အနည်းငယ်လျော့သွားတာကို တွေ့ရပါတယ်။

လက်ရှိအချိန်မှာတော့ မြန်မာနိုင်ငံအောက်ပိုင်းနဲ့ အလယ်ပိုင်းဒေသတွေမှာ အပူချိန်မြင့်မားမှုတွေကြောင့် ရုတ်တရက်သတိလစ်မေ့မြောတာတွေနဲ့ ရုတ်တရက်သေဆုံးမှုအချို့ ဖြစ်ပွားနေတာတွေကြောင့် အပူဒဏ်ကို သတိပြုနေထိုင်ကြဖို့ မိုးလေဝသနှင့်ဇလဗေဒညွှန်ကြားမှုဦးစီးဌာနနဲ့ ကျန်းမာရေးဝန်ကြီးဌာနတို့က အသိပေးနှိုးဆော် ထားပါတယ်။

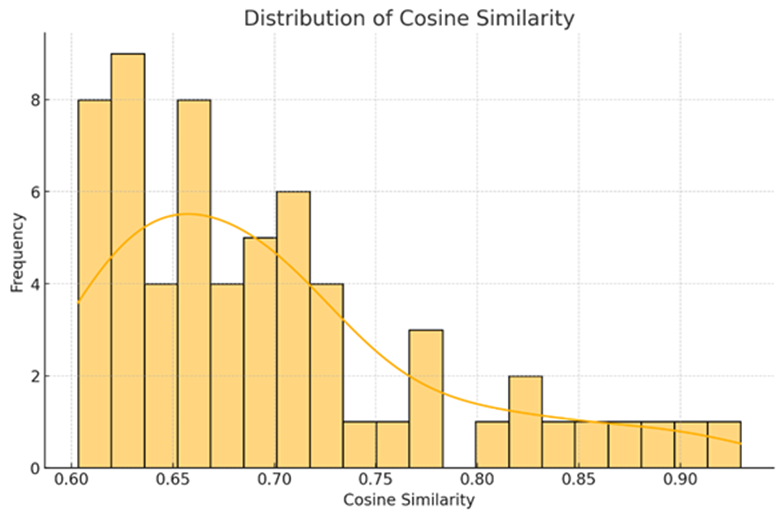
In figure (6) X-axis (Cosine Similarity) represents the cosine similarity scores between pairs of news contents. The cosine similarity score ranges from 0.60 to 0.90, where 0.60 indicates lower similarity and 0.90 indicates higher similarity. And Y-axis (Frequency) represents the number of content pairs that fall within each cosine similarity score range. This axis shows how frequently specific similarity scores occur across the dataset.

Fig. 6. Distribution of cosine similarity

The distribution is right-skewed, with the highest frequency of content pairs having cosine similarity scores between 0.60 and 0.70. There are fewer content pairs with very high similarity scores (above 0.85), suggesting that while some contents are nearly identical, the majority share moderate to high similarity. The peak of the distribution occurs around a cosine similarity score of 0.65, indicating that many contents have a moderate level of similarity. The decline in frequency as the cosine similarity score increases suggests that exact duplicates or highly similar contents are less common compared to those with moderate similarity.

|  |  |
| --- | --- |
| **Platform** | **Average Cosine Similarity** |
| Facebook | 0.716 |
| Telegram | 0.680 |
| YouTube | 0.660 |

Table 3. Average cosine similarity value by platform

In figure (7) X-axis (Cosine Similarity) represents the cosine similarity scores between pairs of news contents, ranging from lower to higher similarity scores. Y-axis (Frequency) represents the number of content pairs that fall within each cosine similarity score range for each platform. And color coding (Platforms): Blue Bars - contents from Facebook, Red Bars - contents from YouTube, Light Blue Bars - contents from Telegram. The distribution of cosine similarity scores varies across different platforms. Facebook: Dominates the frequency of article pairs in most similarity score ranges, especially between 0.60 and 0.80. YouTube: Has a lower frequency of article pairs across the similarity score ranges but shows significant frequency in the 0.65 to 0.70 range. Telegram: Shows a substantial presence in the lower similarity score ranges (0.60 to 0.70) but less so in higher ranges.

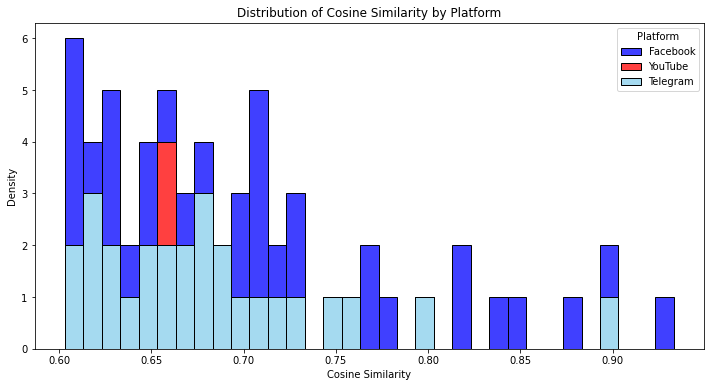
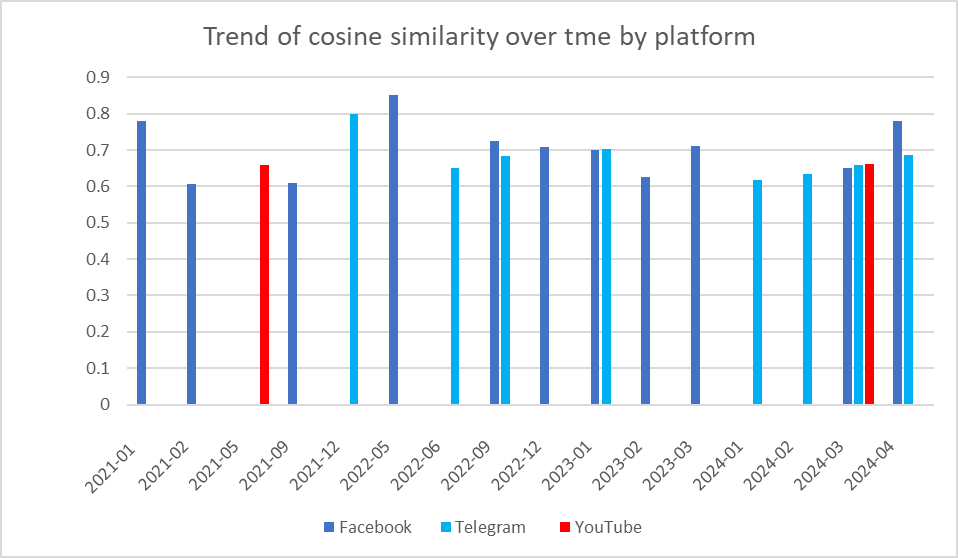
****

Fig. 7. Distribution of cosine similarity by platform

Table 4. Total uploads by platform

|  |  |
| --- | --- |
| **Platform** | **Total Uploads** |
| Facebook | 33 |
| Telegram | 27 |
| YouTube | 2 |

In figure (8) X-axis (Upload Date): represents the timeline from January 2021 to May 2024, indicating the dates when news contents were uploaded. Y-axis (Average Cosine Similarity): represents the average cosine similarity score between pairs of news contents for each platform over time. And Color Coding (Platforms): Yellow Line - Facebook, Orange Line – Telegram, and Red Line - YouTube. The trends for each platform show how the average cosine similarity of news contents changes over time. Facebook: Shows significant fluctuations in average cosine similarity, with peaks around mid-2022 and a noticeable drop towards early 2023. Telegram: Exhibits high variability with sharp increases and decreases, particularly peaking around early 2022 and dipping significantly in mid-2023. YouTube: Maintains a relatively stable trend with slight increases and decreases, staying around an average cosine similarity score of approximately 0.65 throughout the period.



Average cosine similarity

Fig. 8. Trend of cosine similarity over time by platform

Table 5. Cosine similarity value by upload date

|  |  |  |  |
| --- | --- | --- | --- |
| **Upload Date** | **Facebook** | **Telegram** | **YouTube** |
| 2021-01 | 0.780509 | 0.000000 | 0.000000 |
| 2021-02 | 0.607037 | 0.000000 | 0.000000 |
| 2021-05 | 0.000000 | 0.000000 | 0.659171 |
| 2021-09 | 0.608120 | 0.000000 | 0.000000 |
| 2021-12 | 0.000000 | 0.800100 | 0.000000 |
| 2022-05 | 0.851904 | 0.000000 | 0.000000 |
| 2022-06 | 0.000000 | 0.649675 | 0.000000 |
| 2022-09 | 0.723565 | 0.682500 | 0.000000 |
| 2022-12 | 0.707576 | 0.000000 | 0.000000 |
| 2023-01 | 0.699488 | 0.702011 | 0.000000 |
| 2023-02 | 0.625422 | 0.000000 | 0.000000 |
| 2023-03 | 0.710330 | 0.000000 | 0.000000 |
| 2024-01 | 0.000000 | 0.617343 | 0.000000 |
| 2024-02 | 0.000000 | 0.634714 | 0.000000 |
| 2024-03 | 0.650444 | 0.658029 | 0.661727 |
| 2024-04 | 0.780072 | 0.687464 | 0.000000 |

1. **CONCLUSION AND FUTURE WORK**

The proposed system presents a strong method for measuring text similarity in Burmese online news media. The results show that many news articles share a significant amount of content with little or no modification, which can impact content quality and contribute to misinformation. The distribution of cosine similarity scores reveals that a considerable number of news articles are moderately similar, with a notable amount being highly similar. This emphasizes the need to monitor content similarity to address issues such as duplicate content and misinformation in online news media.

The distribution of cosine similarity scores across different social media platforms highlights variations in content sharing and duplication practices. Facebook exhibits a higher tendency for content similarity, while YouTube and Telegram show more moderate levels of similarity. This analysis is crucial for understanding how different platforms contribute to content dissemination and potential duplication, providing valuable insights for media monitoring and management strategies.

Trend analysis uncovers distinct patterns of content similarity for different platforms over time. Facebook and Telegram show more variability in content similarity, while YouTube remains stable. Understanding these trends can help in identifying periods of high content duplication, which is crucial for media monitoring and addressing issues related to misinformation and content quality.

Future work will focus on enhancing tokenization techniques for better handling of the Burmese language, incorporating semantic analysis to capture deeper contextual similarities, implementing real-time analysis capabilities, and optimizing processing time through the use of parallel computing and Kubernetes platforms.

**REFERENCES**

[1] C. Li, L. Zhao, X. Li, and L. Wang, “Text sentiment classification model based on TF-IDF weighted convolutional neural network,” Journal of Chongqing University of Technology (Natural Science), vol. 35, no. 11, pp. 109–115, 2021.

[2] K. Z. Ye, Y. M. Naing, Y. Naung, K. P. Nyein and T. Z. Lin, "Implementation of Burmese Language News Classification System by Using SVM and LSTM Machine Learning Algorithm," 2023 IEEE 6th International Conference on Computer and Communication Engineering Technology (CCET), Beijing, China, 2023, pp. 27-32.

[3] Ye Kyaw Thu, Andrew Finch, Yoshinori Sagisaka and Eiichiro Sumita, "A Study of Myanmar Word Segmentation Schemes for Statistical Machine Translation", in Proceedings of the 11th International Conference on Computer Applications (ICCA 2013), February 26~27, 2013, Yangon, Myanmar, pp. 167-179.

[4] L. Zhang, Y. Jiang, and L. Sun, “An improved TF-IDF text clustering method,” Journal of Jilin University (Science Edition), vol. 59, no. 5, pp. 1200–1204, 2021.

[5] Nabila Febriyanti , Dian Palupi Rinib, Osvari Arsalan, “Text Similarity Detection Between Documents Using Case Based Reasoning Method with Cosine Similarity Measure (Case Study SIMNG LPPM Universitas Sriwijaya)”, Sriwijaya Journal of Informatic and Applications.

Vol. 3, No. 2, August 2022, pp. 85-94

[6] Kapellas, N. and Kapidakis, S. “A Text Similarity Study: Understanding How Differently Greek News Media Describe News Events”, In Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2022)- Volume 2: KEOD, pages 245-252.

[7] M. AlMousa, R. Benlamri, and R. Khoury, “Exploiting nontaxonomic relations for measuring semantic similarity and relatedness in WordNet,” Knowledge-Based Systems, vol. 212, Article ID 106565, 2021.

[8] Y. Zhou and M. Dai, “News recommendation technology combining semantic analysis with TF-IDF method,” Computer Science, vol. 40, no. S2, pp. 267–269, 2013.

[9] S. Yildirim and T. Yildiz, “A comparative analysis of text classification for Turkish language,” Pamukkale University Journal of Engineering Sciences, vol. 24, no. 5, pp. 879–886, 2018.