STMIK ROSMA
School Of Technology Information

IEEE
INDONESIA SECTION

IEEE Conference No. #60125

Conference Book

ICMERALDA

International Conference on Modeling & E-Information Research, Artificial Learning & Digital Applications

Karawang-Indonesia, 24 November 2023

icmeralda.rosma.ac.id

IEEE Conference No. 60125

The 2023 1st International Conference on Modelling & E-Information Research, Artificial Learning and Digital Applications

(ICMERALDA)

24 November 2023

STMIK ROSMA KARAWANG

icmeralda.rosma.ac.id

THE COMMITTEE

STEERING COMMITTEE
Darmansyah, S.Kom., M.Kom.
Chief of STMIK ROSMA Karawang Indonesia

ORGANIZING COMMITTEE
General Chair
Anggi Elanda, S.Kom., M.Kom

Treasurer
Dr. Femmy Effendy, ST., MM

Technical Program Committee
Arif Maulana Yusuf, S.Kom., M.Kom

Publication Chair & Editor
Rahmat Gunawan, S.Kom., M.Kom

Secretary
Vina Maulidina, S.Kom

Program Chair
Yeny Rostiani, S.E., MM

Technical Team
Daniel Richi Roynaldi Simamora, S.Kom
Shiddieq Fathurrozaq, S.Kom
Aji Zulfikar Nugraha

TECHNICAL PROGRAM COMMITTEE

Muthukumaran Vaithianathan (SAMSUNG SSI - USA)
Sandy Kosasi (STMIK Pontianak - Indonesia)
Intan Mutia (IPB University - Indonesia)
Ahmed Kawther (Mustansiriyah Universtiy - Iraq)
Hastari Utama (Universitas Amikom Yogyakarta - Indonesia)
Shashikant Patil (Mumbai University - India)
Rafee Al Ahsan (University of Calgary - Canada)
Hayder AL-Qaysi (University of Diyala - Iraq)
Cucut Susanto (Universitas Dipa Makassar - Indonesia)
Widyastuti Andriyani (Universitas Teknologi Digital Indonesia - Indonesia)
Qingli Zeng (UMKC - USA)
Nurdin Nurdin (Universitas Islam Negeri Datokarama Palu - Indonesia)
Grienggrai Rajchakit (Maejo University - Thailand)
Marzuki Pilliang (Esa Unggul University - Indonesia)
Adwaita Jadhav (Apple - USA)
Wasnaa Kadhim (University of Information Technology and Communications - Iraq)
Channa Khieng (National Polytechnic Institute of Cambodia - Cambodia)
Kirana Fatika (College of Advanced Manufacturing Innovation - Thailand)
Henderi Henderi (University of Raharja - Indonesia)
Neeraj Singh (Senior Technical Lead, Automotive, HCL Technologies - India)
Seng Hansun (Universitas Multimedia Nusantara - Indonesia)
Ajay Shukla (All India Intitute of Ayureveda (AIIA) - India)
Widodo Widodo (Universitas Negeri Jakarta - Indonesia)
Hanny Haryanto (Universitas Dian Nuswantoro - Indonesia)
Andi Wahju Rahardjo Emanuel (Universitas Atma Jaya Yogyakarta - Indonesia)
Basaeir Ahmed (University of Basra - Iraq)
Terlapu Vital (jntuK - India)
Sancha Panpaeng (Chiang Mai Rajabhat University - Thailand)
Mayur Rele (Parachute Health - USA)
Karthik Sivarama Krishnan (Rochester Institute of Technology - USA)
Fajar Hermawati (Universitas 17 Agustus 1945 Surabaya - Indonesia)
Reza Khalilian (MUI Research Assistant and Author - Iran)
Mohd Ashraf Ahmad (Universiti Malaysia Pahang - Malaysia)
Wihayati Wihayati (Satya Wacana Christian University - Indonesia)
Pavel Loskot (ZJU-UIUC Institute - China)
Lala Hucadinota Ainul Amri (Politeknik Negeri Madiun - Indonesia)
Abul Al Arabi (Texas A&M University - USA)
Aakashjit Bhattacharya (Indian Institute of Technology Kharagpur - India)
Mohammad Zedan (Ninaveh University - Iraq)
Saad Chakkor (University of Abdelmalek Essaâdi - Morocco)
Tri Priyambodo (Universitas Gadjah Mada - Indonesia)
Yi-Jen Su (National Penghu University of Science and Technology - Taiwan)
Naveen Aggarwal (Panjab University - India)
Domy Kristomo (Universitas Teknologi Digital Indonesia - Indonesia)
Nor Liyana Mohd Shuib (University of Malaya - Malaysia)
Poorani Shivkumar (ESEC - India)
Anggi Elanda (STMIK ROSMA - Indonesia)
Hozairi Hozairi (Universitas Islam Madura - Indonesia)
Paulus Insap Santosa (Universitas Gadjah Mada - Indonesia)
Luis Alves (Polytechnic Institute of Bragança - Portugal)
Shahzad Ashraf (Hohai University Changzhou, Jiangsu - China)
Intan Ermahani A. Jalil (Universiti Teknikal Malaysia Melaka - Malaysia)
Suvadip Batabyal (NIT Durgapur - India)
Suman Kr. Dey (National Institute of Technology Rourkela - India)
Akhil Gupta (Lovely Professional University - India)
Ramkumar Jaganathan (Sri Krishna Arts and Science College - India)

Lie Jasa (Udayana University - Indonesia)
Dan Milici (University of Suceava - Romania)
Nitish Ojha (Sharda University, Greater Noida, UP - India)
Ilker Ali Ozkan (Selcuk University - Turkey)
Kiran Sree Pokkuluri (Shri Vishnu Engineering College for Women(A) - India)
Somnuk Puangpronpitag (Mahasarakham University - Thailand)
Ali Rafiei (General Motors - Canada)
Zairi Ismael Rizman (Universiti Teknologi MARA - Malaysia)
Aditi Sharma (Parul University, Vadodara - India)
Akbar Sheikh-Akbari (Leeds Beckett University - United Kingdom (Great Britain))
Joey Suba (University of the Assumption - Philippines)
Chakib Taybi (Mohammed First University - Morocco)
Warusia Yassin (Universiti Teknikal Malaysia Melaka - Malaysia)
Chaoqun You (Singapore University of Technology and Design - Singapore)
Nur Zareen Zulkarnain (Universiti Teknikal Malaysia Melaka - Malaysia)
Aashish Gadgil (KLS Gogte Institute of Technology, Belgaum - India)
Roberto Carlos Herrera Lara (Electricity Company of Quito - Ecuador)
Deven Panchal (AT&T - USA)
Alvaro Paricio García (Universidad de Alcala - Spain)
Edy Prayitno (Universitas Teknologi DIgital Indonesia - Indonesia)
Al Gburi Qahtan (Ministry of Education - Iraq)
Suharjito Suharjito (Bina Nusantara University - Indonesia)
Marwah Dabdawb (University of Mosul - Iraq)
Arna Fariza (Politeknik Elektronika Negeri Surabaya - Indonesia)
Rabab Farhan Abbas (University of Technology - Iraq)
George Dekoulis (Aerospace Engineering Institute (AEI) - Cyprus)
Priya Ranjan (University of Petroleum and Energy Studies (UPES) - India)
Sayantam Sarkar (MVJ College of Engineering - India)
Thaweesak Yingthawornsuk (King Mongkut's University of Technology Thonburi - Thailand)
Sultan Alzahrani (KACST - Saudi Arabia)
Maslin Masrom (Universiti Teknologi Malaysia - Malaysia)
Evi Triandini (Institut Teknologi dan Bisnis STIKOM Bali - Indonesia)
Yosua Alvin Adi Soetrisno (Diponegoro University - Indonesia)
Andria Arisal (National Research and Innovation Agency (BRIN) - Indonesia)
Ahmad Chusyairi (IPB University - Indonesia)
Seyyidahmed Lahmer (University of Padova - Italy)
Irfan Pratama (Universitas Mercubuana Yogyakarta - Indonesia)
Rasim Rasim (Indonesia University of Education - Indonesia)
Dedi Rohendi (Universitas Pendidikan Indonesia - Indonesia)

2023 International Conference on Modeling & E-Information Research, Artificial Learning and Digital Applications (ICMERALDA) | 979-8-3503-6935-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICMERALDA60125.2023.10458178

2023 International Conference on Modeling & E-Information Research, Artificial Learning and Digital Applications (ICMERALDA) | 979-8-3503-6935-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICMERALDA60125.2023.10458170

Technical Program Committee

Intan Ermahani A. Jalil	Universiti Teknikal Malaysia Melaka	Malaysia
Rabab Farhan Abbas	University of Technology	Iraq
Sukarya Ade	Indonesian Researcher and Scientist Institute	Indonesia
Naveen Aggarwal	Panjab University	India
Mohd Ashraf Ahmad	Universiti Malaysia Pahang	Malaysia
Basaeir Ahmed	University of Basra	Iraq
Rafee Al Ahsan	University of Calgary	Canada
Hayder AL-Qaysi	University of Diyala	Iraq
Luis Alves	Polytechnic Institute of Bragança	Portugal
Yosua Alvin Adi Soetrisno	Diponegoro University	Indonesia
Sultan Alzahrani	KACST	Saudi Arabia
Lala Hucadinota Ainul Amri	Politeknik Negeri Madiun	Indonesia
Widyastuti Andriyani	Universitas Teknologi Digital Indonesia	Indonesia
Abul Al Arabi	Texas A&M University	USA
Andria Arisal	National Research and Innovation Agency (BRIN)	Indonesia
Shahzad Ashraf	Hohai University Changzhou, Jiangsu	China
Suvadip Batabyal	NIT Durgapur	India
Aakashjit Bhattacharya	Indian Institute of Technology Kharagpur	India
Saad Chakkor	University of Abdelmalek Essaâdi	Morocco
Ahmad Chusyairi	IPB University	Indonesia
Marwah Dabdawb	University of Mosul	Iraq
George Dekoulis	Aerospace Engineering Institute (AEI)	Cyprus

Suman Kr. Dey	National Institute of Technology Rourkela	India
Anggi Elanda	STMIK ROSMA	Indonesia
Andi Wahju Rahardjo Emanuel	Universitas Atma Jaya Yogyakarta	Indonesia
Arna Fariza	Politeknik Elektronika Negeri Surabaya	Indonesia
Kirana Fatika	College of Advanced Manufacturing Innovation	Thailand
Aashish Gadgil	KLS Gogte Institute of Technology, Belgaum	India
Akhil Gupta	Lovely Professional University	India
Seng Hansun	Universitas Multimedia Nusantara	Indonesia
Hanny Haryanto	Universitas Dian Nuswantoro	Indonesia
Henderi Henderi	University of Raharja	Indonesia
Fajar Hermawati	Universitas 17 Agustus 1945 Surabaya	Indonesia
Roberto Carlos Herrera Lara	Electricity Company of Quito	Ecuador
Hozairi Hozairi	Universitas Islam Madura	Indonesia
Stmik Rosma Icmeralda	STMIK Rosma Karawang	Indonesia
Paulus Insap Santosa	Universitas Gadjah Mada	Indonesia
Adwaita Jadhav	Apple	USA
Ramkumar Jaganathan	Sri Krishna Arts and Science College	India
Lie Jasa	Udayana University	Indonesia
Wasnaa Kadhim	University of Information Technology and Communications	Iraq
Ahmed Kawther	Mustansiriyah Universtiy	Iraq
Reza Khalilian	MUI Research Assistant and Author	Iran
Channa Khieng	National Polytechnic Institute of Cambodia	Cambodia
Sandy Kosasi	STMIK Pontianak	Indonesia

Domy Kristomo	Universitas Teknologi Digital Indonesia	Indonesia
Seyyidahmed Lahmer	University of Padova	Italy
Pavel Loskot	ZJU-UIUC Institute	China
Maslin Masrom	Universiti Teknologi Malaysia	Malaysia
Dan Milici	University of Suceava	Romania
Nor Liyana Mohd Shuib	University of Malaya	Malaysia
Intan Mutia	IPB University	Indonesia
Nurdin Nurdin	Universitas Islam Negeri Datokarama Palu	Indonesia
Nitish Ojha	Sharda University, Greater Noida, UP	India
Ilker Ali Ozkan	Selcuk University	Turkey
Deven Panchal	AT&T	USA
Sancha Panpaeng	Chiang Mai Rajabhat University	Thailand
Alvaro Paricio García	Universidad de Alcala	Spain
Shashikant Patil	uGDX School of Technology	India
Marzuki Pilliang	Esa Unggul University	Indonesia
Kiran Sree Pokkuluri	Shri Vishnu Engineering College for Women(A)	India
Irfan Pratama	Universitas Mercubuana Yogyakarta	Indonesia
Edy Prayitno	Universitas Teknologi DIgital Indonesia	Indonesia
Tri Priyambodo	Universitas Gadjah Mada	Indonesia
Somnuk Puangpronpitag	Mahasarakham University	Thailand
Al Gburi Qahtan	Ministry of Education	Iraq
Ali Rafiei	General Motors	Canada
Grienggrai Rajchakit	Maejo University	Thailand

Priya Ranjan	University of Petroleum and Energy Studies (UPES)	India
Rasim Rasim	Indonesia University of Education	Indonesia
Mayur Rele	Parachute Health	USA
Zairi Ismael Rizman	Universiti Teknologi MARA	Malaysia
Dedi Rohendi	Universitas Pendidikan Indonesia	Indonesia
Sayantam Sarkar	MVJ College of Engineering	India
Aditi Sharma	Symbiosis Institute of Technology, Symbiosis International Deemed University, Pune	India
Akbar Sheikh-Akbari	Leeds Beckett University	United Kingdom (Great Britain)
Poorani Shivkumar	ESEC	India
Ajay Shukla	All India Intitute of Ayureveda (AIIA)	India
Neeraj Singh	Senior Technical Lead, Automotive, HCL Technologies	India
Karthik Sivarama Krishnan	Rochester Institute of Technology	USA
Yi-Jen Su	National Penghu University of Science and Technology	Taiwan
Joey Suba	University of the Assumption	Philippines
Suharjito Suharjito	Bina Nusantara University	Indonesia
Cucut Susanto	Universitas Dipa Makassar	Indonesia
Chakib Taybi	Mohammed First University	Morocco
Evi Triandini	Institut Teknologi dan Bisnis STIKOM Bali	Indonesia
Hastari Utama	Universitas Amikom Yogyakarta	Indonesia
Muthukumaran Vaithianathan	SAMSUNG SSI	USA
Terlapu Vital	jntuK	India
Ferry Wahyu Wibowo	Universitas Amikom Yogyakarta	Indonesia

Widodo Widodo	Universitas Negeri Jakarta	Indonesia
Wihayati Wihayati	Satya Wacana Christian University	Indonesia
Warusia Yassin	Universiti Teknikal Malaysia Melaka	Malaysia
Thaweesak Yingthawornsuk	King Mongkut's University of Technology Thonburi	Thailand
Chaoqun You	Fudan University	China
Mohammad Zedan	Ninaveh University	Iraq
Qingli Zeng	UMKC	USA
Nur Zareen Zulkarnain	Universiti Teknikal Malaysia Melaka	Malaysia

A

A C D E F G I K L M O P R S T U V

C

A C D E F G I K L M O P R S T U V

D

A C D E F G I K L M O P R S T U V

2023 International Conference on Modeling & E-Information Research, Artificial Learning and Digital Applications (ICMERALDA) | 979-8-3503-6935-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICMERALDA60125.2023.10458164

E

A C D E F G I K L M O P R S T U V

F

A C D E F G I K L M O P R S T U V

G

A C D E F G I K L M O P R S T U V

I

A C D E F G I K L M O P R S T U V

K

A C D E F G I K L M O P R S T U V

L

A C D E F G I K L M O P R S T U V

M

A C D E F G I K L M O P R S T U V

O

A C D E F G I K L M O P R S T U V

P

A C D E F G I K L M O P R S T U V

R

A C D E F G I K L M O P R S T U V

S

A C D E F G I K L M O P R S T U V

T

A C D E F G I K L M O P R S T U V

U

Using Naïve Bayes and Support Vector Machines Algorithms to Predict Television Channel Viewership

V

Video-Based Lobster Species Identification Using YOLOv7

2023 International Conference on Modeling & E-Information Research, Artificial Learning and Digital Applications (ICMERALDA) | 979-8-3503-6935-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICMERALDA60125.2023.10458198

Trending Topic Detection during Pandemic (Covid- 19) In Indonesian Tweets Using the Document Pivot (Doc-p) Method and BN-grams

1st Indra
Faculty of Technology of Information
Universitas Budi Luhur
South Jakarta, Indonesia
indra@budiluhur.ac.id

2nd Agus Umar Hamdani
Faculty of Technology of Information
Universitas Budi Luhur
South Jakarta, Indonesia
agus.umarhamdani@budiluhur.ac.id

3rd Suci Setiawati
Faculty of Technology of Information
Universitas Budi Luhur
South Jakarta, Indonesia
1811520079@student.budiluhur.ac.id

4th Sukha Vaddhana
Faculty of Technology of Information
Universitas Budi Luhur
South Jakarta, Indonesia
1811520095@student.budiluhur.ac.id

***Abstract*— People are providing a variety of information about the coronavirus outbreak, including the needs of affected communities and the location of the outbreak. Such information can be used as one of the resources to map the coronavirus outbreak events and the needs of affected people in Indonesia. However, the information obtained from social media has an informal structure and has low reliability as an information provider. Unstructured social media data makes it difficult to identify information related to trending topics, especially related to the coronavirus outbreak. Therefore, in this study, we used Document Pivot (Doc-p) and BN-grams (January to May 2020) methods to detect trending topics in Indonesian tweets. In our experiments, we investigate the impact of different topic numbers and master data on the quality of the resulting trending topics. We measure the accuracy of detecting trending topics by comparing both methods to trending topics found in local news and Twitter. Our experimental results show that using 10 topics yields the highest topic recall. Trending topics generated by BN-grams have the highest topic recall values. Stemming also reduces the quality of the resulting trending topics. The topic recall values of Doc-p and BN-gram from the four datasets are 75% and 50%, respectively. Overall, Doc-p has higher topic recall compared to his BN-gram because, unlike BN-gram, the dataset is used without stemming.**

Keywords: Trending Topics Pandemic, Covid-19, BN-grams, Doc-p.

I. INTRODUCTION

According to [11], a trending topic is a term (consisting of one or more words) that appears in a significant number of tweets and originates from a specific source. Trending topics on social media and search engines are presented in the form of Indonesian trends (Twitter), Google Trends (Google), and popular news (Facebook). The rapid development of trending topics in social media has resulted in the emergence of numerous research endeavors aimed at detecting trending topics on social platforms.

The processing methods for trending topics are generally classified into three categories: textual content, social content, and hybrid [3]. The textual content method is a trending topic detection technique based on the processing of text or messages posted on social media.

Indonesia was affected by the Covid-19 outbreak from January to May 2020. The majority of user statuses on Twitter in Indonesia contain information or opinions related to Covid-19. The quantity of Twitter data (tweets) continues to grow, with a progressively extensive discussion about Covid-19. This research proposes trending topic detection as the primary foundation of information for the government, online news, and television. Information originating from Twitter, according to [1], has the characteristic of being original, unedited, and factual. The information generated on Twitter comes directly from the Twitter user community, a characteristic commonly known as Original content. Information from Twitter is unedited, unlike in online news media where there is an editorial team prior to the news being broadcast. Information from Twitter depicts the reality of events happening at specific times and locations, making it factual.

Authorized licensed use limited to: Institut Teknologi Sepuluh Nopember. Downloaded on March 26,2024 at 13:42:05 UTC from IEEE Xplore. Restrictions apply.

The contribution of this study is to examine the effect of varying the number of topics when comparing BN-gram and Doc-p methods on the quality of trending topics related to Covid-19 in Indonesian tweets. Testing BN-gram Doc-p method on four datasets, Impact of implementing stemming in preprocessing, Comparing BN-gram and Doc-p method on resulting trending topics in Indonesian tweets. Impact on quality.

II. LITERATUR REVIEW

Several previous studies have conducted research on trending topics related to Covid-19. [9] This research proposes a model based on the Universal Sentence Encoder to analyze and detect trending topics as well as the primary concerns of the public regarding COVID-19 on Twitter. This model utilizes semantic representations from the Universal Sentence Encoder and the K-means clustering algorithm to group tweets with semantic similarity. The experimental results demonstrate the superiority of this model in detecting informative topics compared to baseline methods such as TF-IDF and Latent Dirichlet Allocation (LDA). This approach is not limited to specific data distributions and can be applied to social media and other contexts beyond COVID-19.

Whereas in research [11] there is no previous research that compares the results of the k-means algorithm, CLOPE clustering, and Latent Dirichlet Allocation (LDA) topic modelling to detect trending topics in tweets. As not all tweets contain hashtags, we considered three sets of training data features: hashtags, keywords, and keywords + hashtags in this study. The proposed methodology demonstrates that CLOPE can also be applied in non-transactional databases, such as Twitter datasets, to address trending topic discovery and can provide more topic patterns than k-means and LDA. The use of additional feature sets has enhanced the results of k-means and LDA, indicating that keywords + hashtags can identify more meaningful topics.

This study discusses the trending topics on Twitter, which are defined as names, phrases, or topics mentioned at a higher rate than others on the platform. Although trending topics on Twitter have been shown to impact various public events and market changes, research that comprehends the dynamics of these trending topics is still lacking. The focus of this research is to examine the trending topics on Twitter in 2018, utilizing the Twitter trends API and evaluating the dataset based on six criteria: lexical analysis, time taken, trend recurrence, trend duration, tweet volume, and language analysis. This research contributes by providing general statistics, top trending topics based on each criterion, and computed distributions to elucidate the data. These findings aim to enhance the understanding of the characteristics and patterns associated with Twitter trending topics, filling gaps in the existing literature on their dynamics and impact [5].

This study presents an in-depth analysis of the discourse changes related to Covid-19 on Twitter during the first wave of the pandemic. Through topic modeling, this research identifies the evolution of topics being discussed, while sentiment analysis reveals a shift from positive to negative sentiments in line with the phases of lockdown and reopening. Furthermore, the increase in tweet subjectivity is noteworthy, and the figurative framework of war undergoes changes as real events enter the discourse. The research findings provide valuable insights into how perspectives and responses to the pandemic are reflected in the language of social media [8].

During the COVID-19 pandemic, understanding public concerns and sentiments through social media, especially Twitter, has become crucial. This study proposes a combined approach of peak detection and clustering to extract insights from COVID-19-related tweets. As a result, important topics were identified, including health status, government policies, economic crisis, COVID-19 updates, prevention, as well as vaccines and treatments. This approach proved to be accurate in capturing relevant topics from the large and noisy stream of tweets in the United States [2].

III. RESEARCH METHOD

This section explains the fundamental concept of the BN-grams and Doc-p methods.

A. BN-gram Method

Trending topic detection using the BN-grams method employs three steps as illustrated in Figure 1 [1], [7].

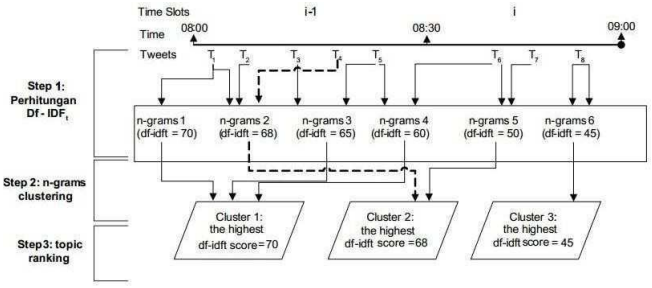

Fig 1. The stages of trending topic detection using the BN-grams method [1].

In Figure 1, during the first step, the calculation of document frequency ($dfidf_t$) is performed for each n-gram. In the second step, a set of n-grams undergoes clustering through hierarchical clustering. In the third step, each cluster undergoes topic ranking to detect trending topics. Before processing the tweets in the first step, the tweets are collected in the current time slot (t) and the previous time slot ($t - 1$). Furthermore, in both of these time slots, the tweets undergo preprocessing, including tokenization, stemming, and aggregation. This study utilizes two types of aggregation, namely time aggregation and topic aggregation. Time aggregation is the consolidation of tweets based on closely related time intervals in each time slot. Topic aggregation involves the merging of tweets based on the similarity between tweets. In this study, the similarity between tweets is calculated using the LSH method. If trending topics are generated from clusters with n-grams with the highest df-idf values (topic ranking) [10].

B. The Document Pivot Method (Doc-p)

Tweets collected at certain time slots with the time aggregation approach are preprocessed with two stages, namely tokenization and stemming. The preprocessing results are subjected to trending topic detection using the Doc-p method. The Doc-p method employs four main steps, namely 1). Clustering tweets using the Locality Sensitive Hashing method, 2). Removal of clusters whose number of members is below the threshold, 3). Weighting of each cluster and 4). Cluster ranking to determine trending topics [1].

Authorized licensed use limited to: Institut Teknologi Sepuluh Nopember. Downloaded on March 26,2024 at 13:42:05 UTC from IEEE Xplore. Restrictions apply.

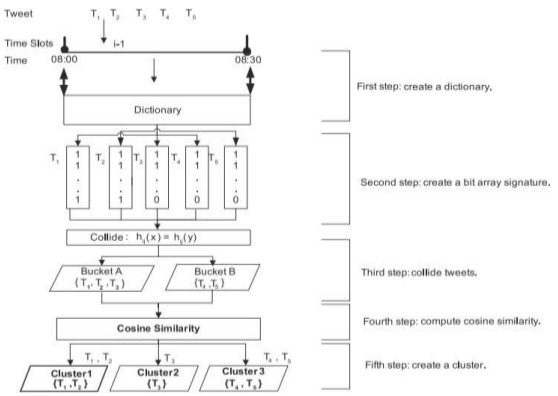

Fig 2. Clustering tweet with LSH [6]

C. Difference between Pivot (Doc-P) and BN-grams Method

According to [1], there is an explanation about DOC-P (Document-pivot) in the context of Topic Detection and Tracking (TDT). DOC-P is used in a general-purpose topic detection method that produces two complementary types of output. In the DOC-P method, a topic is represented by a cluster of documents. There are advantages and disadvantages to this method. The main weakness lies in the issue of cluster fragmentation and dependence on an arbitrary threshold for including new documents into existing topics in the streaming context.

Additionally, there is also the feature-pivot method, where a cluster of keywords is produced. Both of these approaches are considered complementary, and depending on the specific application, one may be more suitable than the other.

DOC-P method commonly employs several approaches. One of them is document clustering based on textual similarity among documents. An example is the method discussed by Phuvipadawat and Murata for detecting breaking news on Twitter. They use a bag-of-words representation with weighting to create clusters of tweets based on text similarity. Some challenges of the DOC-P method include cluster fragmentation and the need for an arbitrary threshold. However, this method can be improved by considering other dimensions, such as temporal proximity, and using clustering algorithms that take both dimensions into account.

BN-grams method with three stages, namely:

1. Extracting keywords in the form of bigrams or trigrams from the Twitter dataset; The first step is to take the Twitter dataset and perform the extraction of n-grams, here in the form of bigrams or trigrams. For example, bigrams are pairs of two words that appear consecutively in the text.
2. Performing ranking for each n-gram (bigrams or trigrams) using df-idft (document frequency invers document term) calculation to obtain scores for each bigram or trigram. After obtaining bigrams or trigrams, the next step is to score each n-gram using the df-idft (document frequency invers document term) calculation. This metric helps assess how important an n-gram is in the overall context of the dataset and identifies keywords that may reflect trending topics. DF-IDF gives higher weight to words that rarely appear in the entire dataset but frequently appear in specific documents, thus distinguishing more relevant keywords.
3. The next step is to perform clustering for each n-gram based on the distance between n-grams to obtain trending topics from each cluster. The last step is to cluster these ranked n-grams based on the distance between them. This can be done using clustering methods such as k-means or hierarchical clustering. Clustering aims to group n-grams that have similarities in context or topic. In this way, high-scoring n-grams that are similar to each other will be grouped together, forming clusters. These clusters can then be interpreted as trending topics or popular topics within the Twitter dataset.

IV. RESULT AND DISCUSSION

A. Dataset

This study uses four datasets namely P1, P2, P3 and P4 and sequentially has a total number of tweets of 2,186, 1,395, 7,175, 1,223. Datasets P1, P2 were collected on June 11-12, 2020 and July 01, 2020 respectively. Datasets P3, P4 were collected on April 17 and June 19, 2020, respectively. Datasets P1-P4 were constructed based on keywords such as “unemployment, layoffs, termination, return to work, OTG, ODP, PDP, quarantine, lockdown, covid-19, covid, corona, BNPB, and Ministry of Health”. In addition, keywords were added in the data collection if there were no keywords corresponding to the Covid-19 events. The trending topic detection in this research is not based on a specific event. Conversely, the anticipated outcome of the proposed method's trending topics is to unveil events that remain unreported by online news media or television.

B. Evalution Method

Testing BN-grams and Doc-p is conducted by comparing the number of topics generated from the proposed method with the ground truth created based on popular news in online media.

Authorized licensed use limited to: Institut Teknologi Sepuluh Nopember. Downloaded on March 26,2024 at 13:42:05 UTC from IEEE Xplore. Restrictions apply.

Table 1. The Ground Truth sample

Dataset	Time Period	Title/Headline News	Keyword
P-1	11 June 2020 (00:00 -23:59)	"The Omnibus Law on Job Creation Was Made Because Unemployment Rates Are Still High." (RUU Cipta Kerja Dibuat Karena Tingkat Pengangguran Masih Tinggi)	unemployment; omnibus law on job; work (pengangguran;ruu cipta;kerja)
P-2	01 July 2020 (15:00-22:00)	Faisal Basri: Luhut Pandjaitan is more dangerous than the Coronavirus COVID-19. (Faisal Basri: Luhut Pandjaitan Lebih Berbahaya dari Coronavirus COVID-19)	Covid; lbp; dangerous (Covid; lbp;berbahaya)
P-3	17 April 2020 (00:00-23:59)	The Head of the Criminal Investigation Department Violators of Covid-19 Handling and PSBB Will Be Punished. (Kabareskrim Pelanggar Penanganan Covid-19 dan PSBB akan Dihukum)	The Head of the Criminal Investigation Department; Violator; psbb; punished (Kabareskrim; pelanggar; psbb; dihukum)
P-4	19 June 2020 (01:00-01:07)	Asymptomatic COVID-19 Positive Cases Need to be Educated About the Transmission of the Coronavirus (OTG Positif Covid-19 Perlu Diberikan Edukasi Penularan Virus Corona)	Asymptomatic;covid;corona (Otg;covid;corona)

The overall test in this article uses three measurements: topic recall (TR), keyword precision (KP), and keyword recall (KR). Topic recall (TR) is the comparison between trending topics and ground truth topics (Equation (1)). Keyword precision (KP) is the ratio of trending topic keywords that match ground truth keywords compared to the total number of trending topic keywords (Equation (2)). Keyword recall (KR) is the ratio of trending topic keywords that match ground truth keywords compared to the total number of keywords in the ground truth (Equation (3)). The three measurements are expressed by the following equation:

$$TR = \frac{|GT \cap BT|}{|GT|} \quad (1)$$

$$KP = \frac{|KGT \cap KBT|}{|KBT|} \quad (2)$$

$$KR = \frac{|KGT \cap KBT|}{|KGT|} \quad (3)$$

Here, GT represents a set of topics in the ground truth, BT represents a set of trending topics, KGT represents a set of keywords in the ground truth, and KBT represents a set of keywords in trending topics.

C. Evalution Result

a. Comparison of BN-gram and Doc-p Across Four Datasets

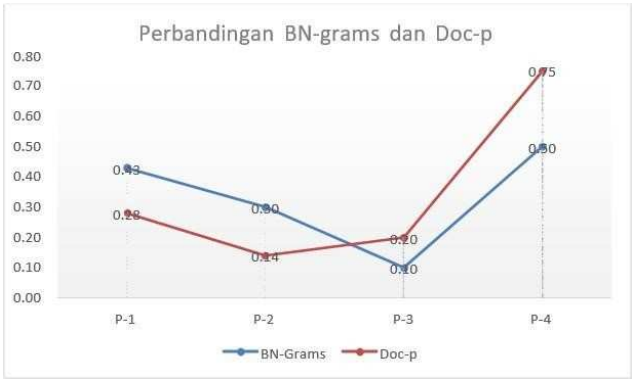

Fig 3. Comparing BN-gram and Doc-p on Four Datasets

Figure 3 explains the comparison of the test results of trending topic detection from the BN-grams and Doc-p methods. The testing results indicate that the Doc-p method has the highest topic recall value and relatively good keyword precision and keyword recall values compared to the other five methods across the entire dataset. The Doc-p method has the highest topic recall value on datasets P-3 and P-4. Whereas in the P-2 and P-1 datasets, the highest topic recall value is owned by BN-grams. Regarding datasets P-1, P-2, P-3, and P-4 have been described in Table 1.

In the BN-grams and Doc-P methods, there is a decrease in recall values at P-1 and P-2. The P-2 and P-3 values in BN-grams experienced a decrease, while Doc-P experienced an increase. The P-3 and P-4 values experienced an increase in recall, both in BN-grams and Doc-P. The large number of records in the P-3 data resulted in a decrease in the recall value of P-3 in BN-grams because there were many diverse or heterogeneous topics, making it difficult to detect trending topics. On the contrary, if the number of records is small, the recall value tends to increase. In contrast to Doc-p, the recall value always increases and is not dependent on the number of records.

Authorized licensed use limited to: Institut Teknologi Sepuluh Nopember. Downloaded on March 26,2024 at 13:42:05 UTC from IEEE Xplore. Restrictions apply.

Table 2. Trending Topic Testing Results BN-grams and Doc-p

Data set	Total	BN-Grams			Doc-p		
	Topic	Topic Recall	Keyword Precision	Keyword Recall	Topik Recall	Keyword Precision	Keyword Recall
P-1	10	0,43	0,24	1,00	0,28	0,20	1,00
P-2	10	0,30	0,32	1,00	0,14	0,57	1,00
P-3	10	0,10	1,00	0,50	0,20	0,23	1,00
P-4	10	0,50	0,67	0,63	0,75	0,45	0,74

Testing on P-3 and P-4 indicates that the Doc-p method produces several topics with high similarity to the trending topics on Twitter, online news, and local news. This occurs because the Doc-p method utilizes a dataset with preprocessing without stemming. Preprocessing without stemming has an impact on the quality of trending topic sentences produced according to the applicable EYD (Indonesian language rules).

Testing on P-2 and P-1 shows that the BN-grams method has a higher topic recall value than Doc-p. This happens because the use of trigrams in trending topic processing has an impact on the results of trending topic sentences that are more SPOK-patterned and in accordance with applicable EYD. However, the accuracy value of BN-grams is still lower compared to Doc-p in P2 (April) and P-3. This occurs because the datasets used in P- 2 (July) and P-1 still incorporate stemming, resulting in trending topics that do not adhere to the Indonesian Language Guidelines (EYD) and are difficult to comprehend, thus yielding trending topics of inferior quality compared to datasets processed without stemming.

Figure 3 illustrates the comparison of topic recall between the BN-grams and Doc-p methods. Overall, the topic recall in P-4 has a higher accuracy value. Figure 4 shows that the topic recall value for BN-Grams in P-1 is higher than that in P-2 across all methods. The highest accuracy values for BN-grams in P-1 and P-2 are 43% and 28% respectively. Meanwhile, in P-4, BN-Gram and Doc- p have accuracies of 50% and 75% respectively. This is because the use of stemming in the preprocessing stage in P-2 deteriorates the quality of trending topics, resulting in a lower accuracy compared to P-1.

Testing on P-2 (April) and P-3 indicates that the Doc- p method produces several topics with high similarity to the trending topics on Twitter, online news, and local news. This occurs because the Doc-p method employs a dataset with preprocessing without stemming. Preprocessing without stemming impacts the quality of the generated trending topic sentences in accordance with the prevailing Indonesian language rules (EYD).

b. The effect of the number of topics on topic recall.

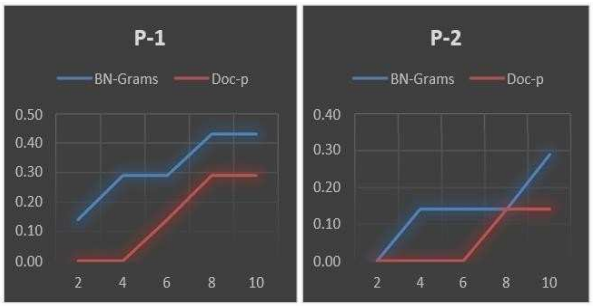

Fig 4. The Effect of Varying Topic Numbers on the Comparison of BN- grams and Doc-p

Figure 4 explains the topic recall comparison between the BN-grams and Doc-p methods [1]. Overall, the topic recall on P-1 has a higher accuracy value than P-2. The first picture shows the BN-Grams P-1 topic recall value is higher than P-2 of the entire method. The highest accuracy values for BN-grams in P-1 and P-2 are 43% and 29% respectively. This happens because the use of stemming in the preprocessing stage in P-2 worsens the quality of trending topics so that the accuracy becomes lower than P-1.

The larger the number of topics tested in BN-grams and Doc-p, the higher the impact on the increase in the topic recall accuracy value. This is evidenced in P-1 and P-2 with the number of topics ranging from 2 to 10 topics, the topic recall accuracy value of BN-grams has increased compared to Doc-p. This happens because the greater the number of topics tested, the more topics that have similarities with popular news on cyber media. Therefore, the highest topic recall values of BN-grams on P-2 and Doc-p on P-1 and P-2 are generated rom the number of topics with the highest value, namely with the number of topics 10.

The clustering in the BN-grams method uses hierarchical group average clustering. This is different from non-hierarchical K-means clustering. The topic modeling in BN-grams uses an approach of clustering between N-grams. This is different from Latent Dirichlet Allocation (LDA), where the topic modeling is based on probability of terms/words.

V. CONCLUSION

Overall, the quality of trending topic detection in Indonesian-language tweets is influenced by the type of dataset and preprocessing employed. Datasets with a more heterogeneous term distribution and using stemming tend to reduce the accuracy of trending topics generated by BN- grams and Doc-p. In addition to being

Authorized licensed use limited to: Institut Teknologi Sepuluh Nopember. Downloaded on March 26,2024 at 13:42:05 UTC from IEEE Xplore. Restrictions apply.

influenced by the type of dataset and preprocessing, the accuracy of trending topics is affected by the number of topics tested.

An increase in the number of topics results in an improvement in the accuracy of trending topics for BN-grams and Doc-p. This occurs because the addition of more topics impacts the number of keywords and sentences detected consistently with the ground truth, thereby enhancing the accuracy value.

Overall, the testing results indicate that the trending topic detection in Indonesian-language tweets from Doc-p demonstrates better accuracy compared to BN-grams. This occurs because the cluster formation principle in Document Pivot, using Locality Sensitive Hashing, exhibits better quality compared to the group average hierarchical clustering in the cluster formation of BN-grams. Therefore, the trending topics generated in each cluster in Doc-p increasingly exhibit similarities with popular online media news.

Opportunities for further research could include incorporating variations in the number of topics and the number of N-Grams in trending topic testing, as well as conducting tests using languages other than Indonesian tweets.

REFERENCES

[1] Aiello, L.M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., Göker, A., Kompatsiaris, I., Member, S., IEEE, Jaimes, A., 2013. Sensing Trending Topics in Twitter. IEEE Transactions on Multimedia (2013) 15, 1268–1282.

[2] C. Comito, "Social Media Mining and Analysis to support authorities in COVID-19 pandemic preparedness," 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Las Vegas, NV, USA, 2022, pp. 2779-2783, doi: 10.1109/BIBM55620.2022.9994914.

[3] Chen, J., Yu, J., Shen, Y., 2012. Towards Topic Trend Prediction on a Topic Evolution Model with Social Connection. 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology 153–157. doi:10.1109/WI-IAT.2012.31

[4] I. Annamoradnejad and J. Habibi, "A Comprehensive Analysis of Twitter Trending Topics," *2019 5th Int. Conf. Web Res. ICWR 2019*, pp. 22–27, 2019, doi: 10.1109/ICWR.2019.8765252.

[5] Indra, Winarko, E., Pulungan, R., 2018. Trending topics detection of Indonesian tweets using BN-grams and Doc-p. Journal of King Saud University - Computer and Information Sciences 1–9. doi:10.1016/j.jksuci.2018.01.005

[6] Kaleel, S.B., Abhari, A., 2015. Cluster-discovery of Twitter messages for event detection and trending. J. Comput. Sci. 6, 47–57. https://doi.org/10.1016/j. jocs.2014.11.004

[7] M. Asgari-Chenaghlu, N. Nikzad-Khasmakhi, and S. Minaee, "Covid-Transformer: Detecting COVID-19 Trending Topics on Twitter Using Universal Sentence Encoder," 2020, [Online]. Available: http://arxiv.org/abs/2009.03947

[8] P. Wicke and M. M. Bolognesi, "Covid-19 Discourse on Twitter: How the Topics, Sentiments, Subjectivity, and Figurative Frames Changed over Time," *Front. Commun.*, vol. 6, no. March, pp. 1–20, 2021, doi: 10.3389/fcomm.2021.651997.

[9] Petkos, G., Papadopoulos, S., Kompatsiaris, Y., 2014. Two-level Message Clustering for Topic Detection in Twitter., in: Social News on The Web (SNOW) 2014 Data Challenge. http://ceur- ws.org, Seoul, Korea, pp. 49–56.

[10] Petrović, Saša, Osborne, M., & Lavrenko, V. (2010). Streaming first story detection with application to twitter. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 181–189.

[11] Riquelme, F., González-cantergiani, P., 2016. Measuring User influence on Twitter : A survey 52, 949–975.

[12] Sakaki, T., Okazaki, M. and Matsuo, Y. (2013) 'Tweet analysis for real-time event detection and earthquake reporting system development', *IEEE Transactions on Knowledge and Data Engineering*, 25(4), pp. 919–931. Available at: https://doi.org/10.1109/TKDE.2012.29.

[13] Sreenivasulu, M. and Sridevi, M. (2020) 'Comparative study of statistical featuresto detect the target event during disaster', *Big Data Mining and Analytics*, 3(2), pp. 121–130. Available at: https://doi.org/10.26599/BDMA.2019.9020021.

Authorized licensed use limited to: Institut Teknologi Sepuluh Nopember. Downloaded on March 26,2024 at 13:42:05 UTC from IEEE Xplore. Restrictions apply.