Corpus Analysis: A Case Study on Kadazandusun Newspaper Archive

Mohd Shamrie Sainin
Knowledge Technology Research Unit,
Faculty of Computing and Informatics,
Universiti Malaysia Sabah, 88400 Kota
Kinabalu Sabah, Malaysia
shamrie@ums.edu.my

Asni Tahir
Knowledge Technology Research Unit,
Faculty of Computing and Informatics,
Universiti Malaysia Sabah, 88400 Kota
Kinabalu Sabah, Malaysia
asnieta@ums.edu.my

Suraya Alias
Knowledge Technology Research Unit,
Faculty of Computing and Informatics,
Universiti Malaysia Sabah, 88400 Kota
Kinabalu Sabah, Malaysia
suealias@ums.edu.my

***Abstract*—This paper presents the analysis of text data acquired from News Sabah Times, which is the only Sabah's newspaper that has a section for Kadazandusun news. Currently, there is no text translation tool available to translate a sentence or large text specifically from Kadazandusun to other languages such as Bahasa Melayu. Thus, the first step is to develop such a system is to analyze the available corpus from the newspaper archive. The objective is to perform text analysis and then providing the possible ways of utilizing the knowledge. In this work, the purpose is to report the methodology and utilization of the fundamental corpus analysis related to text mining and not covering on the linguistics aspects and grammatical context. In addition, this paper also reports on the findings from the newspaper corpus analysis.**

Keywords—Kadazandusun, corpus analysis, text mining, text analysis

I. INTRODUCTION

According to Cambridge dictionary, a corpus (singular of corpora) is a collection of written or spoken material stored on a computer and can be used to find out how language is used [1]. In other words, a corpus can be defined as a collection of pieces of language, selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language [2]. The scope of the corpus in this paper is referring to the collection of Kadazandusun text articles sections which publicly available in the source news website. As mentioned in the abstract, there is less work has been done on the language specifically on their text analysis [3]. Thus, preliminary corpus analysis is carried prior to this paper to explore the underlying concept and how the language is being used in the news.

The term "corpus analysis" is not clearly being discussed as specific term to refer the processes involved to carry out the analysis and to find out the underlying concepts, statistics, or even the visualization of the analysis. Most of literatures that relates the topic of corpus analysis are associated with the term "linguistics". The corpus linguistics discipline has settled down to its core focus on language [4], where corpus analysis can be done using available linguistics tool such as WordSmith [5] and AntConc [6].

In this paper, however, does not focus on the linguistics but as motivation and as a case study that offers the experience on the corpus analysis to other researchers in the similar domain. The discussion inclined towards part of text mining domain which is to explore the significant knowledge and representation which can be used in relation to data science and knowledge discovery. The analysis reported in this paper is carried out using tools such as R and Python, where basic analysis methods are applied. The process of the work that has been done is described in the following sections with results of the analysis..

II. DATA COLLECTION

Preliminary corpus analysis in Kadazandusun newspaper contains news text acquired from online archive of New Sabah Times. The following Table 1 is the broad details of the corpus.

TABLE I. CORPUS DETAIL

Information	Description
Source	New Sabah Times Kadazandusun Section (Dusun News Archives)
Website	http://www.newsabahtimes.com.my/nstweb/category/Kadazan+Dusun
Duration	1/1/2018 – 12/7/2019
Total document	591
Total words (tokens)	186569
Average words per document	315
No of sentence	13641
Average sentence length	23
Total terms (unique)	14531

A. Text Preprocessing

Before corpus analysis can be done, text preprocessing is the key component for any related text analysis. In text preprocessing, several steps that can be performed are tokenization, filtering, lemmatization and stemming [7]. Specifically, the text preprocessing steps that has been applied prior to the analysis is depicted in Fig. 1.

Fig. 1. Preprocessing and text features analysis.

In the filtering phase, it consists of removing numbers and other alphanumeric leaving only the words token in the document. This process is presented in the Table II

978-1-7281-4082-7 /19/$31.00 ©2019 IEEE

pseudocode snippets and description and filtering sample in Table III.

TABLE II. FILTERING STAGE

Information	Description
toLowerCase	Make the whole text as lower case
replaceAll("[^-a-z'A-Z']"," ")	Remove numbers and other alpha numeric characters except single quote and dash
text <- str_replace_all(text, "(^')", "") text <- str_replace_all(text, "(\')", " ")	Remove single quote in front of word

TABLE III. FILTERING SAMPLE

Raw Text	Filtered Text
"Ounsikou yahai do hiti om gumuli' miampai lobi kogumu tambalut do timpu-timpu dumontol," ka disio. Minaganu nogi yolo' ngaawi' do kosiwatan minumbal mangakan taakanon kampung i aiso' abantalan id pomogunan diolo'.	ounsikou yahai do hiti om gumuli' miampai lobi kogumu tambalut do timpu-timpu dumontol ka disio minaganu nogi yolo' ngaawi' do kosiwatan minumbal mangakan taakanon kampung i aiso' abantalan id pomogunan diolo'
Pulis Marin minagahou 334 katon tinumon kaawuk kigatang RM175,940.	pulis marin minagahou katon tinumon kaawuk kigatang rm

Tokenization is simply an extraction of all tokens from the corpus and followed by the process to identify token count, unique tokens, frequency, concordance sample, TF-IDF, N-Gram and Frequent Adjacent Sequential Pattern. The results are derived from different tools such as R and Python text mining package, WordSmith 6.0 Tool and AntConc 3.5.8.

B. Description of Analysis

The should be many possible analysis dimensions for the data, however as mentioned in Fig. 1, there are seven aspects that this paper consider as interesting findings. Token count is the number of words from all documents after filtering step is done. It contains all words or combinations of characters as string whether it is a word from Kadazandusun, Malay, English or other string. Unique tokens are computed from all tokens where repeated tokens are removed. Then, frequency can be calculated based on the number of times those unique tokens appear in all documents. Frequency of top ten terms in several categories are reported (unique tokens, local district, local villages, world places, and names. Sample concordance sample from the perspective of linguistics is also briefly presented.

The TF-IDF of Term Frequency – Inverse Document Frequency is a document representation method that used in text mining task specifically for document classification and has become one of the most common methods [8]. TF-IDF provides the word with a weight using frequency of its usage in a document and the infrequency of its occurrence in the entire corpus. Finally, the N-gram is presented in this case study that it is a continuous sequence of words in a text. It analyzes the relationship between words that tend to follow others or co-occur within a document [9]. Another text representation based on Sequential Pattern-Growth technique is also briefly discussed based on the study from [10].

III. ANALYSIS DIMENSIONS

As discussed in previous section, the selected analysis provides an insight of the corpus that may interest or useful for other investigation and application development. In this analysis, several categories of term were reviewed and presented. The following subtopics are derived term from the top order in frequency in the corpus.

A. Unique words

The first top highest frequency of the unique words from the selected corpus as shown in Table IV, below.

TABLE IV. TOP TEN HIGHEST FREQUENCY UNIQUE WORDS

Words	Frequency
do	8354
i	5662
di	5473
om	4786
id	4747
dilo	2554
disio	1859
ka	1736
montok	1642
diti	1629

The unique words in Table 4 shows that those are the most common words in Kadazandusun language specifically in written text. As an example, 'do' is an example of grammatical marker (particle) of a complement type and 'id' is a prepositional as word connector [11]. Examples of sentences using those top ten unique words can be found in further subtopics.

B. Unique words with dash

Repeated words or partial utterances (known as 'kata ganda') in Kadazandusun has strong similarity to Bahasa Malaysia. Table V below shows top ten such words from the selected corpus.

TABLE V. TOP TEN REPEATED WORDS WITH DASH

Words	Frequency
talun-alun	90
sinding-sinding	49
nunu-nunu	45
tanak-anak	30
iman-imanon	23
nombo-nombo	23
iso-iso'	20
hombo-hombo	18

Words	Frequency
timpu-timpu	15
monguni-uni	12

C. *Unique words with quote (apostrophe) as last character*

An apostrophe as part of the word in Kadazandusun has an important contribution to pronunciation as well as in writing. Table VI shows the top usage of such words.

TABLE VI. TOP TEN WORDS WITH APOSTROPHE

Words	Frequency
dilo'	755
nogi'	420
nga'	340
tongo'	291
diolo'	240
amu'	141
iso'	138
ko'	121
ngaawi'	98
ogumu'	87

D. *Unique words of places*

In this topic, list of local district, local villages and countries are examined from set of terms from predefined keywords. Table VII and VIII shows the results.

TABLE VII. TOP TEN LOCAL PLACES

Local District	Frequency	Local Villages	Frequency
kinabalu	310	kampung molisau	7
kota kinabalu	275	kampung bongkol	6
labuan	253	kampung landung	6
keningau	182	kampung mangaris	6
ranau	174	kampung tambatuon	6
sandakan	152	kampung tanjung	6
tamparuli	117	kampung kapa	5
tawau	92	kampung lapai	5
tuaran	86	kampung tenghilan	5
kota marudu	75	kampung tinusa	5

It is clearly shown that ‘Kinabalu’ as single word is mostly the highest frequency and “Kota Kinabalu” is expected as more news related to the locality of the newspaper. Term ‘China’ is also a prevalence that the country has an important relation within the newspaper topics. Interestingly, Kampung Molisau in Tenghilan was mentioned more than the other out of 205 villages appear in the selected news corpus.

TABLE VIII. TOP TEN COUNTRIES

Country	Frequency
china	30
filipina	29
brunei	14
amerika	9
korea	9
australia	6
singapura	6
thailand	5
jepun	3
london	2

E. *Unique words of names*

Names appeared in the news is another interesting revelation where from 143 single names, Table IX shows the top ten names.

TABLE IX. TOP TEN NAMES

Country	Frequency
mohd	207
ewon	89
robert	74
shafie	74
james	71
ahmad	68
raphiel	64
abdul	56
joanes	54
mohamad	54

According to table IX, ‘mohd’ is a popular name which is the first name and normally combined with other names. Example of specific names which appears in the selected news corpus are Mohd Haniff, Mohd Ismi and Mohd Zain. The second ranking of names is ‘Ewon’ which mainly refer to Datuk Ewon Benedick who is the current State Rural Development Minister of Sabah state government at the time of this paper is written.

F. *Unique words not in dictionary*

Words that were not found means that the word does not exist in Kadazandusun dictionaries or the spelling is wrong. In this case study, the dictionaries; [12] and [13] were referred to validate the spelling of the words as well as several other Kadazandusun literatures in [14], [15][16], [17]. Some of the words are listed in Table X.

TABLE X. SAMPLE LIST OF WORDS NOT IN DICTIONARY

Word	Freq
dilo	2554
ababayan	979
diolo	535
amu	493
ilo	360
iso	257
karaja	217
aiso	200
babaino	177
boyoon	66

Based on Table X, it can be seen that quite a number of words that either not in dictionary or has spelling problem. As an example, 'dilo' ('that' in English) is mostly used in the selected news corpus where the news writer omits the apostrophe at the end of the word as listed in [12]. In contradiction, 'dilo' without an apostrophe is used in literatures such as [18] and [16]. This problem is due to the fact that there is no complete dictionary yet available although few already available as dictionary book. Furthermore, there is no text processing application with dictionary built-in. Thus, it is an important future works to enable complete dictionary or Kadazandusun corpus so that the community could utilize the resources. The possible example is a language pack which can be developed in text processing tool such as Microsoft Word.

G. TF-IDF

The main reason why TF-IDF included in this analysis is to explore words importance to a document in the selected new corpus, which also supports natural language processing and information retrieval domain. This representation is defined as:

$$\text{TF-IDF}_{ij} = tf_{ij} \times log\left\{\frac{N}{df_i + 1}\right\} \quad (1)$$

where tf_{ij} is the Term Frequency and $log\{N/df_i + 1\}$ is the Inverse Document Frequency [8]. Top score of the TF-IDF based on the selected news corpus is presented in Table XI using R and TF-IDF package, while Table XII shows the sample of lowest TF-IDF score.

Based on the result of TF-IDF, tf_idf score will be high when the word appeared in a single document and become low if the word occurs in large numbers of documents. The tf_idf score given word 'malaria' is the highest because it is only existed in one document. Meanwhile, word 'do' is among the lowest score because it appears in all documents. In text mining, these scores can be utilized for document classification [19] or topic modelling [20]. For example, document '1269.txt', '1452.txt' and '1078.txt' can be classified in similar class such as 'Animal'.

TABLE XI. HIGHEST TF-IDF OF THE WORDS GIVEN THE DOCUMENT

Document	word	n	tf	idf	tf_idf
1095.txt	malaria	15	0.0568	6.39	0.363
1378.txt	dodopiton	15	0.0577	5.29	0.305
1269.txt	buayo	10	0.0535	5.7	0.305
1123.txt	peka	12	0.043	6.39	0.275
1078.txt	buayo	9	0.0464	5.7	0.264
1302.txt	mesej	23	0.0545	4.78	0.261
1388.txt	noos	10	0.0488	5.29	0.258
1464.txt	sesb	9	0.0559	4.6	0.257
1452.txt	tapir	16	0.0402	6.39	0.257
1440.txt	sikaut	16	0.0479	5.29	0.253

TABLE XII. LOWEST TF-IDF OF THE WORDS GIVEN THE DOCUMENT

Document	word	n	tf	idf	tf_idf
1042.txt	do	12	0.0591	0.00169	0.0000999
1042.txt	i	12	0.0591	0.00169	0.0000999
1228.txt	dilo	2	0.0118	0.00848	0.0000998
1079.txt	i	19	0.059	0.00169	0.0000998
2170.txt	id	4	0.0147	0.00678	0.0000997
1169.txt	id	3	0.0147	0.00678	0.0000997
1278.txt	di	13	0.0294	0.00338	0.0000995
2098.txt	di	6	0.0294	0.00338	0.0000995
1469.txt	do	18	0.0588	0.00169	0.0000994
1477.txt	do	12	0.0588	0.00169	0.0000994

Using the tf-idf score from Python 'sklearn' (based on Non-Negative Matrix Factorization), sample of 6 topic models are shown in Table XIII below with the inferring topics. Inferring topics are manually assigned based on the available keywords which may not be the best topic, however it is the nearest assignment considering the combination and meaning of the keywords.

TABLE XIII. SAMPLE OF TOPIC MODEL

No.	Keywords	Topic
Topic 1	disio nogi sabah ka diti montok ababayan nga pogun koporintaan	Governance
Topic 2	bot marin pulis maritim semporna kusai kowoigan jaam sada tinahan	Crime
Topic 3	sinding dau album artis ku fh production suminding kaganaan studio	Entertainment
Topic 4	sikul susumikul mongiingia sk molohing ababayan molohingon palajaran tenghilan smk	Education
Topic 5	raha kimpin monoluod hospital talasemia pongubatan beaufort susumakit keningau alumni	Health
Topic 6	dadah pulis tinahan syabu notondosan kusai gram operasi tulun habibi	Crime

H. N-Gram

One of the most popular text mining approaches is based on N-Gram which also called as 'bag of words' [21]. There are normally two N-Grams representations which can be derived from a corpus analysis; unigram and bigrams. Unigram is a variable containing occurrences of single words in each document. Two-words sequence is a variable called bigrams. The list of words found in Table IV are also the unigram variables translated into total occurrences on all documents. The result of corpus analysis using bigrams (two consecutive words) based on documents is presented in Table XIV.

TABLE XIV. LIST OF BIGRAMS (N=2) PER-DOCUMENT

Document	Word 1	Word 2	Frequency
2104.txt	sabah	tea	13
1046.txt	monoluod	raha	12
1123.txt	peka	b40	12
1352.txt	labus	kakadayan	12
2167.txt	malim	gunung	12
1276.txt	ukuman	matai	11
1452.txt	tapir	malaya	11
0998.txt	labus	kakadayan	10
1081.txt	pingludaan	sangod	10
1298.txt	batu	sumpah	10

The occurrence of the two words is quite low considering the word average in a document is about 315. However, it is interesting to find out that "sabah tea" and "monulud raha" (blood donation) are among the important words based on bigrams. Furthermore, bigrams can also be computed to examine the two words that came together in corpus. The bigrams based on all documents are shown in Table XV.

TABLE XV. LIST OF BIGRAMS (N=2) BASED ON ALL DOCUMENTS

Word 1	Word 2	Frequency
ka	disio	1245
nopo	nga	696
po	nogi	685
di	tongo	430
ka	di	383
do	hiti	372
dilo	i	350
nogi	do	338
suai	ko	336
kota	kinabalu	293

The word 'ka disio' ('he said' – referring to masculine gender) is mostly used in the selected corpus with 1245 occurrences to address third-person speaking. Based on Table XV, as an example in reference to [22] and [18], the words are type of particle determiners ('i' – definite pivot, 'di' – definite non-pivot, 'do' – indefinite non-pivot), dietetic adverbs ('do hiti' – near speaker non-pivot) and particles common parts of speech ('no', 'po', 'nopo', 'nga', 'nopo nga').

I. Text representation model using Sequential Pattern-Growth

While Bags-of-Words (BOW) and N-Gram model are popular text representation in text mining, recent study investigating the use of pattern-based model called Frequent Adjacent Sequential Pattern or FASP was presented in [10]. Raising the issue of inaccurate semantic representation of text using BOW model and high dimensionality using N-Gram model, FASP was proposed to represent the text using a set of frequent sequence of adjacent word discovered in the document. FASP can be used to represent the textual pattern similarity between documents and the result can be converted to a set of rules for describing the main event in the news. However, notable setting from the study was on the news domain collection, where the English and Malay news dataset focuses on natural disasters and events tragedy. In contrast, the Kadazandusun news collection corpus in this study explores very high variety of domains/topics i.e. criminology, education, entertainment and others.

The purpose of applying the FASP algorithm in this paper is to explore the word representation in comparison with N-Gram top word generation. Table XVI shows the words with the highest normalized FASP score (support) for one sequence and Table XVII provides the sample FASP score given the word 'om' ('and' in English). Based on the result, single word 'om' with score 0.8 is the most frequent word and the frequent adjacent pattern as displayed in Table XVII using the word 'om' across all documents. Compared to the results in Table XV, N-Gram represents the frequent combination of words from the whole documents, however most of the words found are type of particles with non-descriptive representation. Exception on the word combination of 'kota kinabalu' where it is found that Kota Kinabalu has the probability that the news is discussing or take place in this location. The list of words form N-Gram model can be considered as stop words for Kadazandusun language (future work should be conducted to support this findings).

FASP on the other hand, although that stopwords is not applied (because no available stopwords at this point for Kadazandusun corpus), the algorithm is able to produce few possible insights and descriptive representation such as 'mogiigiyon' (resident/community), 'kobolingkahangan' (problem) and 'kogoogonopan' (the needs) as important words that may explain the corpus. Further study on FASP algorithm implementation and stopword for Kadazandusun is required to find more meaningful word representation by the FASP or other representation methods is very much required in future.

TABLE XVI. FASP SCORE FOR ONE WORD SEQUENCE

Word	Frequency	FASP score
om	1231	0.815231788
montok	732	0.484768212
kumaa	546	0.361589404
diolo	413	0.273509934
hiti	376	0.249006623
mogiigiyon	332	0.21986755
sabaagi	295	0.195364238
mogisusuai	190	0.125827815
kobolingkahangan	90	0.059602649
kogoogonopan	79	0.052317881

TABLE XVII. FASP SCORE FOR TWO OR THREE SEQUENCE GIVEN THE WORD 'OM'

Word	Frequency	FASP score
mogiigiyon om	27	0.017880795
om mogiigiyon	20	0.013245033
om montok	18	0.01192053

om mogisusuai	11	0.007284768
om kogoogonopan	10	0.006622517
kogoogonopan om	8	0.005298013
om kumaa	7	0.004635762
hiti om	6	0.00397351
om kobolingkahangan	4	0.002649007
kumaa mogiigiyon om	2	0.001324503

IV. CONSLUSION

Earlier, this paper described the approach in analysing the selected corpus of Kadazandusun news archive from New Sabah Times. According to the preprocessing and feature extraction, several dimensions were presented. The top tens of several categories as frequent words, repeated words with dash, words with apostrophe, places, names and not in dictionary or wrong spelling words were listed and discussed. Then, followed by presenting the significant words with TF-IDF, N-Gram and FASP score were also deliberated. In line with this case study exploration, the insight on the potential richness in knowledge discovery acquired from the corpus. What more if more news archive is added to the collection that until the very last document in the exercise, there were still new words added to the list of words. Though, this case study is still lacking on other analysis approaches that were not discussed in depth or considered such as linguistics analysis (not the focus of this paper as none of the researchers are expert in the field), named entity parsing, part-of-speech tagging and Kadazandusun stopwords. Due to limited resources for validating the words other than the available dictionaries as mentioned before and the scale of the corpora, the integrity may draw questions and critiques from the expert. However, this paper draws the attention that this is a case study to spark more discussions and possible caution to discuss the results related to the noise that may influence the conclusion instigating from the corpus data.

ACKNOWLEDGMENT

This study is partly supported by UMS Innovation Grant Scheme (SGI0063-2018).

REFERENCES

[1] Corpus, "Cambridge online dictionary," *Cambridge Dictionary online*, 2019. [Online]. Available: https://dictionary.cambridge.org/dictionary/english/corpus. [Accessed: 29-Aug-2019].

[2] J. Sinclair, "Preliminary Recommendations on Corpus Typology," *EAGLES Document TCWG-CTYP/P*, 1996. [Online]. Available: http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html. [Accessed: 28-Sep-2019].

[3] A. Omar, "Processing Malaysian Indigenous Languages: A Focus on Phonology and Grammar," *Open J. Mod. Linguist.*, vol. 4, pp. 728–738, 2014.

[4] E. Vaughan and A. O'Keefee, "Corpus Analysis.," in *The International Encyclopedia of Language and Social Interaction, First Edition*, C. I. and T. S. (Associate E. Karen Tracy (General Editor), Ed. John Wiley & Sons, Inc., 2015.

[5] M. Scott, "WordSmith Tools version 7." Stroud: Lexical Analysis Software, 2016.

[6] L. Anthony, "AntConc 3.4.1." 2014.

[7] M. Allahyari *et al.*, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," *KDD Bigdas*, 2017.

[8] D. Kim, D. Seo, S. Cho, and P. Kang, "Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec," *Inf. Sci. (Ny).*, vol. 477, pp. 15–29, Mar. 2019.

[9] D. Robinson and J. Silg, *Text Mining with R*. O'Reilly Media, Inc., 2017.

[10] S. Alias, S. K. Mohammad, G. K. Hoon, and T. T. Ping, "A text representation model using Sequential Pattern-Growth method," *Pattern Anal. Appl.*, vol. 21, no. 1, pp. 233–247, 2018.

[11] G. Pius, "Introduction to the Kadazandusun Prepositional Phrase," 2017. [Online]. Available: https://www.researchgate.net/publication/316029510_INTRODUCTION_TO_THE_KADAZANDUSUN_PREPOSITIONAL_PHRASE_PENGENALAN_KEPADA_FRASA_SENDI_NAMA_BAHASA_KADAZANDUSUN. [Accessed: 03-Sep-2019].

[12] *Daftar Kata Bahasa Kadazandusun - Bahasa Malaysia*. Kadazandusun Language Foundation (KLF), 2015.

[13] *Kamus Malay-Dusun-English*. 2015.

[14] M. Sintian, "Kadaaton Tinaru Kadazan Dusun: Koubasanan om Kotumbayaan di Mongowit Korutumon?," in *Seminar Bahasa, Kesusasteraan dan Kebudayaan Kadazandusun 2013*, 2013.

[15] M. Sintian, "Kogingohon Boros Kadazandusun Tomposio Tokou / Martabatkan Keindahan Bahasa Kadazandusun," in *Seminar Bahasa, Kesusasteraan dan Kebudayaan Kadazandusun*, 2014.

[16] M. Sintian, "Boros Kadazandusun: Pogiroto' Piagalan Pantango' Pisuayan," in *Seminar Kebangsaan Budaya, Bahasa dan Sastera Kadazandusun (MAKADUS)*, 2018.

[17] M. Sintian, "Kopointutunan Sintaksis Boros Kadazandusun (Pengenalan Sintaksis Bahasa Kadazandusun)," 2015.

[18] *Puralan boros Kadazandusun id sikul*. Putrajaya: Bahagian Pembangunan Kurikulum, Kementerian Pelajaran Malaysi, 2008.

[19] B. Das and S. Chakraborty, "An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation," *arXiv Prepr.*, vol. arXiv:1806, 2018.

[20] G. Zhao, Y. Liu, W. Zhang, and Y. Wang, "TFIDF Based Feature Words Extraction and Topic Modeling for Short Text," in *Proceedings of the 2018 2Nd International Conference on Management Engineering, Software Engineering and Service Sciences*, 2018, pp. 188–191.

[21] M. Schonlau, N. Guenther, and I. Sucholutsky, "Text mining with n-gram variables," *Stata J.*, vol. 17, no. 4, pp. 866–881, Dec. 2017.

[22] D. C. Price, "Bundu Dusun Sketch Grammar," 2007. [Online]. Available: http://blog.thetelegraphic.com/wp-content/uploads/2008/12/bundu-dusunsketch-grammar.pdf. [Accessed: 28-Jul-2019].