IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Effects of Pre-processing Phases in Sentiment Analysis for Malayalam Language

Deepa Mary Mathews[1*], Sajimon Abraham[2]

[1]Research Scholar, Mahatma Gandhi University, Kottayam, India
[2] School of Management and Business Studies, Mahatma Gandhi University, Kottayam, India

[*]Corresponding Author: *deepamarymathews@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 20/July/2018, Published: 31/July/2018

***Abstract*—** Over the last few years, the generation of computerized information has increased exponentially. Most people use digital media to share news and their views on a topic. To analyze this outsized web information, new analytical techniques are required which automatically portrays the data open on the Web. Most of us are more comfortable in expressing our viewpoints and outlooks in Mother tongue. Sentiments of the social users on various topics expressed in their own mother tongue leads to the necessity of mining the sentiments in various dialects. In fact, some data do not have an effect on the classification result even removing them and some carries similar meanings, therefore a pre-processing phase has to accomplish and thus the dataset can be more precise. In this paper, the authors are focusing on pre-processing the words given by the user through their reviews in the social networking sites expressed in Malayalam language. The authors calculated the reduction in word count after performing the preprocessing processes and the experiments shows that more than 20% of word count reduction occurred.

***Keywords*—** Opinion Mining, POS Tagging, Stemming, Stopword Removal, Malayalam

I. INTRODUCTION

Since English is a universal language, many research works in Opinion mining is mostly done in English dialect. Most of the social users are more comfort in expressing our views and opinions in their own local dialect. In India, 30 official local languages are there and a great many individuals convey viably in their local dialect hence creating huge amount of information. These information can well be routed to extract many valuable patterns like the customers' buying pattern, product feedback, and so on at a regional level. The aforementioned work is lagging in these local dialects. Standpoints of the social users on diverse topics communicated in their own mother tongue leads to the necessity of mining the sentiments in various dialects. More than 35 million people spreading along the regions of Kerala, Pondicherry and Lakshadweep are using this Malayalam.

The article is framed as follows. The subsequent section explains the works related to the study followed by the section which explains the various methodologies used to implement the work. The implementation section portrayed the framework of the proposed work and in the Results and Discussion section the results are analyzed which is followed by the Conclusion section which concludes the work.

II. RELATED WORK

Various works done in Indian dialects are - Kannada Morphology Analyzer is brought in by Shambhavi et.al in 2001 [1]. A Stemmer for Hindi was proposed by Ramanathan et al. [2]. Sentiment analysis on Punjabi News Articles using Support Vector Machine is done by Gagandeep Kaur in [3].

A stemmer for Bengali language is introduced by Khan et al. in [4]. An Urdu stemmer was proposed by QuratUlAin et al. in [5]. In 2013, Dutta, P. K., introduced an online POS tagging method for Assamese [6]. In 2014, Kasthuri, M., & Kumar, S. developed a rule based stemmer for Tamil language[7]. So many works have been proposed for Sentiment Analysis in Universal Languages like English, although it is comparatively less for Malayalam. Prajitha, U et.al in the year 2013 introduced a light weight Malayalam stemmer called LALITHA [8] and Pragisha et.al developed a stemmer called STHREE [9]. Jisha P Jayan et.al in the year 2013 demonstrates that TnT and rule based suit combination is better for Malayalam [10]. In the year 2014, Deepu S Nair in his paper, proposed a rule based approach for sentiment analysis from Malayalam movie reviews [11]. Manju K, et al stated that they face a lot of difficulties while dealing with Malayalam because of the inflectional and morphological variations of the language [12]. Renjith S R, Sony P in their

paper, used page ranking method for ranking the sentences in the document [13].

III. ROLE OF PREPROCESSING IN OPINION MINING

Opinion Mining intends to computationally recognize and extricate subjective data in the content. The polarity of the text can be either positive or neutral or negative. This process is extremely useful in social media monitoring to ascertain mood as happy/ angry/ sad/ excited. The pre-processing phase of the sentiment analysis process reduces the size of opinionated text and thereby enhances performance. This step is required to speed up the process while dealing with the large corpora. The pre-processing phase may consist of Tokenization, POS Tagging, Stopword Removal, Stemming and Lemmatization. Even though these normalization processes may introduce noise, it is often done to reduce the size of the corpora and thereby to simplify the analysis.

IV. DATASET

The user comments from various Malayalam online news websites like www.deepika.com, www.mathrubhumi.com, www.manoramaonline.com etc related to the News Title "സംസ്കൃതം പഠിച്ചാല് ബുദ്ധി കൂടുമെന്ന് അവകാശപ്പെട്ട് ഗവേഷകന്....." are collected for the experimentation using beautiful soup method. The dataset contains comments in English language also. The various pre-processing phases are required to clean the dataset before proceeding to sentiment analysis. A sample of reviews collected is shown in the figure 1.

V. IMPLEMENTATION

The preprocessing phase is primarily dictated by the objective of the problem. This phase in the sentiment analysis process reduces the size of opinionated text and thereby augments performance. This step is required to pace up the process while dealing with the large corpora which considerably reduces the size of the corpora. The overview of the preprocessing steps is shown in figure 2. The various preprocessing process in the case of text/opinion analytics consists of

i) Bilingual dataset to Unilingual Dataset,
ii) POS Tagging,
iii) Stopwords Removal
iv) Stemming

i) Bilingual dataset to Unilingual Dataset

Some users are comfortable in both the English and Malayalam languages. They used to use a mix of linguistic words while expressing their opinion. The user reviews that we collected contain words of both Malayalam and English. So the first step is to make all the reviews into a single language, here Malayalam. Using Google translate package in the Python, we converted the bilingual dataset into unilingual dataset.

ii) POS Tagging

This method spot out a term in the review as relating to a specific grammatical form, based on definition and the context [6]. It identifies the category of the word. Tagging is based on the relationship of the word with adjacent and related words in a phrase or sentence by considering its

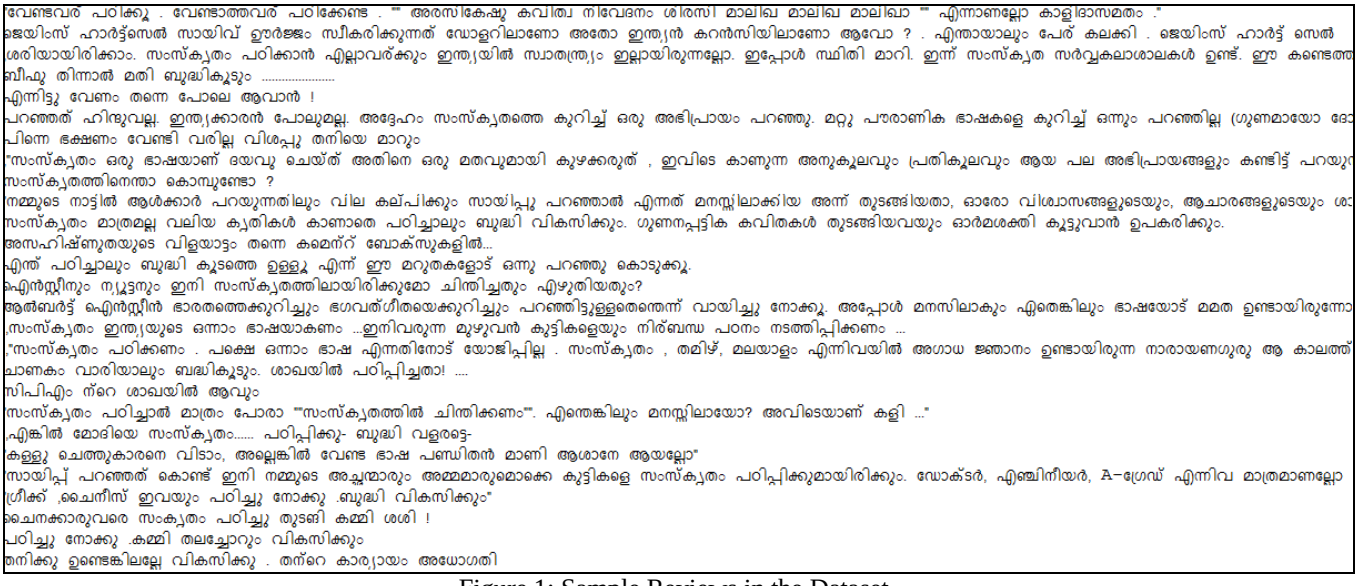
വേണ്ടവർ പഠിക്കൂ . വേണ്ടാത്തവർ പഠിക്കേണ്ട . "" അരസികേഷു കവിത്വ നിവേദനം ശിരസി മാലിഖ മാലിഖ മാലിഖാ "" എന്നാണല്ലോ കാളിദാസമതം ."
ജെയിംസ് ഹാർട്ട്സെൽ സായിവ് ഊർജ്ജം സ്വീകരിക്കുന്നത് ഡോളറിലാണോ അതോ ഇന്ത്യൻ കറൻസിയിലാണോ ആവോ ? . എന്തായാലും പേര് കലക്കി . ജെയിംസ് ഹാർട്ട് സെൽ
ശരിയായിരിക്കാം. സംസ്കൃതം പഠിക്കാൻ എല്ലാവർക്കും ഇന്ത്യയിൽ സ്വാതന്ത്ര്യം ഇല്ലായിരുന്നല്ലോ. ഇപ്പോൾ സ്ഥിതി മാറി. ഇന്ന് സംസ്കൃത സർവ്വകലാശാലകൾ ഉണ്ട്. ഈ കണ്ടെത്ത
ബീഫു തിന്നാൽ മതി ബുദ്ധികൂടും .....
എന്നിട്ടു വേണം തന്നെ പോലെ ആവാൻ !
പറഞ്ഞത് ഹിന്ദുവല്ല. ഇന്ത്യക്കാരൻ പോലുമല്ല. അദ്ദേഹം സംസ്കൃതത്തെ കുറിച്ച് ഒരു അഭിപ്രായം പറഞ്ഞു. മറ്റു പൗരാണിക ഭാഷകളെ കുറിച്ച് ഒന്നും പറഞ്ഞില്ല (ഗുണമായോ ദോ
പിന്നെ ഭക്ഷണം വേണ്ടി വരില്ല വിശപ്പു തനിയെ മാറും
"സംസ്കൃതം ഒരു ഭാഷയാണ് ദയവു ചെയ്ത് അതിനെ ഒരു മതവുമായി കൂഴക്കരുത് , ഇവിടെ കാണുന്ന അനുകൂലവും പ്രതികൂലവും ആയ പല അഭിപ്രായങ്ങളും കണ്ടിട്ട് പറയു
സംസ്കൃതത്തിനെന്താ കൊമ്പുണ്ടോ ?
നമ്മുടെ നാട്ടിൽ ആൾക്കാർ പറയുന്നതിലും വില കല്പിക്കും സായിപ്പു പറഞ്ഞാൽ എന്നത് മനസ്സിലാക്കിയ അന്ന് തുടങ്ങിയതാ, ഓരോ വിശ്വാസങ്ങളുടെയും, ആചാരങ്ങളുടെയും ശാ
സംസ്കൃതം മാത്രമല്ല വലിയ കൃതികൾ കാണാതെ പഠിച്ചാലും ബുദ്ധി വികസിക്കും. ഗുണനപ്പട്ടിക കവിതകൾ തുടങ്ങിയവയും ഓർമശക്തി കൂട്ടുവാൻ ഉപകരിക്കും.
അസഹിഷ്ണുതയുടെ വിളയാട്ടം തന്നെ കമെന്റ് ബോക്സുകളിൽ...
എന്ത് പഠിച്ചാലും ബുദ്ധി കൂടത്തെ ഉള്ളൂ എന്ന് ഈ മറുതകളോട് ഒന്നു പറഞ്ഞു കൊടുക്കൂ.
ഐൻസ്റ്റീനും ന്യൂട്ടനും ഇനി സംസ്കൃതത്തിലായിരിക്കുമോ ചിന്തിച്ചതും എഴുതിയതും?
ആൽബർട്ട് ഐൻസ്റ്റീൻ ഭാരതത്തെക്കുറിച്ചും ഭഗവത്ഗീതയെക്കുറിച്ചും പറഞ്ഞിട്ടുള്ളതെന്തെന്ന് വായിച്ചു നോക്കൂ. അപ്പോൾ മനസിലാകും ഏതെങ്കിലും ഭാഷയോട് മമത ഉണ്ടായിരുന്നോ
,സംസ്കൃതം ഇന്ത്യയുടെ ഒന്നാം ഭാഷയാകണം ...ഇനിവരുന്ന മുഴുവൻ കുട്ടികളെയും നിർബന്ധ പഠനം നടത്തിപ്പിക്കണം ...
,"സംസ്കൃതം പഠിക്കണം . പക്ഷെ ഒന്നാം ഭാഷ എന്നതിനോട് യോജിപ്പില്ല . സംസ്കൃതം , തമിഴ്, മലയാളം എന്നിവയിൽ അഗാധ ജ്ഞാനം ഉണ്ടായിരുന്ന നാരായണഗുരു ആ കാലത്ത്
ചാണകം വാരിയാലും ബദ്ധികൂടും. ശാഖയിൽ പഠിപ്പിച്ചതാ! ....
സിപിഎം ന്റെ ശാഖയിൽ ആവും
സംസ്കൃതം പഠിച്ചാൽ മാത്രം പോരാ ""സംസ്കൃതത്തിൽ ചിന്തിക്കണം"". എന്തെങ്കിലും മനസ്സിലായോ? അവിടെയാണ് കളി ..."
,എങ്കിൽ മോദിയെ സംസ്കൃതം..... പഠിപ്പിക്കു- ബുദ്ധി വളരട്ടെ-
കള്ളു ചെത്തുകാരനെ വിടാം, അല്ലെങ്കിൽ വേണ്ട ഭാഷ പണ്ഡിതൻ മാണി ആശാനേ ആയല്ലോ"
സായിപ്പ് പറഞ്ഞത് കൊണ്ട് ഇനി നമ്മുടെ അച്ഛന്മാരും അമ്മമാരുമൊക്കെ കുട്ടികളെ സംസ്കൃതം പഠിപ്പിക്കുമായിരിക്കും. ഡോക്ടർ, എഞ്ചിനീയർ, A-ഗ്രേഡ് എന്നിവ മാത്രമാണല്ലോ
ഗ്രീക്ക് ,ചൈനീസ് ഇവയും പഠിച്ചു നോക്കു .ബുദ്ധി വികസിക്കും"
ചൈനക്കാരുവരെ സംകൃതം പഠിച്ചു തുടങ്ങി കമ്മി ശശി !
പഠിച്ചു നോക്കു .കമ്മി തലച്ചോറും വികസിക്കും
തനിക്കു ഉണ്ടെങ്കിലല്ലേ വികസിക്കു . തന്റെ കാര്യായം അധോഗതി

Figure 1: Sample Reviews in the Dataset

context. Parts Of Speech (POS) labeling for Malayalam writings can be prepared either using Bureau of Indian Standards tagset (BIS) or International Institute of Information Technology Hyderabad tagset. The utmost benefit is its pace, when managing with big corpora. In this article, user reviews are tagged using BIS tagset which is shown in figure 3. Sample of the tagged review file is shown in the figure 4

iii) Stopwords Removal

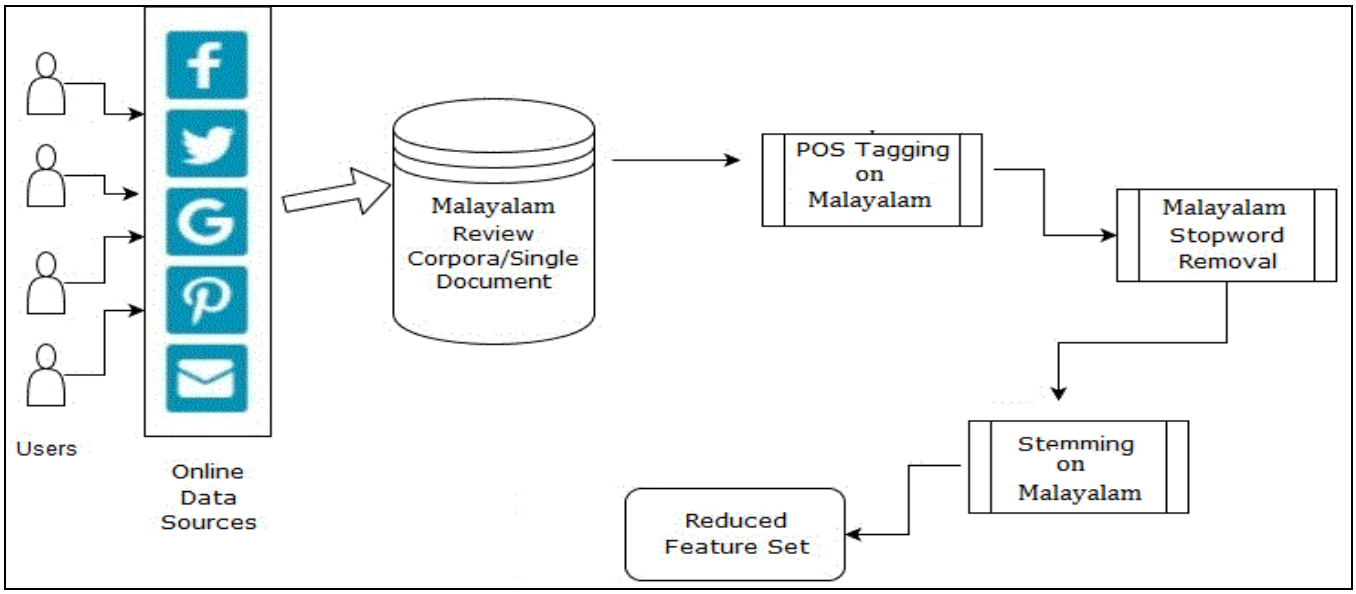

Figure 2: Framework of the Pre-processing phases

SL No	Category			Label	Annotation
	Top Level	Subtype (level 1)	Subtype (level 1)		
1	Noun			N	N
1.1		Common		NN	N_NN
1.2		Proper		NNP	N_NNP
1.3		Nloc		NST	N_NST
2	Pronoun			PR	PR
2.1		Personal		PRP	PR_PRP
2.2		Reflexive		PRF	PR_PRF
2.3		Relative		PRL	PR_PRL
2.4		Reciprocal		PRC	PR_PRC
2.5		Wh-word		PRQ	PR_PRQ
3	Demonstrative			DM	DM
3.1		Deictic		DMD	DM_DMD
3.2		Relative		DMR	DM_DMR
3.3		Wh-word		DMQ	DM_DMQ
4	Verb			V	V
4.1		Main		VM	V_VM
4.1.1			Finite	VF	V_VM_VF
4.1.2			Non-Finite	VNF	V_VM_VNF
4.1.3			Infinite	VINF	V_VM_VINF
4.2		Verbal		VN	V_VN
4.3		Auxiliary		VAUX	V_VAUX
5	Adjective			JJ	
6	Adverb			RB	
7	Postposition			PSP	
8	Conjunction			CC	CC

Figure 3: BIS Tagset

stopwords list which is already manually created and that stopword is compared with the tokens of the review document using sequential search technique. If it matches, the token in the array is removed, and the comparison is continued till the length of array and removes that particular stopword completely from the array. The process iterates until all the stopwords are compared. The Review tokens devoid of stopwords are thus left out in the array and are stored back into the corpora.

Stopwords are frequently occurring words in a natural language which are considered as unimportant in certain Natural Language Processing applications like Clustering, Text Summarization, Information Retrieval, etc. The removal of stop words could be an important step as its elimination reduces the feature space, thus helps in reducing time and space complexity. A dictionary based approach is been utilized to remove stopwords from the review document. A generic stop-word list containing approximately 75 Malayalam stopwords is created in-house. The list is shown in figure 5. A single stop word is read from the Malayalam

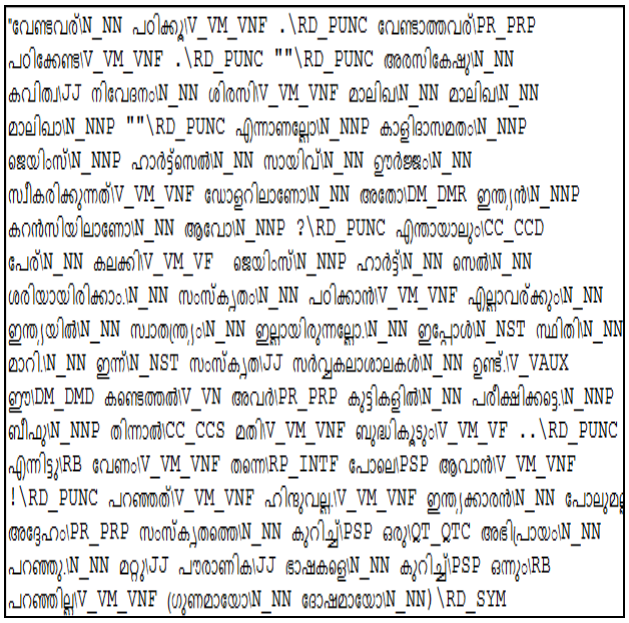
"വേണ്ടവർ\N_NN പഠിക്കൂ\V_VM_VNF .\RD_PUNC വേണ്ടാത്തവർ\PR_PRP
പഠിക്കേണ്ട\V_VM_VNF .\RD_PUNC ""\RD_PUNC അരസികേഷു\N_NN
കവിത\JJ നിവേദനം\N_NN ശിരസി\V_VM_VNF മാലിഖ\N_NN മാലിഖ\N_NN
മാലിഖാ\N_NNP ""\RD_PUNC എന്നാണല്ലോ\N_NNP കാളിദാസമതം\N_NNP
ജെയിംസ്\N_NNP ഹാർട്ട്സെൽ\N_NN സായിപ്പ്\N_NN ഊർജ്ജം\N_NN
സ്വീകരിക്കുന്നത്\V_VM_VNF ഡോളറിലാണോ\N_NN അതോ\DM_DMR ഇന്ത്യൻ\N_NNP
കറൻസിയിലാണോ\N_NN ആവോ\N_NNP ?\RD_PUNC എന്തായാലും\CC_CCD
പേര്\N_NN കലക്കി\V_VM_VF ജെയിംസ്\N_NNP ഹാർട്ട്\N_NN സെൽ\N_NN
ശരിയായിരിക്കാം.\N_NN സംസ്കൃതം\N_NN പഠിക്കാൻ\V_VM_VNF എല്ലാവർക്കും\N_NN
ഇന്ത്യയിൽ\N_NN സ്വാതന്ത്ര്യം\N_NN ഇല്ലായിരുന്നല്ലോ.\N_NN ഇപ്പോൾ\N_NST സ്ഥിതി\N_NN
മാറി.\N_NN ഇന്ന്\N_NST സംസ്കൃത\JJ സർവ്വകലാശാലകൾ\N_NN ഉണ്ട്.\V_VAUX
ഈ\DM_DMD കണ്ടെത്തൽ\V_VN അവർ\PR_PRP കുട്ടികളിൽ\N_NN പരീക്ഷിക്കട്ടെ.\N_NNP
ബീഫു\N_NNP തിന്നാൽ\CC_CCS മതി\V_VM_VNF ബുദ്ധികൂടും\V_VM_VF ..\RD_PUNC
എന്നിട്ടു\RB വേണം\V_VM_VNF തന്നെ\RP_INTF പോലെ\PSP ആവാൻ\V_VM_VNF
!\RD_PUNC പറഞ്ഞത്\V_VM_VNF ഹിന്ദുവല്ല.\V_VM_VNF ഇന്ത്യക്കാരൻ\N_NN പോലുമല്ല
അദ്ദേഹം\PR_PRP സംസ്കൃതത്തെ\N_NN കുറിച്ച്\PSP ഒരു\QT_QTC അഭിപ്രായം\N_NN
പറഞ്ഞു\N_NN മറ്റു\JJ പൗരാണിക\JJ ഭാഷകളെ\N_NN കുറിച്ച്\PSP ഒന്നും\RB
പറഞ്ഞില്ല\V_VM_VNF (ഗുണമായോ\N_NN ദോഷമായോ\N_NN) \RD_SYM

Figure 4: Sample of the reviews after POS tagging

iv) Stemming

This process expels the affixes from inflections and to restore the root form [7]. Applying stemming process during preprocessing is used to reduce words to their basic form and to get good TF-IDF (Term frequency -Inverse Document Frequency) score[9]. Even in a single sentence, same word exists in different forms. Many emphases are feasible for each Malayalam word as it is agglutinative. Stemming process groups all terms that are derived from a similar stem (ex: നല്ലത്, നല്ലതാണ്, നല്ലതാ from the stem നല്ല) [14] and this grouping will increase the occurrence of this stem because frequencies are calculated using the stemmed words not by the actual words. The stemmer uses the algorithm that, upon giving a word as the input, gives the base word as the output. To facilitate and automate the process of matching morphological term variants, the Indicstemmer is used to extract stems of the words in the given review document. This application uses a rule based approach. The stemmer utilizes a set of suffix stripping rules. It follows iterative suffix stripping to handle multiple levels of inflection. The framework is written from the scratch using python language.

The rules for handling the Malayalam language inflections like പ്രതിഗ്രാഹിക, സംയോജിക, ഉദ്ദേശിക, പ്രയോജിക, സംബന്ധിക, ആധാരിക, സംബോധന, plurals, verbs, numbers

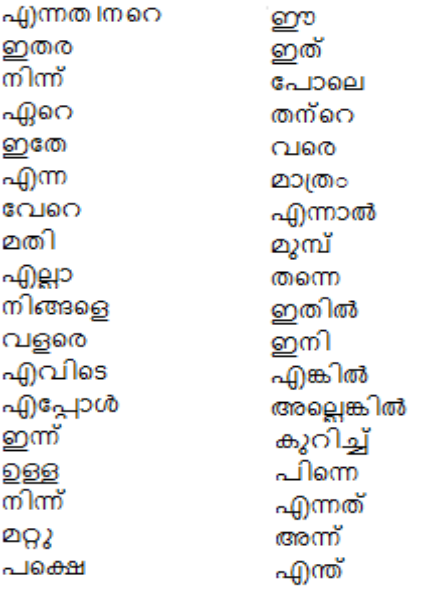
എന്നത നറെ
ഇതര
നിന്ന്
ഏറെ
ഇതേ
എന്ന
വേറെ
മതി
എല്ലാ
നിങ്ങളെ
വളരെ
എവിടെ
എപ്പോൾ
ഇന്ന്
ഉള്ള
നിന്ന്
മറ്റു
പക്ഷെ
ഈ
ഇത്
പോലെ
തന്റെ
വരെ
മാത്രം
എന്നാൽ
മുമ്പ്
തന്നെ
ഇതിൽ
ഇനി
എങ്കിൽ
അല്ലെങ്കിൽ
കുറിച്ച്
പിന്നെ
എന്നത്
അന്ന്
എന്ത്

Figure 5: Sample of Malayalam Stopwords

etc are defined in the rule set. The sample output of review document after stemming is shown in the figure 6.

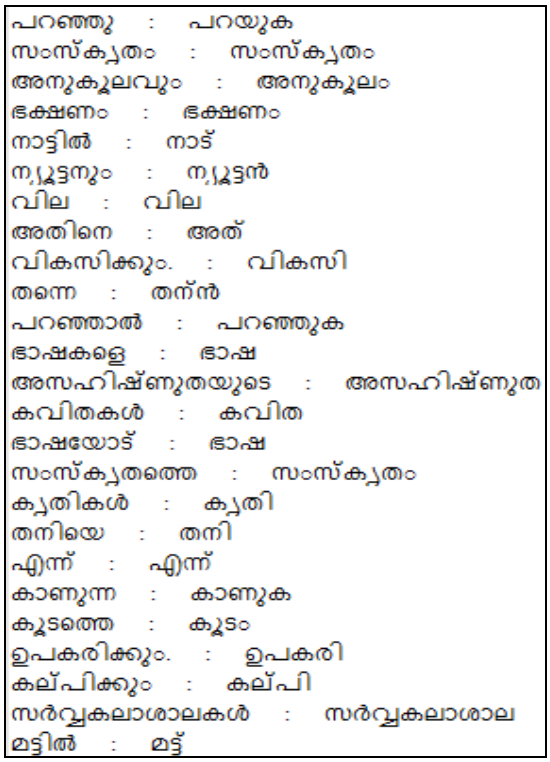
പറഞ്ഞു : പറയുക
സംസ്കൃതം : സംസ്കൃതം
അനുകൂലവും : അനുകൂലം
ഭക്ഷണം : ഭക്ഷണം
നാട്ടിൽ : നാട്
ന്യൂട്ടനും : ന്യൂട്ടൻ
വില : വില
അതിനെ : അത്
വികസിക്കും. : വികസി
തന്നെ : തന്ൻ
പറഞ്ഞാൽ : പറഞ്ഞുക
ഭാഷകളെ : ഭാഷ
അസഹിഷ്ണുതയുടെ : അസഹിഷ്ണുത
കവിതകൾ : കവിത
ഭാഷയോട് : ഭാഷ
സംസ്കൃതത്തെ : സംസ്കൃതം
കൃതികൾ : കൃതി
തനിയെ : തനി
എന്ന് : എന്ന്
കാണുന്ന : കാണുക
കൂടത്തെ : കൂടം
ഉപകരിക്കും. : ഉപകരി
കല്പിക്കും : കല്പി
സർവ്വകലാശാലകൾ : സർവ്വകലാശാല
മട്ടിൽ : മട്ട്

Figure 6: Sample output after Stemming

VI. Results and Discussion

The table 1 represents the number of reduction in the word count after the pre-processing process done on the dataset. The pre-processing process consists of tokenization, POS tagging, Word Inflection and Stemming and Stopwords removal process. The figure 7 shows that 27% of word count reduction occurred for the document 3 and for the documents 4 and 5, 25% and 22% respectively. It is obvious that the pre-processing steps done on the Malayalam review documents considerably reduced the no: of words to be considered and thus the time required for the sentiment analysis in the large corpus.

Table 1: Number of words after each pre-processing phases

Malayalam Review Document #	Total No: of Words in the Document	No: of words after Stemming	No: of words after Stopword Removal
1	424	384	361
2	552	465	443
3	657	509	479
4	512	411	383
5	479	408	375

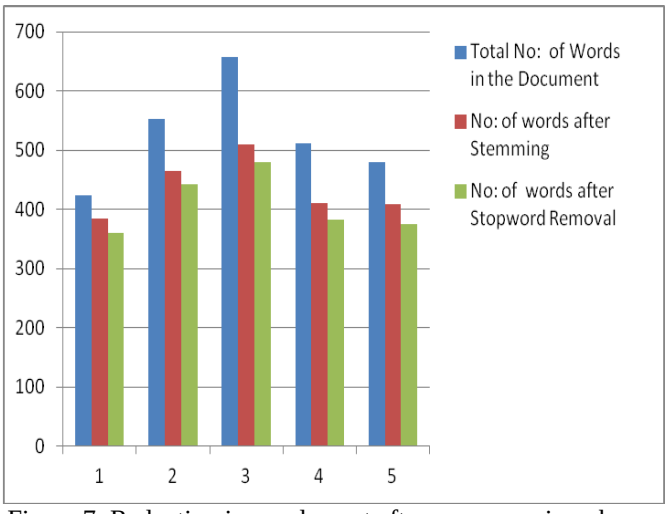

Figure 7: Reduction in word count after preprocessing phases

VII. Conclusion

This work demonstrates various pre-processing phases done on the words given by the user through their reviews in Malayalam language. It is obvious that the pre-processing steps done on the Malayalam review documents considerably reduced the no: of words to be considered and thus the time required for the sentiment analysis in the large corpus. A number of different technical approaches can be taken to calculate the accuracy of each phase which can be consider as a future work

References

[1] Shastri, G., "Kannada morphological analyser and generator using trie",. *IJCSNS*, *11*(1), 112, 2011

[2] Ramanathan, A., & Rao, D. D., "A lightweight stemmer for Hindi", In *the Proceedings of EACL*, 2003

[3] Gagandeep Kaur, Kamaldeep Kaur, "Sentiment Detection from Punjabi Text using Support Vector Machine", International Journal of Scientific Research in Computer Science and Engineering, 5(6), 39-46., 2017

[4] Islam, M., Uddin, M., & Khan, M., "A light weight stemmer for Bengali and its Use in spelling Checker", 2007.

[5] Akram, Q. U. A., Naseer, A., & Hussain, S. "Assas-Band, an affix-exception-list based Urdu stemmer", In *Proceedings of the 7th workshop on Asian language resources* (pp. 40-46). Association for Computational Linguistics, 2009

[6] Dutta, P. K., "An Online Semi Automated Part of Speech Tagging Technique Applied To Assamese" (Doctoral dissertation), 2013.

[7] Kasthuri, M., & Kumar, S. B. R., "An improved rule based iterative affix stripping stemmer for Tamil language using K-mean clustering", *International Journal of Computer Applications*, *94*(13), 2014

[8] Prajitha, U., Sreejith, C., & Raj, P. R., "LALITHA: A light weight Malayalam stemmer using suffix stripping method", In *Control Communication and Computing (ICCC), 2013 International Conference on* (pp. 244-248). IEEE, 2013.

[9] Pragisha, K., & Reghuraj, P. C., "STHREE: Stemmer for Malayalam using three pass algorithm", In *Control Communication and Computing (ICCC), 2013 International Conference on* (pp. 149-152). IEEE, 2013.

[10] Jayan, J. P., Rajeev, R. R., & Sherly, E.. "A hybrid statistical approach for named entity recognition for malayalam language". In *Proceedings of the 11th Workshop on Asian Language Resources* (pp. 58-63), 2013

[11] Nair, D. S., Jayan, J. P., & Sherly, E., "SentiMa-sentiment extraction for Malayalam", In *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on* (pp. 1719-1723). IEEE, 2014.

[12] K, Manju & Peter S, David & Mary idicula, Sumam, "An Extractive Multi-document Summarization System for Malayalam News Documents". 10.4108/eai.27-2-2017.152340.

[13] Renjith, S. R., & Sony, P, "An automatic text summarization for Malayalam using sentence extraction". In *Proceedings of 27th IRF International Conference, 14th June*, 2015

[14] Willett, P, "The Porter stemming algorithm: then and now. *Program"*, *40*(3), 219-223, 2006

Authors Profile

Ms. Deepa Mary Mathews, is presently Assistant Professor of the Department of Computer Applications, FISAT. She was graduated in Chemistry from Mahatma Gandhi University, had her post graduation in Computer Applications from Madurai Kamaraj University, and gained second post graduation in M.Tech in Computer Science and Engineering from Dr.M.G.R.University in the year 2006. She is currently pursuing Ph.D and her research area includes Data Mining, Social Data Analytics and Machine Learning. She has published 9 research papers in the International Journals, National and International Conferences incuding IEEE.

Mr. Sajimon Abraham, MCA, MSc(Mathematics), MBA, PhD (Computer Science). He has been working as Faculty Member in Computer Applications & IT, School of Management and Business Studies, Mahatma Gandhi University, Kottayam, Kerala, India. He currently holds the additional charge of Director(Hon), University Center for International Co-operation. He was previously working as Systems Analyst in Institute of Human Resource Development, Faculty member of Computer Applications in Marian College, Kuttikkanam and Database Architect in Royal University of Bhutan under Colombo Plan on deputation through Ministry of External Affairs, Govt of India. His research area includes Data Science, Spatio-Temporal Databases, Mobility Mining, Sentiment Analysis, Big Data Analytics and E-learning and has published 52 articles in National, International Journals and Conference Proceedings

