A Thesis Report

On

Generating Opinion Lexicon for Nepali using Microblogs

For Partial Fulfillment of the Requirements for the Degree of
Master of Computer Information System Awarded by
Pokhara University

Submitted by

Sawan Vaidya
MSCS
15526

Under the Guidance of

Bal Krishna Bal
Associate Professor
Kathmandu University
Dhulikhel

Department of Graduate Studies
Nepal College of Information Technology
Balkumari, Lalitpur

April 2018

ABSTRACT

A basic ingredient of Sentiment Analysis that uses Bag-of-words approach is an Opinion Lexicon containing words/phrases along with their positivity and negativity ratings. Many methods have been suggested for building such a lexicon. In this thesis a method utilizing micro-blogs (user comments from Facebook pages) is suggested. While, this lexicon can be domain specific, it will be a first in the Nepalese context and can lay the foundation for development of Sentiment Analysis resources. Major challenges are that data generated by users on the Internet is subject to informal language. Also misspellings, abbreviations, emoticons and multi-language content pose additional hurdles. The benefits are that micro-blogs are richer in subjective content compared to text found in formally written sources. The thesis will investigate the use existing lexicons (such as emoticon lexicon) and manually generated lexicons as seed lexicon for enriching and expanding them. In this thesis statistical approach is taken into building a microblog specific lexicon. The output lexicon has been evaluated using Sentiment Analysis on two sets of test data - a manually labelled microblog subset and a set containing word meanings of SentiWordNet words.

Keywords: Opinion Lexicon, Sentiment Analysis, Micro-blog, Twitter, Classification, Nepali,Nepalese, Multiple Languages

ACKNOWLEDGMENT

First and foremost, I would like to thank my supervisor Asst. Prof. Bal Krishna Bal, of Kathmandu University for his support, ideas and guidance.

I would also like to thank Assoc. Prof. Dr. Balaram Prasain for reviewing the report and providing valuable suggestions for further work.

I would like to express my sincere gratitude to Assoc. Prof. Saroj Shakya, Graduate Program Coordinator, Department of Graduates Studies for providing me the opportunity to carry out this thesis midterm work.

I would also like to thank my family, friends and colleagues for their support and encouragement. Last but not the least, I wish to record my appreciation to all the people who directly or indirectly contributed their help during the course of this thesis.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Background

Sentiment Analysis (SA) is a classification process where the opinion presented by a writer or a speaker is categorized into opinion bearing (subjective) or factual (objective) classes. Subjective content can be classified further into positive, negative and neutral classes. Other classification examples can be detection of disgust, love, humor etc. SA can also be used to determine aspects of a text. These aspects could be political, historical, description etc. One of the predominant methods for performing SA is using an Opinion Lexicon (OL), where the analysis performed checks the proportion of opinion bearing words, phrases or patterns in text.

OL can be used along with other tools such as WordNets, translational systems, language semantics and machine learning algorithms to perform several types of analysis of text based content. A direct application is to study sentiment of users across social media. Some applications of sentiment analysis are as follows:

- Study of sentiments related to a certain political entity or ideology
- Study of awareness and effectiveness of an implemented government policy
- Study of effectiveness of advertising campaigns
- Analysis of reaction to certain events such as natural disasters, strikes etc.
- Analysis of reviews of customers in e-commerce portals
- Behavioral analysis of users of social media

- Analysis of reaction to music, movies and other types of entertainment channels

There has been prior research on SA in Nepal (See Literature Review), but work in this field is rather limited. Hence, resources for someone trying to start such an analysis in Nepal is scarce. Some resources that are considered the bread and butter of SA and that are available in English but not in Nepali are WordNets, digital dictionary, synsets and POS taggers. Most methods used for developing OL utilize the aforementioned resources. This thesis will attempt to solve the problem by developing an OL development process that does not require any of the above.

There are challenges to using micro-blogs as opinion corpus, especially in the Nepali context. Nepalese users generate text content on the Internet in Devanagari, English and Romanized Nepali. Second, there is much difference in written and spoken Nepali which is also reflected in language used in formal and informal channels. Micro-blogs are more likely to contain text written in an informal manner. However, from a large collection of words found in micro-blogs, a good portion will be useful.

This research will not use machine learning techniques but attempt to build an Opinion Lexicon using statistical methods. Measurements will be made on how words and phrases occur with one another in positive and negative texts obtained from microblogs. Symbolic and semantic analysis of text and grammar will not be made, and instead any text will be assumed to be bag of words. Statistical techniques will be used to filter out useful and non-useful words. The obtained lexicon will be domain dependent. As such not all terms marked as sentiment bearing in the obtained lexicon will be useful for other domains. Since, the corpus (microblogs) used represent language used in an informal manner in text form, the terms found in the lexicon should be useful for other informal domains.

1.2 Problem Statement

One of the most basic methods of determining sentiment expressed in a text is to use an opinion lexicon which contains frequently used words with their positivity, negativity and neutrality ratings. Opinion Lexicons could be generated by translating an existing lexicon from English to Nepali, but a translated lexicon

would not accurately represent the words, phrases, expressions and slangs used by online users and hence would not be effective. In absence of language resources such as WordNets, digital dictionary, synsets and POS taggers, many other methods of lexicon building are also unavailable. The proposed method attempts to dynamically generate an opinion lexicon by classification of text found in micro-blogs. A domain specific method of performing sentiment analysis using multiple given lexicons in English, Nepali languages and Emoticons can be suggested. This method is not language specific and can be used in absence of other resources such as WordNets and labeled corpus, which is useful for many non-major languages that have not made much advancement in terms of sentiment analysis.

1.3 Research Objectives

The primary objective is to develop a method to generate an opinion lexicon consisting of words and phrases written by Nepalese users. The words and phrases will also contain statistic on how frequently the word was used in a positive or negative context. The opinion lexicon generated can be used for sentiment analysis in the domain of micro-blogs written by Nepalese users.

- Develop a method for generating opinion lexicon using text available in microblogs
- Refine a strategy for filtering out mistakes and useless terms from the microblog lexicon
- Generate an opinion lexicon for Nepali, specify its pros and cons.

1.4 Research Questions

- How effective is the proposed method compared to existing solutions?
- What would be a good sentiment lexicon for representing general Nepali content?
- What are the challenges that need to be overcome to develop a lexicon from microblogs?

- How to provide a statistic (score) that represents the degree of subjectivity in words and how can it be calculated using this method?

Chapter 2

LITERATURE REVIEW

2.1 Related Work

There are two mainly used methods to generate Opinion Lexicons [22]. First method is dictionary based which uses relations between words found in WordNets and/or dictionaries. Second method uses a sentiment corpus to produce domain specific lexicons.

2.1.1 Dictionary Based Examples

WordNet is a lexical resource containing words, their parts of speech and their meanings in gloss form. It also defines semantic relations between words by means of synonymy, antonymy, hyponymy, meronymy, troponymy and entailment . Polysemous (having multiple meanings) words can be distinguished by means of the gloss provided. WordNet provides a dictionary readable to machines and has become the key resource for building Sentiment Lexicons [25].

One of the most popular resource used for sentiment analysis in English is SentiWordNet, currently in version 3.0 [2] [11]. This resource can be considered as a hydrid of WordNet and an Opinion Lexicon. The construction of SentiWordNet was done by traversing the relationship between words in WordNet such as synsets and antonyms. This semi-supervision was used to expand upon a small known opinion lexicon. Additionally a random walk method traverses and visit the words through their links and the more visitations mean stronger linkages between words, and this has been used to determine the numerical polarity of the words. SentiWordNet gives every word a positive and a negative rating. Neutral rating can be calculated by subtracting the sum of the prior from 1.

Sentiment Analysis is not just about Subjective and Objective or Positive and Negative. For example, in [20], an opinion is described by [Topic, Holder, Claim, Sentiment] quadruplet. They use a technique similar to SentiWordNet's for measuring the strength of words. They start with an initial seed lexicon. Then, a new word is given a numerical strength based on the proportion of synonyms that are present in an existing seed lexicon. In [16], the authors mine customer reviews, identify opinion sentences and summarize the results. In the process, they have used a simple technique for determining orientation of sentences, based on adjectives. They start with a seed opinion set and look for synonyms or antonyms of the adjectives and when found, the orientation of the words founds are either kept the same or reversed respectively.

2.1.2 Corpus Based Examples

While dictionary based approach is great at determining polarity at word level, it follows a bag of words model, where opinions of words are not set in relation to one another. Corpus based methods determine polarity of words in relation to one another. However, it's drawback is that the resulting lexicon can be domain specific [22].

One of the most used metrics in corpus based approaches is pointwise mutual information (PMI). [34] suggested an unsupervised method of rating reviews as positive or negative based on the difference between the PMI a review receives with respect to the word "excellent" and the PMI received with respect to the word "poor". The assumption is that words with similar polarities occur together.

Another popular method is based on adjectives connected by conjunctions such as 'and' and 'but'. In [15] adjectives joined by 'and' tend to have similar polarities whereas, adjectives connected by 'but' have opposite polarities. This can be used to expand a lexicon with adjectives of known polarity.

Another PMI based method used in [14] is to build a vector space model of words based on PMI. A singular vector decomposition method is used to break down the vector space matrix into U and V matrices. The U matrix is then used as a the column vector to represent every words. A graph of words is then created where the edges are cosine similarities between words. A random walk method is then used to then rate the words as positive and negative based on visits.

2.2 Non English Languages

2.2.1 Nepali

[13] translated SentiWordNet into Nepali. However, they did not convert the polarities in SentiWordNet to Nepali. They also developed a subjectivity clue list allowing identification of subjective texts. Other than that work on opinion lexicon development in Nepali was not found. Other resources developed for sentiment analysis include Nepali Sentiment Corpus with subjective and objective labels [13].

2.2.2 Other Languages

[35] uses an n-gram pattern finding technique to build a domain specific lexicon for Indonesian language. They look for top n-grams patterns and top POS disambiguated n-gram patterns. Patterns are classified based on availability of sentiment seed words, which were words translated from English. New candidate words found in the patterns are rated based on their occurrence in positive context or negative context. A PMI method is used for scoring.

[29] attempts to align English WordNet with Spanish counterpart. A translated lexicon is hence generated in Spanish. [19] uses a similar approach. It translates four English sentiment lexicons - ANEW, AFINN, SenticNet and NRC Word Emotion Association Lexicon. [6] uses a Malaysian WordNet called Bahasa to develop its sentiment lexicon.

2.3 Study Area

As already mentioned in the Introduction section, resources for Nepali language are rare. [13] attempted a WordNet translation based approach. Disadvantage of translation approach is that only a small amount of words in the target lexicon preserve their original sentiment polarities, post-translation [6]. A corpus based approach has to be used.

The approach taken in this thesis will attempt to use a method similar to [35]. However, this research will not look for frequent patterns but look for microblogs, which are short user generated texts are assumed to be packed with opinions. Like most other corpus based approaches, this research will also assume that words with similar polarities tend to occur together. Based on this, a method to generate

micro-blog domain specific lexicon will be proposed. Additionally, ways to refine the generated lexicon will be investigated and the advantages/disadvantages of using such a lexicon will be studied.

It would have been easier if a corpus such as IMDB's movie ratings was available for Nepali. Such a pre-classified corpus would allow building a supervised framework for lexicon development. A semi-supervised method will be investigated instead. The framework will use a seed lexicon with Devanagari words and/or a emoji lexicon. [27] mentions that emoji rankings were not seen to differ significantly in 13 European languages. This research will assume that maybe the emoji lexicon they built can be used in Nepali as well.

Chapter 3

RESEARCH METHODOLOGY

3.1 Related theory

3.1.1 Opinion Lexicon

There can be different ways to represent an opinion lexicon. A simple list of positive and negative words can be called a lexicon. For example a list of stopwords is a simple lexicon containing objective/neutral words. Some lexicons provide multiple classes for positivity such as 1,2,3,4 and 5 where 5 could be the most positive class and 1 could be the most negative class. Some other lexicons provide positive, negative and neutral strengths. This section will describe the features of the opinion lexicon that will be built during this research.

In this research two manually built and one generated opinion lexicon will be experimented with. The manually built ones are the 2-word (A.1) and the 32-word opinion lexicon (A.2). The emoticon lexicon (fig:emoji-sentiment-rankings-full) is obtained from a research on sentimetn of emoticons [27]. These lexicons will be used as seed lexicons and used in an attempt to obtain new sentiment bearing words from microblog corpus.

It was found in many researches that adjectives are important indicators of opinions [22]. The probability of a sentence being subjective based on the presence of at least one adjective was found to be 55.8% [3]. In the manually built lexicons most words used are adjectives and their positive and negative strengths are assigned to indicate their leaning towards negative or positive sentiments. The absolute value of their strengths are arbitrary and their purpose is to induce positive and negative strengths onto other words found in the microblog corpus. The strengths can be assumed to be the supposed probability.

The emoticon lexicon has positive and negative strengths that can be interpreted as probability of the emoticons to occur in texts bearing a certain positive, negative or neutral sentiment. While the 2-word and 32-word manually built lexicons have arbitrary strengths, the emoticon lexicon strengths were calculated from a large number of observations [27].

3.1.1.1 The simplest opinion lexicon

Table 3.1: Simple 2-word lexicon

Word (w)	Positive (p)	Negative (n)
राम्रो	0.9	0.1
नराम्रो	0.1	0.9

A two class, two word lexicon shown in Table 3.1 will be used in this section for explanation of concepts. The lexicon provides positive and negative strengths of the words राम्रो and नराम्रो which are opposites of each other. The lexicon that is attempted to be built in this research will just like Table 3.1

3.1.1.2 Score/Positivity

Table 3.2: Simple 2-word lexicon with Score instead of Strength

Word (w)	Score
राम्रो	0.8
नराम्रो	-0.8

Table 3.2 shows an alternate representation of the table 3.1. Here positive and negative strengths of words has been replaced by a single score. This score can be calculated by using the formula in Eqn. 3.1a. If we assume that the positive and negative strength of a word sum up to one (Eqn. 3.1b), we can also get strength if we know the score (Eqn. 3.1c, 3.1d). If positive class was labelled as 1 and negative class as -1, score can be interpreted as the expected class value.

$$strength(positive) - strength(negative) = score \tag{3.1a}$$

$$strength(positive) + strength(negative) = 1 \tag{3.1b}$$

$$strength(positive) = \frac{1 + score}{2} \tag{3.1c}$$

$$strength(negative) = \frac{1 - score}{2} \tag{3.1d}$$

Score ranges between -1 and 1, where 1 is the most positive and -1 is the most negative. The advantages of using score instead of positive and negative strengths is that score provides a continuous range of positivity where higher number is positive. Neutral words can be supposed to be those words that have scores close to zero.

3.1.1.3 Subjectivity

Table 3.3: Simple 2-word lexicon and Subjectivity

Word (w)	Subjectivity
राम्रो	0.6
नराम्रो	0.6

Another way to interpret positivity score shown in the previous section is to derive a value that represents subjectivity from it. One way is shown in Eqn. 3.2. Subjectivity of a word is it's tendency to show opinion. So the formula puts highly subjective words near 1 and highly objective words near -1. Table 3.3 shows the simple opinion lexicon and it's calculated subjectivity. Notice that both words have the same subjectivity because of the absolute value in the formula.

$$subjectivity = 2 * |score| - 1 \tag{3.2a}$$

While some opinion lexicons have a separate class for neutral, for this thesis it will be assumed that neutral/objective words are words with low subjectivity i.e. words which are neither too positive nor too negative.

3.1.1.4 Usage

In order to understand the strategy presented in this thesis, an explanation of how an Opinion Lexicon is used for sentiment analysis is necessary. The following steps show one way in which this is performed. Other methods may have minor variations. This is the method used throughout this thesis.

- All words in a target text that are contained in the Opinion Lexicon are extracted
- The score (described in the previous section) of each word found is computed
- The score is averaged
- Based on whether the average score is greater than or less than 0, the text is classified as positive or negative respectively

Eqn 3.3 shows the formula used for calculating average score of words in a piece of text.

$$averagescore = \frac{\sum_{i=1}^{N_p} score(w_i)}{count(w_i)} \tag{3.3a}$$

(3.3b)

Averaging the score allows comparing documents and prevents bias towards longer documents. Table 3.4 shows sentiment analysis of some sentences using these equations. Table 3.1 is used as the lexicon for this analysis. Words not found in the lexicon do not have any scores hence contributing to a score of zero (0). Words that are not found in the lexicon are not included in the average. In Table 3.4, those sentences with equal positive and negative scores are marked as unknown, but this is a matter of choice.

Note that Tables 3.2 and 3.4 will be referred to repeatedly in this section.

Table 3.4: Sentiment Analysis of various sentences using Eqn 3.3, and Lexicon in Table 3.2

Id	Sentences	score	verdict
1	म राम्रो र सफा	0.8	Positive
2	म नराम्रो र फोहर	-0.8	Negative
3	म सफा र शुद्ध	0	Unknown
4	म फोहर र गनाउने	0	Unknown
5	म राम्रो र नराम्रो	0	Unknown

3.1.2 The Problem

Building a 2-word lexicon such as the one in Table 3.2 can be done easily. But to build a larger lexicon requires many hours of manual work. The scores assigned is also a matter of perspective and depends on human judgment. Also in time, meanings of words change and so do their popularity. To address these needs continually at a human level is neither practical nor feasible. Hence, an automated solution is required.

In Table 3.4, words such as सफा, फोहर, शुद्ध and गनाउने are not included in the lexicon, and hence their scores are unknown. The proposed method in the following sections will try to address this problem and devise a method to approximate the scores of the unknown words.

3.1.3 The Core Insight

The intuition behind the Opinion Lexicon development method is the observation that if documents can be segregated to classes based on the words they contain, then it should be possible to classify words in document based on the class it receives. A positive text is more likely to contain positive words and a negative text is more likely to contain negative words. Words that occur in both classes are more likely to be neutral. In Table 3.4 a clever algorithm would be able to infer that सफा is positive and फोहर is negative because they occur within positive and negative sentences respectively. Moving ahead in this direction, word शुद्ध can be inferred as positive and गनाउने can be inferred as negative. Words म and र occur in both positive and negative classes and can hence be inferred as neutral

In reality, an opinion word does not necessitate that a text comprising it is of the same opinion. For example adjectives can easily be negated to change the

opinion polarity of a word. Hence, a large number of opinion bearing text needs to be used so that errors that result from such ambiguities will be averaged out. And as it is not practical to manually develop such large classified corpus, text in micro-blogs are used. Micro-blogs such as user comments on Facebook can be expected to contain more opinionated text compared to text in media such as news, articles, blogs and books.

3.1.4 The Proposed Solution

Table 3.5: Classification of Words from Table 3.4

Word	Positive	Negative
राम्रो	1	0
सफा	1	0
म	1	1
र	1	1
नराम्रो	0	1
फोहर	0	1

A solution to the problem pointed out in Section 3.1.2 will be elaborated in this section. First a table containing count of words in positive and negative classes is computed (See Table 3.5).

Now we know which words occurred in which classes and how many times. We could additionally conduct some calculations which would give a numerical strength to each word showing it's belongingness to each class. First the probability of each word given a class is computed (See Eqn. 3.4a). The strength is then set to be the proportion of the probability across all classes (See Eqn. 3.4b).

$$P(w_i|class_j) = \frac{count(w_i)}{count(class_j)} \tag{3.4a}$$

$$strength(w_i, class_j) = \frac{P(w_i|class_j)}{\sum_j^{Pos,Neg} P(w_i|class_j)} \tag{3.4b}$$

$$score(w_i, class_i, class_j) = strength(w_i, class_i) - strength(w_i, class_j) \tag{3.4c}$$

A Score for positive vs negative can be calculated by simply subtracting the strength (See Eqn. 3.4c). In Eqn. 3.4c, when $class_i$ is positive and $class_j$ is negative, score represents the tendency of a word to be positive than negative.

At the same time, words with scores close to zero (0) can be expected to be neutral words.

Table 3.6: Lexicon Built from sample sentences in 3.4. New opinion words detected are underlined

Word (w)	Pos	Neg	P(w|Pos)	P(w|Neg)	Strength (pos)	Strength (neg)	Score
राम्रो	1	0	0.25	0	1.00	0.00	1
सफा	1	0	0.25	0	1.00	0.00	1
म	1	1	0.25	0.25	0.50	0.50	0
र	1	1	0.25	0.25	0.50	0.50	0
नराम्रो	0	1	0	0.25	0.00	1.00	-1
फोहर	0	1	0	0.25	0.00	1.00	-1
Total	4	4					

Scores for words सफा and फोहर were calculated in Table 3.6. We can repeat the calculation with the newly obtained lexicon to determine ratings for words गनाउने and शुद्ध as well.

3.1.4.1 Limitations

The solution mentioned in Section 3.1.4 has overlooked several details:

- It assumes thata the sentences have no error in spellings
- A bag of words approach is used and the order of words is not considered important.
- The sample sentences only have subjective words and stopwords. It does not have proper nouns
- The sample sentences have been devised to give a nice answer

Real microblogs comprise of text which are far more complicated than those shown in the sample sentences. The proposed method is able to approximate which words occur with or are linked to an initial lexicon but in reality this method may lead to mistakes. For example, a popular celebrity might always be talked about positively. The proposed method will mark the celebrity name as a positive word.

Despite these limitations, the thesis will optimistically assume that with a large set of data and paired with various statistical techniques, the mistakes will be minimized and most terms in the lexicon will be useful.

3.2 Data Collection

Other microblog sources such as Twitter feeds, or comments on various blogs could have been included but only comments from Facebook pages will be used because they present a consistent way of for scrapping, and would allow easier categorization, which could be useful in the future stage of analysis. The following items are required to retrieve comments from facebook:

- Facebook PageID
- appId
- appSecret

Every page on facebook has a PageID. The PageID can we accessed by visiting the facebook page and looking for the term following http://www.facebook.com/PageID. For example in https://www.facebook.com/xyz/, xyz is the PageID.

In order to retrieve appId and appSecret one needs to visit
https://developers.facebook.com/apps
and create an app. The process of obtaining the id and the secret is straightforward from there onwards.

After obtaining all the required items, a GET request to https://graph.facebook.com can be sent to obtain allowed page data. For comments, the URL needs to be formed in the following way:

https://graph.facebook.com/<PageID>/feed/?access_token=<appId>
|<appSecret>&fields=comments

A JSON document is returned. The data is paginated. To obtain data from other pages, the link given in 'paging' section of the JSON document can be followed

3.3 System Model and Work Flow

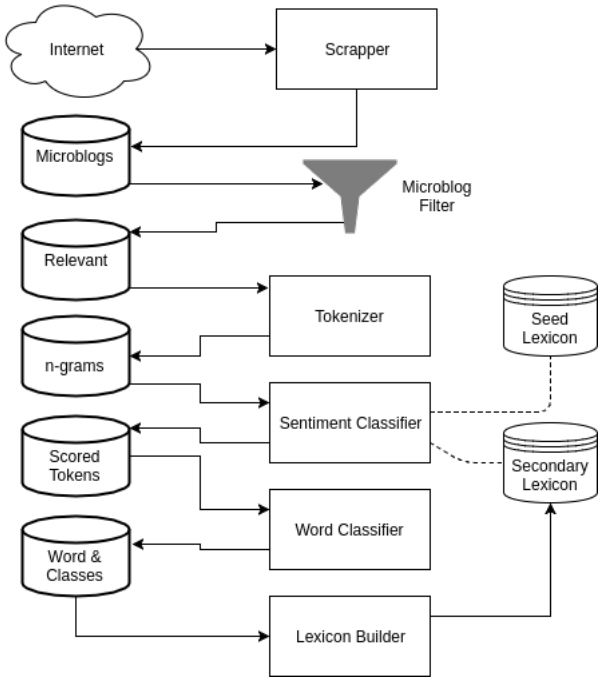

Figure 3.1: Block Diagram of the proposed solution

In Figure 3.1, the rectangular blocks represent processes, the cylindrical blocks represents the database, the funnel represents data filter, the cloud represents the internet, and the cylindrical blocks with striped lines represent the lexicons. The arrows show the direction of the program flow. The portion starting from Sentiment Classifier and below repeats for a specified number of iterations.

3.3.1 Scrapper

A scrapper collects the microblogs from the web and saves them to the database. As mentioned in the Data Collection section, comments from facebook pages will be scrapped. The tool used for scrapping is called Scrapy and will be talked about in Tools section.

3.3.2 Microblog Filter

The text in comments can be composed in different ways. They contain emoticons, devanagari text, latin text and a mixture of the previous. Although this poses a difficulty in cleaning up, some of these features is also of advantage. For example

in the absense of devanagari opinion lexicon, sentences containing english words can be used to classify the comments and infer the opinion in devanagari words. Emoticons contained in the text along with devanagari can be used in similar way.

3.3.3 Tokenizer

The tokenizer's responsibility is to break down comments into individual terms or ngrams. For this, special characters are removed and only devanagari and emoticon characters are kept. In addition to tokenizing the comments, the tokenizer will also filters words based on TF-IDF and document-count such that much of words that undergoes analysis is void of sparely used words, typos or superfluously used words. Eqn. 3.5 were used for computing TF-IDF of the words. At the end of this stage each microblog is broken down into token collections. The tokenizer works in two steps:

3.3.3.1 Word Count and TF-IDF

$$TF = 1 + log_{10}(F) \tag{3.5a}$$

$$IDF = log_{10}(\frac{N}{D}) \tag{3.5b}$$

where, TF = Term Frequency, IDF = Inverse Document Frequency
F = Occurences of a word, N = Total Documents
and D=Documents where the word occurred

Two statistics for each microblog is computed at this stage. The word document count is the number of documents/microblogs a word occured in. At the same time the formula in 3.5 is used to compute a Term Frequency - Inverse Document Frequency of each word. These statistics are used in the next step

3.3.3.2 Ngram sets

In this step, each microblog is first broken down into sentences. Ngrams are then extracted from each sentence. Ngrams must satisfy the greater than a set minimum word document count threshold. This allows typos to be removed. TF-IDF of each Ngram should fall within a prespecified range. This allows

selection of important words and reduce unimportant words such as stop words from the Ngram set.

At the end of this step, a token (ngram) collection is produced for each sentence and for each microblog.

3.3.4 Sentiment Classifier

The sentiment classifier uses the tokens from the previous step and applies the method specified in Eqn. 3.3 to classify the token collections. Those token collections that do not have any linkages with the supplied lexicon are ignored.

Two lexicons are used for this stage. The first lexicon or the primary lexicon is supplied at the beginning of the processing. The second lexicon or the secondary lexicon is generated and changes in each step. If we consider the the primary lexicon to be more reliable than the generated lexicon, it can be weighted higher than the secondary lexicon.

There are multiple options for primary lexicon:

- A 2 word devanagari lexicon with राम्रो and नराम्रो shown in Table 3.1
- A devanagari lexicon was manually developed and contained 16 positive and 16 negative words (Table A.2)
- Emoticon Lexicon was obtained from [27]. It contained emoticons and their corresponding sentiment scores
- English Lexicon can be obtained from sources such as SentiWordNet. (This method was not used)

For this research, English Lexicon will not be used and experiments will be restricted to devanagari and emoticon lexicons. The output of the sentiment classifier step is a collection of tokens for each microblog along with its class (positive or negative)

3.3.5 Word Classifer

The word classifier determines the counts of ngrams from token collections. The class of an ngram is set to be the same as the class of the token collection in which the ngram was found (determined in the previous step). A word may appear in one class in a microblog and in another class in another microblog.

The output of the word classifier is a list of words along with the number of classifications to each class. For sample sentences in Table 3.4, the word classification output is shown in Table 3.5.

In addition to generating counts for ngrams in positive and negative contexts, the word classifier also generates a statistics such as mean, standard deviation, skewness and range. The purpose of these statistics is to filter the lexicon in the next step.

3.3.6 Lexicon Builder

After the ngrams have been classified, and their positive and negative counts are developed in the Word Classifier step, a lexicon can be built. This is the primary purpose of the lexicon builder. The lexicon builder uses the calculations shown in Eqn. 3.4 to provide strengths and/or scores to the words.

In addition to to providing numerical scores to ngrams, several filters can be added at this stage to refine the lexicon. The secondary purpose of the Lexicon Builder is to filter the data based on statistics obtained in the previous step so as to refine the obtained lexicon. The filters that are applied as as follows:

3.3.6.1 Skewness Filter

Pearson's median skewness (Eqn. 3.6) has been used to see whether an ngram's average score is positively or negatively skewed. Each sentence classified produces an average score for the ngrams found in the sentence (See Eqn. 3.3). Hence an ngram has as many average scores as the number of sentences it is found it. The mean, median and standard deviation of the average scores are computed in the Word Classifier step and the skewness of this data is computed.

In general an ngram appearing uniformly across positive and negative classes will not be skewed. The skewness filter attempts to capture only those ngrams which have a threshold skewness.

$$skewness = \frac{3 * (mean - median)}{standarddeviation} \tag{3.6a}$$

3.3.6.2 Range Filter

As mentioned before, multiple average scores are obtained for each ngram. The difference between the highest and lowest average score gives the range (See

Eqn. 3.7)

This filter specifies that the range should be greater than a minimum threshold range. This is put to ensure that ngrams that do not have a variety of average scores are eliminated. Range filter was put so that only those ngrams that appear in a diverse number of situations are selected

$$range = highestscore - lowestscore \tag{3.7a}$$

3.3.6.3 Score Filter

Like skewness, the ngrams with very high or very low scores are probably more likely to be subjective than objective. Scores near +1 represent ngrams used more in positive context and scores near -1 represent those used in negative context. Scores near 0 represent ngrams that are neither positive or negative. So this filter keeps only those ngrams likely to be subjective by filtering out those ngrams between a specified range.

3.3.6.4 Count Filter

The word classifier creates a table of positive and negative counts for all ngrams. Based on the counts the Lexicon builder computes a score. Some ngrams have more counts and some have sparse counts. The scores computed from ngrams with high count can be considered to be more reliable as there is more evidence of this score. The count filter is used to set a minimum count threshold. Ngrams with count less than the minium count are discarded

3.4 Tools

3.4.1 Python Programming Language

Python was chosen as the programming language tool because of the numerous libraries available to it enabling text processing. Some of these libraries - numpy and matplotlib were very appealing. Python is also compatible across multiple operating systems and comes readily integrated into Linux environments which would be a plus for running experiments on cloud servers.

3.4.2 MongoDB

MongoDB was picked as the database for storage and processing because of the possibility of having to store unstructured data. MongoDB uses document based storage where each document is in a JSON-like format and does not require that each document have the same keys. MongoDB also has capabilities of Map Reduce which would be required to run heavy computations like Word-Count.

3.4.3 Scrapy

Scrapy is the most popular tool used for Web Scraping in Python and was hence picked for scrapping data such as comments from facebook.

Chapter 4

EXPERIMENTS AND EVALUATION

Before talking about the results, a set of experiments need to be defined and the parameters they used need to be explained. This is done in Section 4.1. This is followed by the explanation of evaluation methods used for testing the results in Section 4.2. Next chapter (5) will talk about the outcome of these experiments.

4.1 Experimental Setup

This section will explain the rationale behind using the initial parameters that were set before running the experiments. Explanation of some parameters will also require understanding the evaluation procedure talked about in Section 4.2

Table 4.1 describes the parameters used for every experiment. The only difference is the primary (seed) lexicon. Table 4.2 describes the filters used during Build Lexicon step. The experiments and the lexicon used are as follows:

- Experiment 1 : Emoticon Lexicon (Fig. A.1)
- Experiment 2 : Manually build 32 word lexicon (Table A.2)
- Experiment 3 : 2 word lexicon (Table A.1)

Table 4.1: Parameters used for experiments

Parameter	Value	Parameter	Value
Devanagari	Mandatory	TF_IDF_min	8
Latin	Disallow	TF_IDF_max	10
Test_Data	Disallow	Doc_count_min	32
Emoticons	Optional	Iterations	20
TestData	1059	Ngrams	Ngram Method 2
Microblog_Count	108566	Build_Parameters	See Table 4.2

Table 4.2: Build Filter Parameters

Parameter	Value
Range	range>=0.05
Count	count>=Mean/4
Score	|score|>=0.5
Skewness	|skewness|>=Mean/8

4.1.1 Dataset

The microblog dataset used for the analysis consists of comments extracted from various facebook pages. This includes popular pages from political parties, celebrities and news organisations. Table 4.3 shows the comments extracted from each facebook page. Table 4.4 shows the dataset that was used for the experiments in underlines. When test data is subtracted from this 108566 microblogs are obtained for use. When broken down, 290171 sentences are obtained. The Tokenizer, as shown in Fig: 3.1 breaks down these sentences into token collections and Sentiment Analysis is performed on them thereafter.

Table 4.3: Comments retrieved from various facebook pages

PageID	Comments	PageID	Comments
MaoistCenterNp	2901	bbcsajhasawal	34320
BibeksheelSajha	7590	NepalPolicePHQ	38768
yourMadanKrishnaShrestha	8433	dc.nepal	42433
nepalicongresshq	9864	DrBaburamBhattarai	51384
communistpartynepal	11986	rekhathapa.net	55116
hari.bamsha.acharya	14286	PriyankaKarkiofficial	60660
setopati	15162	BBCnewsNepali	64062
nepal8thwonder	23158	nagariknews	71181
onlinekhabarnews	30267	eKantipur	73286

Table 4.4: Latin, Devanagari and Emoticon Composition of comments

Latin	Devanagari	Emoticons	Count
No	No	No	16694
No	No	Yes	890
No	Yes	No	106506
No	Yes	Yes	3180
Yes	No	No	448296
Yes	No	Yes	11317
Yes	Yes	No	26227
Yes	Yes	Yes	1747
No	Yes	Don't Care	109686

4.1.2 Character Selection

The dataset is composed mainly of devanagari, emoticons and latin characters. Table 4.1 shows that only microblogs that had characters in Devanagari character set in them were selected. No microblog selected had characters from Latin character set expect punctuations and white spaces. Emoticons were allowed but was not a requirement. Microblogs that were labelled as Test Data was disallowed in the selection. As discussed in the previous section, eventually, 108566 microblogs were deemed suitable for processing. 1059 microblogs were used for testing

4.1.3 Seed Lexicon

The Opinion Lexicon (OL) development method used in this thesis requires an initial (seed,primary) lexicon. The process described in the Research Methodology (Section 3.1.4), enriches the lexicon in every iteration. Three types of initial lexicon were used for this research - emoticon lexicon (Sample in Table: 4.1, full in Appendix: A.1), 2-word lexicon (Table: 3.1), and the 32-word lexicon (Appendix: A.2). The positive, negative and neutral strengths from Emoji Sentiment Ranking V1 (http://kt.ijs.si/data/Emoji_sentiment_ranking/) was used to build the Emoticon Lexicon.

Top Negative				Top Positive			
Emoticon	Positive	Negative	Neutral	Emoticon	Positive	Negative	Neutral
😕	0.08	0.79	0.13	😲	0.68	0.05	0.27
😒	0.08	0.79	0.13	🎈	0.60	0.04	0.36
😩	0.08	0.79	0.13	🎁	0.66	0.11	0.23
😐	0.06	0.75	0.19	🏆	0.62	0.09	0.29
😫	0.10	0.78	0.12	❤	0.61	0.09	0.30
☹	0.13	0.77	0.09	💕	0.63	0.11	0.26
😑	0.08	0.72	0.20	🎉	0.62	0.11	0.27
😣	0.15	0.77	0.08	❤	0.63	0.13	0.24
👎	0.13	0.73	0.15	💃	0.62	0.12	0.25
🔫	0.12	0.72	0.16	💗	0.59	0.12	0.29

Figure 4.1: Top and Bottom from Emoji Sentiment Rankings V1

4.1.4 TF-IDF and minimum document count

Some words are more important that others even if they have low occurrence frequencies and some words are very frequent but are not as important. To achieve a right balance of important words and decently frequent words selecting the right values of TF-IDF is important. User comments also have various typos which could be reduced by setting a minimum document count. For the dataset selected above a TF-IDF between 8 and 10 and a minimum document count of 32 seemed right. The word count distribution before and after filtering can be seen from the graph in Appendix B.6 and B.7

4.1.5 Iterations

In Fig. 3.1, the steps before Sentiment Classifier, runs once and the steps after and including Sentiment Classifier runs for a sets number of times. The iteration value was set to 20 because after this many iterations, changes in ngram strengths were seen to not change much. This can be seen in both Figures. 4.2 and 4.3

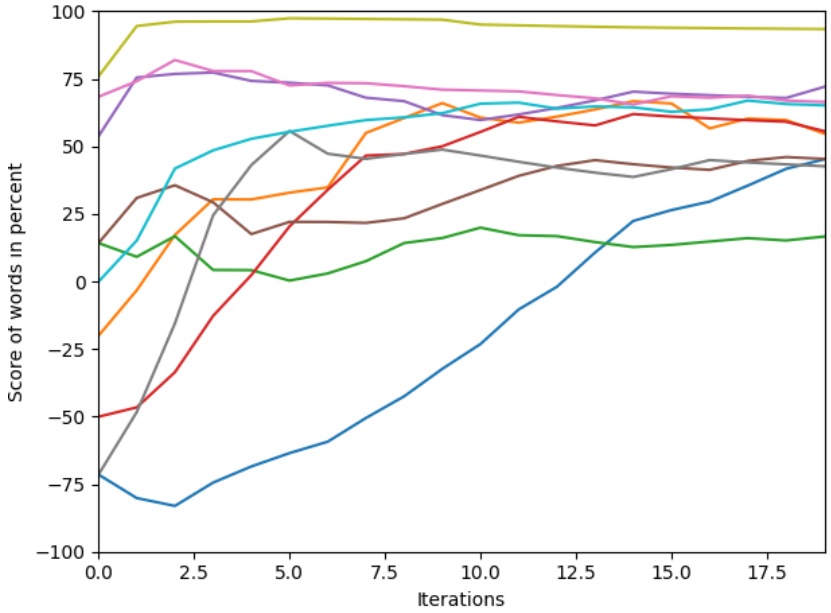

Figure 4.2: Score change over 20 iterations in 10 random words found positive

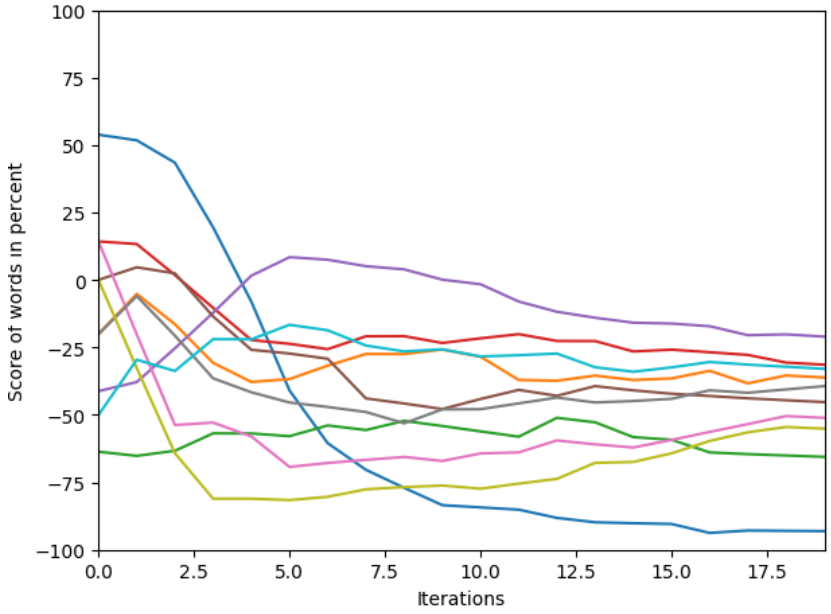

Figure 4.3: Score change over 20 iterations in 10 random words found negative

4.1.6 Ngram Composition

The tokens used are composed of 1-gram, 2-gram and 3-grams. The process is documented in Section 3.3.3.2. Various ngram techniques were attempted shown from . The meaning of these tests will be explained in the Evaluation section.

The objective of this test was to see whether tokenizing to more than 1 gram produced better results. The tokenization methods tried as as follows:

- Unigram Method: The comments were first broken down into sentences and the sentences were broken down into words.
- Bigram Method: The comments were first broken down into sentences and the sentences were broken down into bigrams. Bigrams are constituted from two consecutively occuring words
- Trigram Method: The comments were first broken down into sentences and the sentences were broken down into trigrams. Trigrams are constituted from three consecutively occuring words
- Combined Method: The lexicon obtained in the previous steps were combined
- Ngram Lexicon Method 1: All frequent[1]trigrams are extracted from a sentence first. All frequent[1] bigrams are extracted from the remainder. Lastly all frequent unigrams are extracted from the remainder
- Ngram Lexicon Method 2: All frequent trigrams are extracted from a sentence first. All frequent bigrams are extracted from the sentence (not from the remainder). Lastly all frequent unigrams are extracted from the sentence

Figure C.1 shows that tokenizing the comments into bigrams gave better test results than tokenizing into unigrams. Trigram results were however inferior to unigram and bigram approaches. This was possibly because trigram count was much lower than the unigram and bigram counts after being put through TF-IDF and minimum document count filter.

When the three lexicons were combined, the results were slightly inferior to the bigram lexicon method. Ngram Lexicon Method 2 was seen to be the best and was used in subsequent tests.

[1]Frequent refers to the words seen in at least 32 documents and having TF-IDF between 8 and 10

4.1.7 Build Filters

Appendix B shows the effects of applying various filters during the Lexicon Builder stage (See Section 3.3.6). The testing procedure will be explained in the Evaluation section.

- Figure B.2 shows that applying a score filter may or may not improve the results and the best results are obtained when ngrams with absolute value of score > 4/8 is taken.
- Figure B.3 shows that applying a range filter generally improves test results. Applying a range filter with range > 0.9 improve precision by a large margin but decreases accuracy. Range > 0.05 was selected
- Figure B.4 shows that applying a skewness filter can improve test results. Ngrams with absolute value of skewness > mean skewness / 8 was selected
- Figure B.5 shows that applying a count filter can improve test results. Count > mean count /4 was selected as the filter for the experiments

4.2 Validation Technique

Owing to the absence of pre-existing opinion lexicons in Nepali, the developed lexicon cannot be tested directly. An indirect approach can be taken where the opinion lexicon is used to perform sentiment analysis on a known dataset and the outcome is studied.

Two such tests were used. The first test data consisted of a subset of the microblog dataset. This test data shows how good is the generated lexicon in classifying microblogs. The second test data contained meanings of top positive and negative words from SentiWordNet. Most tokens in this test data are domain independent in their sentiments. Mosts tests conducted have used the first test data. This is because the second test data was put together only at the later stages of this research.

4.2.1 Test Data 1 - Microblog Subset

Table 4.5: Sample Test Data (Microblog Subset)

Text	Sentiment
म मंगल ग्रह बाट सुन्दै छु	0
प्रधानमंत्री को भारत यात्रा ले नेपाळ को फाइदा होइन नोकसान मात्र हुने छ किनकी देउवा भारत को शिष्य हो। कुनै पनि संधि समझेउता नग्रोस भारत संग	-1
हाम्रो देशको बाढी पहिरो कस्ले हेरिदिने कस्ले बुझिदिने ??	-1
धन्य हो हाम्रो नेपाली सेना सबै थाउ मा उहाँ हरू सहयोग ले नै देस बचि राछ मेरो सलाम छ	1

The test data was developed by randomly labelling devanagari microblogs. 109686 microblogs were available for developing the lexicon. Out of these, 1059 were randomly selected for creating a test dataset. The comments were manually read and labelled 1 for Positive and -1 for Negative. Those comments labelled 0 comprise of microblogs which were neutral or comments which could not be classified as positive or negative.

4.2.1.1 Properties

The test data overall was of the following nature:

- Total Unique Tokens: 10277
- Total 1 grams matching TF-IDF criteria : 4546
- Total 2 grams matching TF-IDF criteria : 2195
- Total 3 grams matching TF-IDF criteria : 246
- Positive Comments (labelled 1) : 353
- Negative Comments (labelled -1): 353
- Unknown and Neutral Comments (labelled 0): 353

4.2.2 Test Data 2 - Word Meanings

Table 4.6: Sample Test Data (Word Meanings)

Word	Text	Sentiment
controvert	वाद विवाद गर्नु;खण्डन गर्नु;अस्वीकार गर्नु;विवाद गर्नु;खण्डन गर्नु	-1
convivial	रमाइलो माहोल;बातावरण वा उत्सव;उत्सव-सम्बन्धी;खुशीको;उत्सव-सम्बन्धी	1
coolie	कुल्ली;भरिया;कुली (m);मजदूर	-1
cooly	कुली	-1
coquetry	नखरेबाजी (f);सौकीन	1
cordless	तार रहित;बेतार	-1
corking	धेरै राम्रो	1

This test data was developed by first selecting top positive and negative words from SentiWordNet then translating the English words to their individual meanings in Nepali. For translation, englishnepalidictionary.com and shabdakosh.com.np meanings were merged together. A sample of this test data can be seen in Table 4.6

4.2.2.1 Properties

The test data overall was of the following nature:

- Total Unique Tokens: 4192
- Total 1 grams matching TF-IDF criteria : 1160
- Total 2 grams matching TF-IDF criteria : 66
- Total 3 grams matching TF-IDF criteria : 1
- Positive Test Data (labelled 1) : 595
- Negative Test Data (labelled -1): 595
- Unknown and Neutral Comments (labelled 0): 0

4.2.3 Confusion Matrix

Since the terms ‘Positive’ and ‘Negative’ coincide with the terminology used in confusion matrix, the following labels are set for the classes:

- Letter A for Negative Class
- Letter B for Positive Class
- Letter C for Unknown Class

Since the classification is not binary, multiple precision, recall, fscores and accuracy values are computed. Also an overall accuracy value and a mean precision value is calculated to represent the overall effectiveness of the classification.

$$Precision = \frac{TP}{TP + FP} \tag{4.1a}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.1b}$$

$$Fscore = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4.1c}$$

$$Accuracy = \frac{TP + FN}{TP + FP + TN + FN} \tag{4.1d}$$

where,

$$TP = TruePositive, \qquad TN = TrueNegative,$$
$$FP = FalsePositive, \qquad FN = FalseNegative$$

Formula 4.1 are used for the calculations for each class, whereas, formula 4.2 are used for overall calculations.

$$OverallAccuracy = \frac{Sumofdiagonals}{OverallSum} \tag{4.2a}$$

$$MeanPrecision = \sqrt{Precision(A) * Precision(B)} \tag{4.2b}$$

Formula 4.2a takes into account all three negative (A), positive (B) and unknown (C) classes. Overall Accuracy determines how good is the classifier at labelling microblogs to their correct classes.

The formula 4.2b only takes negative (A) and positive (B) classes into account and ignores unknown class (C). This is because the lexicon developed is only able to tell positive and negative terms and was not designed for neutral terms. A microblog is labelled C when it is unable to accurately state whether it is of class A or of class B. Also, emphasis is placed on precision rather than recall because a small but correct lexicon would be preferred rather than a large but

incorrect one. A geometric mean ensures that the precision of both A and B classes should be high and not just one of them.

As there are too many metrices, the comparisions henceforth will be simplified by focusing mostly on Overall Accuracy(4.2a) and Mean Precision (4.2b).

4.2.3.1 Random Classification

In order to understand what the numbers mean, a frame of reference is necessary. A system that randomly classifies microblogs to A,B and C classes can be used as the reference.

Table 4.7 shows the confusion matrix and corresponding f-score and accuracy calculations for a randomly classified test data. As there are 3 classes, the overall accuracy and the mean precision are almost equal to a third of 1 (33%).

Table 4.7: Confusion matrix of randomly classified test data

		PREDICTED			
	Class	A	B	C	Total
ACTUAL	A	116	100	137	353
	B	103	122	128	353
	C	123	118	112	353
	Total	342	340	377	
	Class	TP	TN	FP	FN
	A	116	480	226	237
	B	122	488	218	231
	C	112	441	265	241
	Class	Precision	Recall	F-score	Accuracy
	A	0.34	0.33	0.33	0.56
	B	0.36	0.35	0.35	0.58
	C	0.3	0.32	0.31	0.52
	Overall Accuracy	0.33	Mean Precision	0.35	

4.2.3.2 Limitations

The method suggested for evaluating the obtained lexicon is limited because of indirect approach taken. The generated lexicon is not perfect but the indirect

approach makes it look worse.

The method that will be used performs sentiment analysis on the test data using the obtained lexicon. The sentiment analysis assumes that the tokens in the test data are bag of words[1], which does not always hold.

There are also numerous spelling mistakes in the testdata which will not be corrected. Section 4.2.1.1 shows that out of 10277 unique tokens in the test data, only 4546 unigrams have correct spelling (as per the TF-IDF and minimum document count) criteria. This means that on average, only about 40% unigrams will be useful. Similarly in 4.2.2.1 it can be seen than out of 4192 unique tokens 1160 (less than 30%) unigrams are useful in the test data

The error inherent to the sentiment analysis procedure and the error due to misspellings is expected to decrease the reported accuracy and precision during the tests. As such the tests cannot be used to exactly state the correctness of the lexicon obtained. However, it can be good enough for comparing variations in parameters.

[1]Bag of words assumes that words/tokens occur independent of each other. This assumptions allows determining the sentiment of a text using the sentiment of words/tokens it contains regardless of the order in which the words/tokens appear

Chapter 5

RESULTS AND DISCUSSION

This chapter will discuss the findings from the experiments talked about in Chapter 4. Test results and outcomes of the experiments are included in the Appendix D

5.1 Top Words Found

Top 30 positive and negative words obtained after each experiment listed in D.2. In the untruncated version, Emoticon approach built 4004 total ngrams; 2-word approach built 4237 ngrams; and 32-word approach built 3924 ngrams.

The words can also be visualized from the word clouds in Appendix D.3. Note that the plotting tool plotted ि for all ी . This appears to be a bug in the plotting tool. The following subsections will focus on the top 30 list and not the word clouds.

5.1.1 Mistakes

5.1.1.1 Proper nouns

References to proper nouns such as organisations and people can be seen : बिबिसी.नेपाली, बाबुराम.भट्टराई.ले, ने.क.पा, नारायण.जी, साझा.सवाल.कार्यक्रम etc. It appears that the lexicon building algorithm misclassifies propernouns which are dominantly used in one of positive or negative contexts.

5.1.1.2 Neutral words

Some positive terms such as उहाँहरु, गर्छु, छ.हजुरलाई, गर्नु.हुने, यदि.तपाई.लाई can actually be used in positive, negative and neutral contexts but are labelled positive.

Similarly some negative terms such as तिमिहरुलाइ, गर्छस, गए.हुन्छ, के.गर्छौं, जानेको.छ, लगायो are also context specific but are labelled negative. Nepali language has separate words to address seniors and older people (e.g. तपाई, हजुर etc), and juniors and younger people (e.g. तिमी, तं etc). Some words are used more when used in derogatory contexts despite the fact that they independently don't have any negative meanings (E.g. तं, तेरो). This can be seen reflected in the top words.

5.1.1.3 Ngram subsets

If a ngram of a certain size is frequent, ngrams that are contained in the larger ngrams are also frequent. This may have been reflected in the top 30s. For example गर्दछु may have been reported as positive because it occurs frequently within ngrams such as शुभकामना.व्यक्त.गर्दछु or श्रद्धान्जली.व्यक्त.गर्दछु. The ngrams बि.बि and सी.नेपाली were reported as positive because they were segments of a larger ngram बि.बि.सी.नेपाली. Simarly सवालको must have been a part of साझा.सवालको. These issues can be fixed by changing the way sentences are tokenized as talked about in Section 3.3.3.2

5.1.2 Achievements

5.1.2.1 Agreement between experiments

Positive Unigrams दामि, उत्तरोत्तर, हिममत, सून्दर, सूपार, लक and बदाई occur in all three lexicons. Negative Unigram कौडिको occur in all three lexicons. Positive Bigram धेरै.शुभकामना occurs in all three lexicons. Positive trigrams बि.बि.सी, बधाई.तथा.सफल, रेखा.मैसाप.जी and धेरै.धेरै.शुभकामना occur in all three lexicons. Negative trigrams राजा.महेन्द्र.ले, कानमा.तेल.हालेर and रगत.र.पसिना occur in all three lexicons. Around 20% ngrams can be found in the top 30 of at least two experiments. This shows that there is a some agreement with regards to top words even when three different seed lexicons were used.

5.1.2.2 Domain Independent Words

While mosts ngrams classified appear correct under context, some ngrams were also correct independent of the context. For example positive unigrams लाभको, दामि, उत्तरोत्तर, हिममत, अभिवादन, अगृम, सुखलाई, क्षमतावान, सून्दर, शिघ्र, सूपार, शान्तीको, श्रद्धान्जली, सफल्ता, लक etc are definitely correct while others like टीम, शिघ्र, उहाँहरु,

पराएको, गर्छु etc are context specific. The context specific positive words occurred in the Top 30 because they were seen used more along with other postive words.

On manually counting the following was revealed:

- On using Emoticon Seed Lexicon, 41% positive, 23% negative and 32%overall ngrams were found to be domain independent and correct
- On using 32-word Seed Lexicon, 44% positive, 17% negative and 31%overall ngrams were found to be domain independent and correct
- On using 2-word Seed Lexicon, 41% positive, 12% negative and 27%overall ngrams were found to be domain independent and correct

The positive correct and domain independent percentage was higher in the top 30 words compared to it's negative counterpart. Overall, Emoticon Seed Lexicon gave the most domain independent and correct words.

5.1.2.3 Proverb Pieces

Some trigrams can be seen as popular proverbs or proverb pieces such as सबैलाई.चेतना.भया, जोगी.आए.पनि, आखामा.छारो.हालेर, कानमा.तेल.हालेर. No such proverb pieces were seen in the top 30 positive context

5.2 Test Results

Filters played a significant role in improving the test results. This can be seen from the figure in Appendix B.1. Another important factor contributing to accuray and precision could have been the effect of data size. To check this multifold experiments were conducted shown in Section 5.2.3. In almost all accuracy-precision graphs, test data 1 was used. This is because test data 2 was only introduced later in the research to see how domain independent words were classified in the generated lexicon. The following two sections compare primary/seed lexicons in classified test data.

5.2.1 Test Data 1

Figure 5.1 shows that in the test results of the experiments conducted, results from Emoticon Seed Lexicon was the best , and the results from 2-word seed

lexicon was the worst. However the differences were not too large; the maximum difference was just around 5%. This shows that all three lexicons perform similarly in terms of accuracy and precision.

The test data contained equal portions of positive, negative and unknown class data. However, Figure 5.2 shows that the classifier classfied much of the unknown test dataset as Positive or Negative. This indicates that the classifier is not good at predicting unknown or neutral comments.

The full confusion matrix of the experiments can be seen in Appendix D.1.

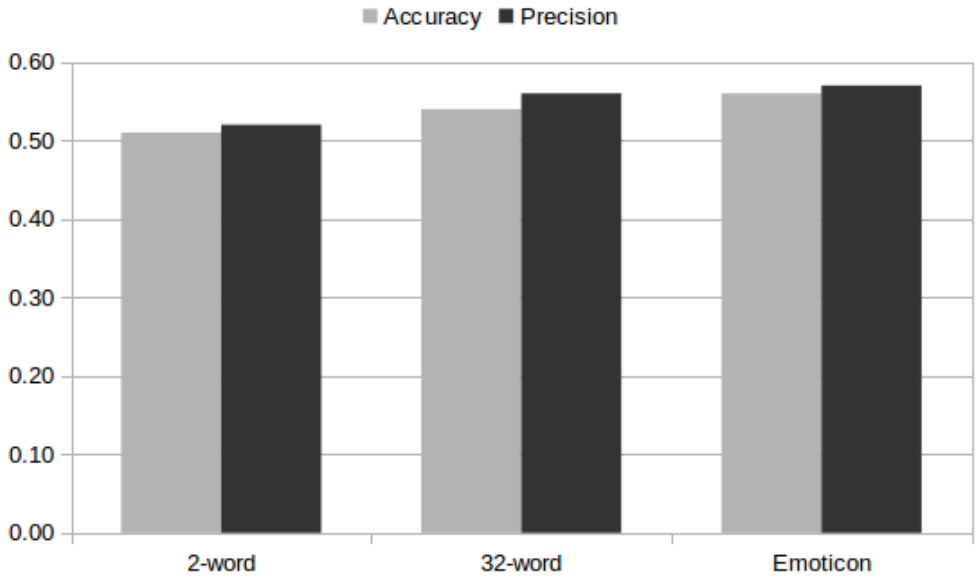

Figure 5.1: Emoji lexicon vs Manually developed text lexicon used as primary lexicon. Tested on Test Data 1

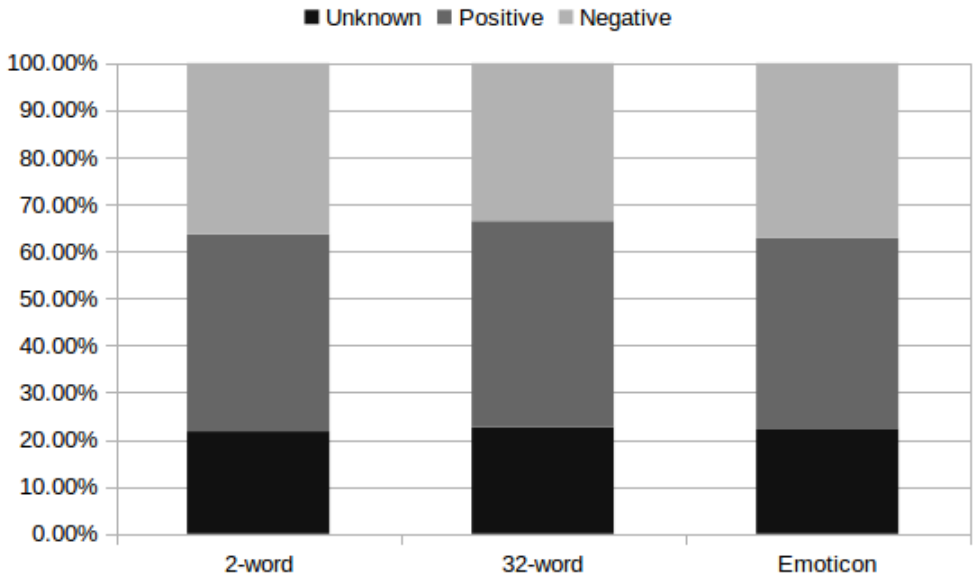

Figure 5.2: Percent of Positive, Negative and Unknown classifications using Test Data 1

5.2.2 Test Data 2

Figure 5.4 shows that the output of all three seed lexicons gave good precision and poor accuracy when the generated lexicon was used to classify Test Data 2. Low acccuracy shows that overall classification of test data was mostly incorrect. High precision shows that out of the test data that were classified as positive and negative, most of them were correct.

Test Data 2 is different that Test Data 1 in it's proportion of actual Positive, Negative and Unknown class data. Test Data 2 does not have any data labelled unknown. Test Data 2 is also not representative of the vocabulary used in microblogs as it contains definitions of terms and word meanings instead of user comments which were contained in Test Data 1. As such microblog specific generated lexicon is unable classify most of the data in Test Data 2. This is shown in Fig. 5.3 where around 60% of the data is classified as unknown.

The generated lexicon was built using microblog vocabulary, whereas Test Data 2 contains vocabulary used in dictionaries. Hence most words in Test Data 2 were not contained in the generated lexicon resulting in high percentage of unknowns in the test results. The good precision in the classification shows that the domain independent terms in the generated lexicon have correct polarities.

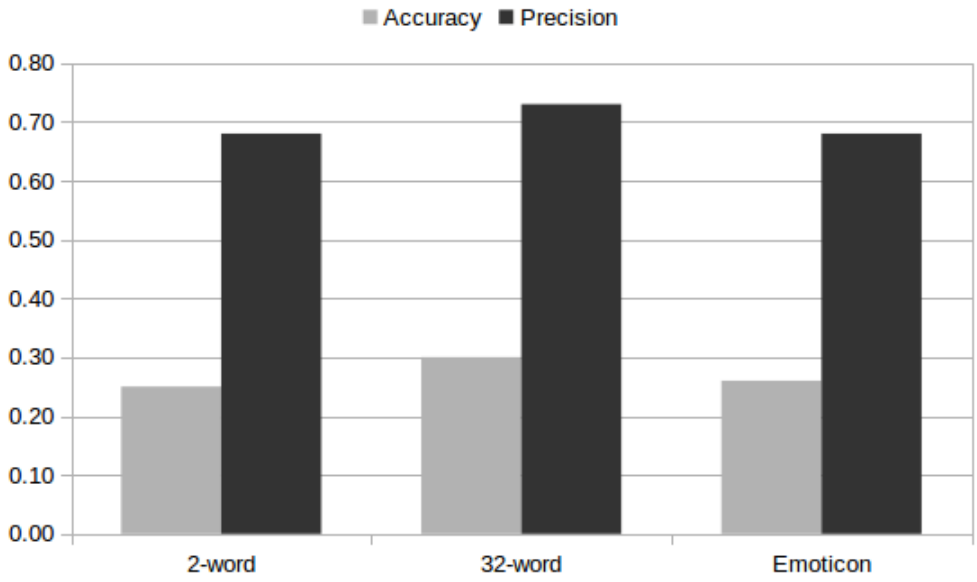

Figure 5.3: Emoji lexicon vs Manually developed text lexicon used as primary lexicon. Tested on Test Data 2

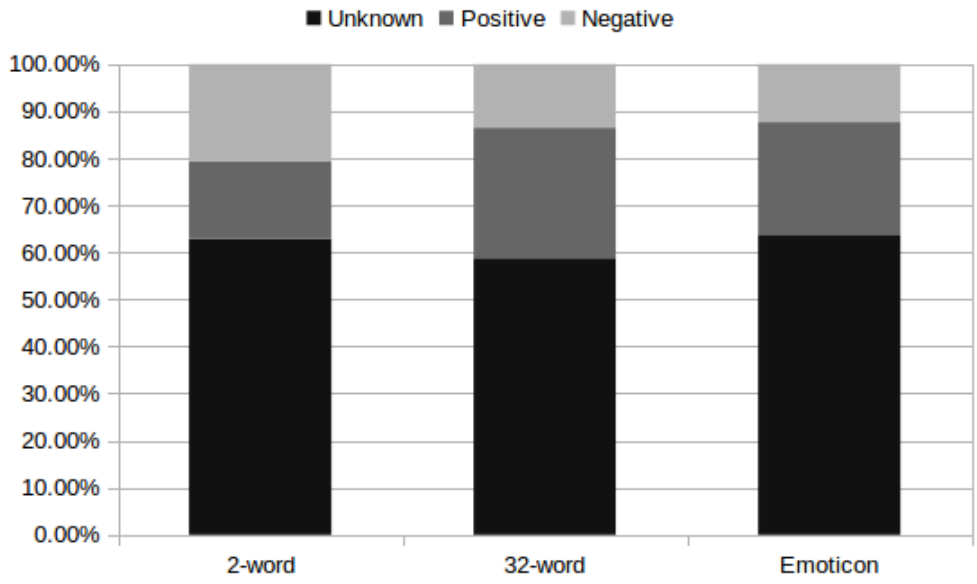

Figure 5.4: Percent of Positive, Negative and Unknown classifications using Test Data 2

5.2.3 Multifold Experiments

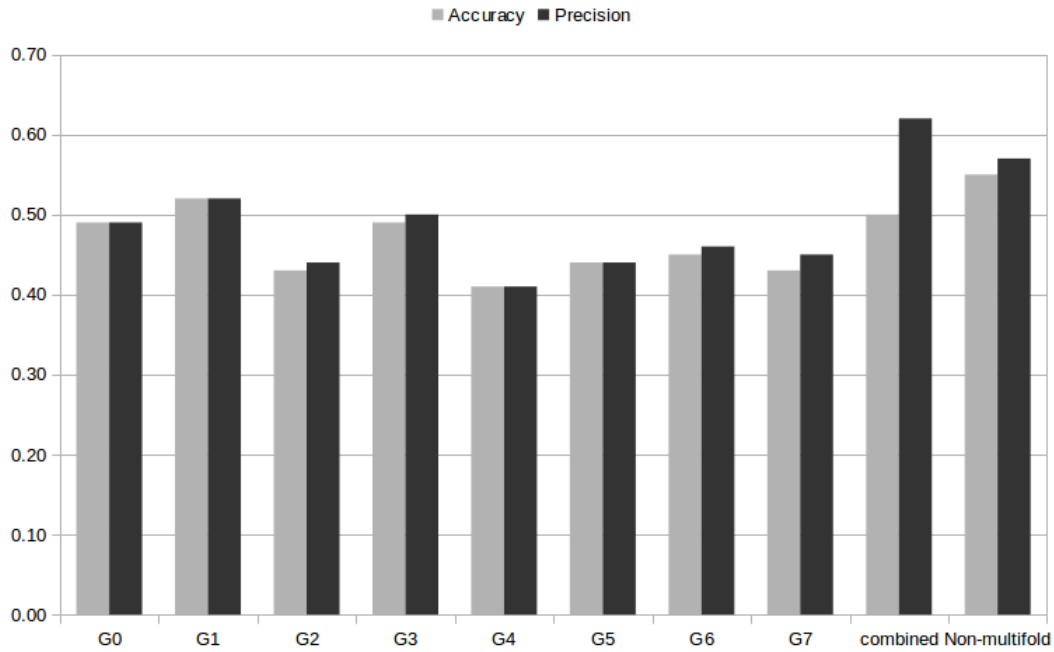

Figure 5.5: Effect of splitting the dataset into 8 groups and combining

The first objective of these tests were to see whether the size of data used to build the lexicon makes a difference. The other objective was to see if results from multiple folds can be averaged to make the generated lexicon better.

In Fig 5.5, G0 through G7 represent the 8 folds that the dataset was broken down into. The test results of these individual folds are inferior to the non-multifold experiment shown in the last column. Merging the lexicon from the folds and averaging the score obtained produced a combined lexicon which was superior in precision but inferior in accuracy compared to the non-multifold experiment.

Chapter 6

CONCLUSIONS

6.1 Pros and Cons

The research attempted to build a Nepali language lexicon using microblog data. This was done in absense of resources such as POS taggers, without making any assumption about the semantics of Nepali language and with a simple Bag of Words assumption. It was seen that with even a simple 2-word seed lexicon the process can generate a large number of positive and negative ngrams.

The lexicon obtained after this research is domain-specific and specific to the language used by commenters on social media platforms such as facebook. A qualitative analysis shows that this lexicon can give insights on not just words, but popular phrases, organisations and people. It can give an idea about how commenters on the internet perceive such entities. Since the output lexicon captures the nuances of language used my most users online, it's usage is not only limited to sentiment analysis. The output lexicon if studied along with a temporal dimension can be used to see how Nepali language evolves over time.

The disadvantages of the lexicon obtained is that only a small percentage of ngrams obtained in the lexicon was domain independent and correct, i.e. correct regardless of the situation in which the ngrams are used. The fact that the lexicon is unable to distinguish parts of speech such as proper nouns from truely subjective terms is another drawback.

6.2 Conclusions

The primary objective of this research was to develop a method for generating opinion lexicon for text available in microblogs. The research uses a statistical

approach instead of using a machine learning approach. The research focused on Devanagari Nepali text and used text and emoticon based seed lexicons to achieve the objectives. At the same time, statistical filters involving skewness, range, minimum count and score related metrices were appended to the process and shown to work.

A lexicon containing around 4000 words were produced using each seed lexicon. Table 6.1 shows a sample of the generated lexicon. The words are the same as in the 32 word seed lexicon. The table only contains 1-grams whereas the actual generated lexicon contains 1,2 and 3 grams. Each ngram in the lexicon is given a score that ranges between -1 and 1.

Table 6.1: Sample Output Lexicon

Word	Score	Word	Score	Word	Score	Word	Score
अनियमित	-1	हस्तक्षेप	-0.9	नियमित	0.84	सक्षम	0.91
हचुवा	-1	भ्रष्ट	-0.89	शुद्ध	0.86	राम्रो	0.92
आतंक	-0.97	लुटेरा	-0.89	चम्किलो	0.87	सभ्य	0.92
अपहरण	-0.93	जबरजस्ती	-0.81	शान्त	0.87	व्यवस्थित	0.93
चोर	-0.93	वाक्क	-0.77	स्वच्छ	0.89	निष्पक्ष	0.95
तोडफोड	-0.93	नकारात्मक	-0.65	सफा	0.9	स्वस्थ	0.95
फटाहा	-0.91	नराम्रो	-0.25	ताजा	0.91	पौष्टिक	1
सुस्त	-0.91	झुर	-0.15	शान्ति	0.91	विश्वसनीय	1

6.2.1 Further Work

It was shown that the size of data does impact the accuracy and precision of the lexicon obtained. So one way to improve the lexicon would be to gather more data. Another way is to autocorrect various obvious spelling mistakes in the data which in effect would be the same as increasing the data size. Comments from 18 facebook pages were scraped for this research and more can be achieved. In addition to microblogs other texts can also be used.

It is hard to determine the true accuracy of the proposed method because of lack of reference to compare against. An indirect method was chosen by performing sentiment analysis with the output lexicon. The limitations of this method are that the accuracy reported is only an approximation. Even an 100% correct lexicon may not be able to predict sentiments of test data perfectly. The

proposed method is not the only way to build an opinion lexicon. Other methods should be explored and the results of the future lexicon can be compared with the lexicon from this research. To produce a domain independent lexicon a dictionary or WordNet based approach should be taken.

A lexicon is not the only way to perform sentiment analysis. A lexicon contains scores for words and phrases in an isolated sense. When using a domain specific lexicon the word गर्दछु comes up as positive because it is usually seen to be used in a positive context. But to use गर्दछु with a negative word such as बदमासी making बदमासी गर्दछु is also possible. This shows that measuring the sentiment of words and phrases in an isolated way is not the best way. In the future, a lexicon building approach incorporating combinations of words and phrases and word patterns used to express various sentiments (seen in [35]) would be more useful that using a purely statistical approach.

BIBLIOGRAPHY

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In Proceedings of the workshop on languages in social media, pages 30--38. Association for Computational Linguistics, 2011.

[2] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In LREC, volume 10, pages 2200--2204, 2010.

[3] Bal K Bal. Computational linguistic model for analyzing opinionated texts. PhD dissertation, Kathmandu University, 2015.

[4] Lee Becker, George Erhart, David Skiba, and Valentine Matula. Avaya: Sentiment analysis on twitter with self-training and polarity lexicon expansion. In Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), volume 2, pages 333--340, 2013.

[5] Fermín L Cruz, José A Troyano, F Javier Ortega, and Fernando Enríquez. Automatic expansion of feature-level opinion lexicons. In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, pages 125--131. Association for Computational Linguistics, 2011.

[6] Mohammad Darwich, Shahrul Azman Mohd Noah, and Nazlia Omar. Minimally-supervised sentiment lexicon induction model: A case study of malay sentiment analysis. In International Workshop on Multi-disciplinary Trends in Artificial Intelligence, pages 225--237. Springer, 2017.

[7] Misbah Daud, Rafiullah Khan, Aitazaz Daud, et al. Roman urdu opinion mining system (ruomis). arXiv preprint arXiv:1501.01386, 2015.

[8] Xiaowen Ding, Bing Liu, and Philip S Yu. A holistic lexicon-based approach to opinion mining. In Proceedings of the 2008 international conference on web search and data mining, pages 231--240. ACM, 2008.

[9] Tang Duyu, Qin Bing, Zhou LanJun, Wong KamFai, Zhao Yanyan, and Liu Ting. Domain-specific sentiment word extraction by seed expansion and pattern generation. arXiv preprint arXiv:1309.6722, 2013.

[10] Hady ElSahar and Samhaa R El-Beltagy. A fully automated approach for arabic slang lexicon extraction from microblogs. In International Conference on Intelligent Text Processing and Computational Linguistics, pages 79--91. Springer, 2014.

[11] A Esuli. et f sebastiani.<<<<. SentiWordNet: a Publicly Available Lexical Resource for Opinion Mining, 2006.

[12] Chadan P Gupta. Sentiment analysis of nepali texts. Master's thesis, Kathmandu University, 2011.

[13] Chandan Prasad Gupta and Bal Krishna Bal. Detecting sentiment in nepali texts: A bootstrap approach for sentiment analysis of texts in the nepali language. In Cognitive Computing and Information Processing (CCIP), 2015 International Conference on, pages 1--4. IEEE, 2015.

[14] William L Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. Inducing domain-specific sentiment lexicons from unlabeled corpora. arXiv preprint arXiv:1606.02820, 2016.

[15] Vasileios Hatzivassiloglou and Kathleen R McKeown. Predicting the semantic orientation of adjectives. In Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, pages 174--181. Association for Computational Linguistics, 1997.

[16] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 168--177. ACM, 2004.

[17] Dan Jurafsky and James H Martin. Speech and language processing, volume 3. Pearson London:, 2014.

[18] Hiroshi Kanayama and Tetsuya Nasukawa. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In Proceedings of the 2006 conference on empirical methods in natural language processing, pages 355--363. Association for Computational Linguistics, 2006.

[19] Muhammad Yaseen Khan, Shah Muhammad Emaduddin, and Khurum Nazir Junejo. Harnessing english sentiment lexicons for polarity detection in urdu tweets: A baseline approach. In Semantic Computing (ICSC), 2017 IEEE 11th International Conference on, pages 242--249. IEEE, 2017.

[20] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In Proceedings of the 20th international conference on Computational Linguistics, page 1367. Association for Computational Linguistics, 2004.

[21] Efthymios Kouloumpis, Theresa Wilson, and Johanna D Moore. Twitter sentiment analysis: The good the bad and the omg! Icwsm, 11(538-541):164, 2011.

[22] Bing Liu and Lei Zhang. A survey of opinion mining and sentiment analysis. In Mining text data, pages 415--463. Springer, 2012.

[23] Kun-Lin Liu, Wu-Jun Li, and Minyi Guo. Emoticon smoothed language models for twitter sentiment analysis. In Aaai, 2012.

[24] Raheleh Makki, Stephen Brooks, and Evangelos E Milios. Context-specific sentiment lexicon expansion via minimal user interaction. In Information Visualization Theory and Applications (IVAPP), 2014 International Conference on, pages 178--186. IEEE, 2014.

[25] George A Miller. Wordnet: a lexical database for english. Communications of the ACM, 38(11):39--41, 1995.

[26] Sascha Narr, Michael Hulfenhaus, and Sahin Albayrak. Language-independent twitter sentiment analysis. Knowledge discovery and machine learning (KDML), LWA, pages 12--14, 2012.

[27] Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. Sentiment of emojis. PloS one, 10(12):e0144296, 2015.

[28] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In LREc, volume 10, 2010.

[29] Veronica Perez-Rosas, Carmen Banea, and Rada Mihalcea. Learning sentiment lexicons in spanish. In LREC, volume 12, page 73, 2012.

[30] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Expanding domain sentiment lexicon through double propagation. In IJCAI, volume 9, pages 1199--1204, 2009.

[31] Pavel Smrž. Using wordnet for opinion mining. In Proceedings of the Third International WordNet Conference, pages 333--335, 2006.

[32] Marlo Souza, Renata Vieira, Débora Busetti, Rove Chishman, and Isa Mara Alves. Construction of a portuguese opinion lexicon from multiple resources. In Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, 2011.

[33] Duyu Tang, Furu Wei, Bing Qin, Ming Zhou, and Ting Liu. Building large-scale twitter-specific sentiment lexicon: A representation learning approach. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 172--182, 2014.

[34] Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics, pages 417--424. Association for Computational Linguistics, 2002.

[35] Clara Vania, Moh. Ibrahim, and Mirna Adriani. Sentiment lexicon generation for an under-resourced language. Int. J. Comput. Linguistics Appl., 5(1):59--72, 2014.

[36] Payal Yadav and Dhatri Pandya. Sentireview: Sentiment analysis based on text and emoticons. In Innovative Mechanisms for Industry Applications (ICIMIA), 2017 International Conference on, pages 467--472. IEEE, 2017.

Appendices

Appendix A

Primary/Seed Lexicons

Table A.1: Simple 2-word seed lexicon

Word (w)	Positive (p)	Negative (n)
राम्रो	0.9	0.1
नराम्रो	0.1	0.9

Table A.2: Manually developed 32-word seed lexicon

word	positive	negative	neutral	word	positive	negative	neutral
राम्रो	0.9	0.05	0.05	नराम्रो	0.05	0.9	0.05
स्वच्छ	0.9	0.05	0.05	भ्रष्ट	0.05	0.9	0.05
चम्किलो	0.9	0.05	0.05	सुस्त	0.05	0.9	0.05
सभ्य	0.9	0.05	0.05	हचुवा	0.05	0.9	0.05
सफा	0.9	0.05	0.05	अनियमित	0.05	0.9	0.05
सक्षम	0.9	0.05	0.05	झुर	0.05	0.9	0.05
शान्त	0.9	0.05	0.05	जबरजस्ती	0.05	0.9	0.05
पौष्टिक	0.9	0.05	0.05	हस्तक्षेप	0.05	0.9	0.05
ताजा	0.9	0.05	0.05	नकारात्मक	0.05	0.9	0.05
निष्पक्ष	0.9	0.05	0.05	फटाहा	0.05	0.9	0.05
विश्वसनीय	0.9	0.05	0.05	लुटेरा	0.05	0.9	0.05
शान्ति	0.9	0.05	0.05	चोर	0.05	0.9	0.05
शुद्ध	0.9	0.05	0.05	वाक्क	0.05	0.9	0.05
स्वस्थ	0.9	0.05	0.05	आतंक	0.05	0.9	0.05
नियमित	0.9	0.05	0.05	तोडफोड	0.05	0.9	0.05
व्यवस्थित	0.9	0.05	0.05	अपहरण	0.05	0.9	0.05

Emoji	Pos	Neg	Emoji	Pos	Neg	Emoji	Pos	Neg	Emoji	Pos	Neg
	0.68	0.05		0.44	0.18		0.36	0.26		0.26	0.50
	0.60	0.04		0.42	0.16		0.40	0.32		0.24	0.49
	0.66	0.11		0.44	0.18		0.37	0.29		0.24	0.50
	0.62	0.09		0.26	0.01		0.22	0.13		0.15	0.41
	0.61	0.09		0.46	0.21		0.30	0.22		0.24	0.54
	0.63	0.11		0.34	0.09		0.40	0.32		0.20	0.51
	0.62	0.11		0.31	0.07		0.38	0.30		0.22	0.54
	0.63	0.13		0.42	0.18		0.09	0.02		0.22	0.56
	0.62	0.12		0.43	0.19		0.35	0.29		0.23	0.58
	0.59	0.12		0.42	0.18		0.36	0.31		0.16	0.55
	0.60	0.12		0.46	0.23		0.35	0.30		0.18	0.63
	0.58	0.11		0.31	0.10		0.13	0.08		0.19	0.70
	0.60	0.13		0.46	0.25		0.27	0.24		0.13	0.64
	0.57	0.11		0.44	0.23		0.12	0.10		0.15	0.66
	0.62	0.15		0.40	0.21		0.35	0.34		0.13	0.64
	0.57	0.12		0.31	0.13		0.37	0.36		0.15	0.67
	0.59	0.15		0.41	0.23		0.35	0.34		0.15	0.67
	0.53	0.10		0.42	0.25	ZZZ	0.32	0.31		0.15	0.68
	0.47	0.05	TOP	0.38	0.20		0.00	0.01		0.13	0.66
	0.53	0.11		0.42	0.25		0.13	0.14		0.13	0.70
	0.56	0.14		0.41	0.24	✔	0.23	0.26		0.15	0.72
	0.53	0.12		0.22	0.06		0.27	0.30		0.11	0.68
	0.55	0.14		0.43	0.27		0.35	0.38		0.13	0.71
	0.54	0.13		0.40	0.25		0.22	0.28		0.11	0.69
	0.55	0.16		0.33	0.19		0.24	0.31		0.14	0.73
	0.51	0.11		0.19	0.05		0.16	0.24		0.14	0.72
	0.51	0.12		0.26	0.11		0.30	0.39		0.11	0.70
	0.43	0.05		0.30	0.16		0.15	0.24		0.12	0.72
	0.54	0.17		0.37	0.23		0.23	0.33		0.13	0.73
	0.53	0.16		0.32	0.19	©	0.14	0.25		0.15	0.77
	0.50	0.13	✖	0.23	0.09		0.07	0.20		0.08	0.72
	0.55	0.20	★	0.20	0.07	■	0.03	0.17		0.13	0.77
	0.42	0.08		0.19	0.07		0.26	0.40		0.10	0.78
	0.51	0.17		0.24	0.12	!	0.14	0.30		0.06	0.75
	0.45	0.13		0.38	0.26		0.27	0.43		0.08	0.79
	0.51	0.20		0.37	0.26		0.13	0.31		0.08	0.79
	0.50	0.21		0.41	0.30		0.25	0.47		0.08	0.79
☀	0.34	0.07		0.38	0.28		0.14	0.38			

Figure A.1: Complete Emoticon Seed Lexicon (Source: Emoji Sentiment Rankings V1)

Appendix B

Filters

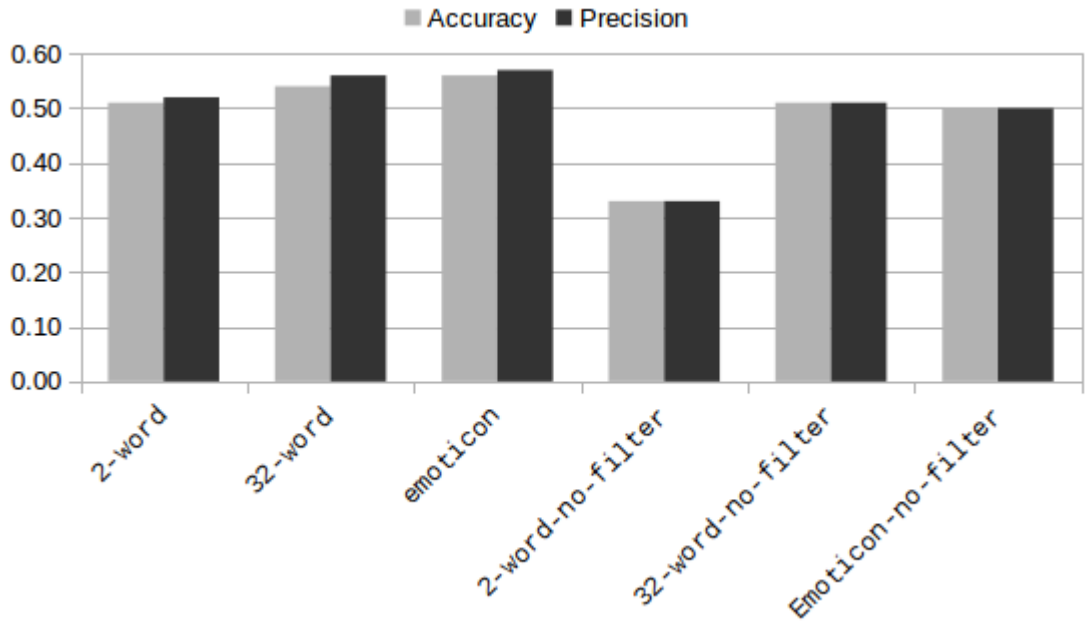

Figure B.1: Results using 3 primary(seed) lexicons with and without filters

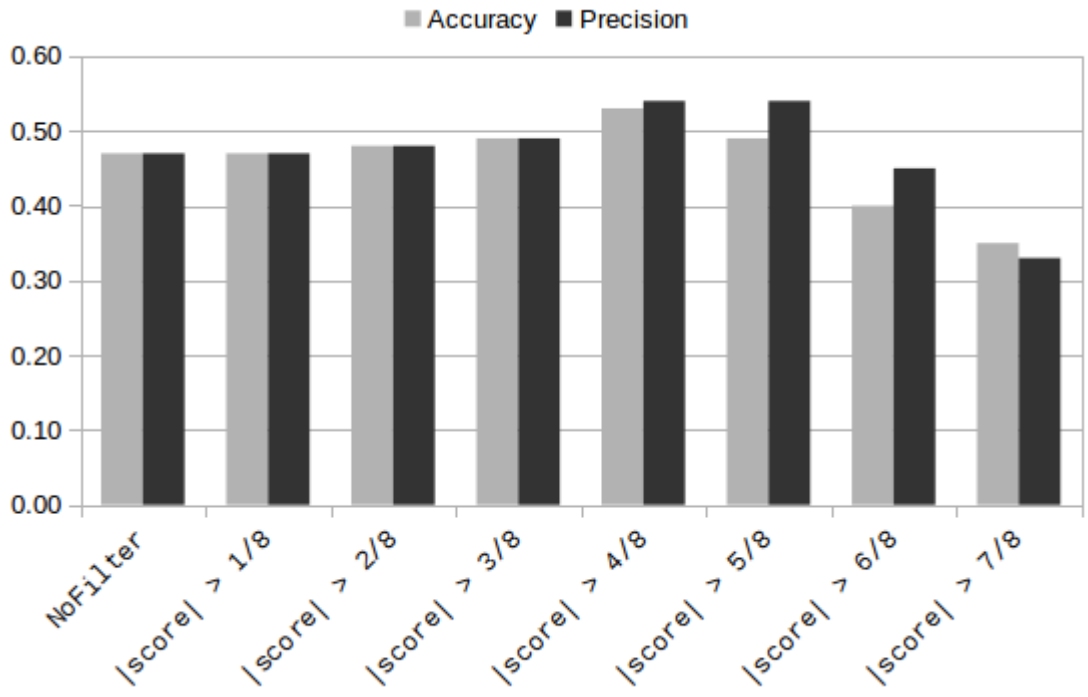

Figure B.2: Results of various Score Filters

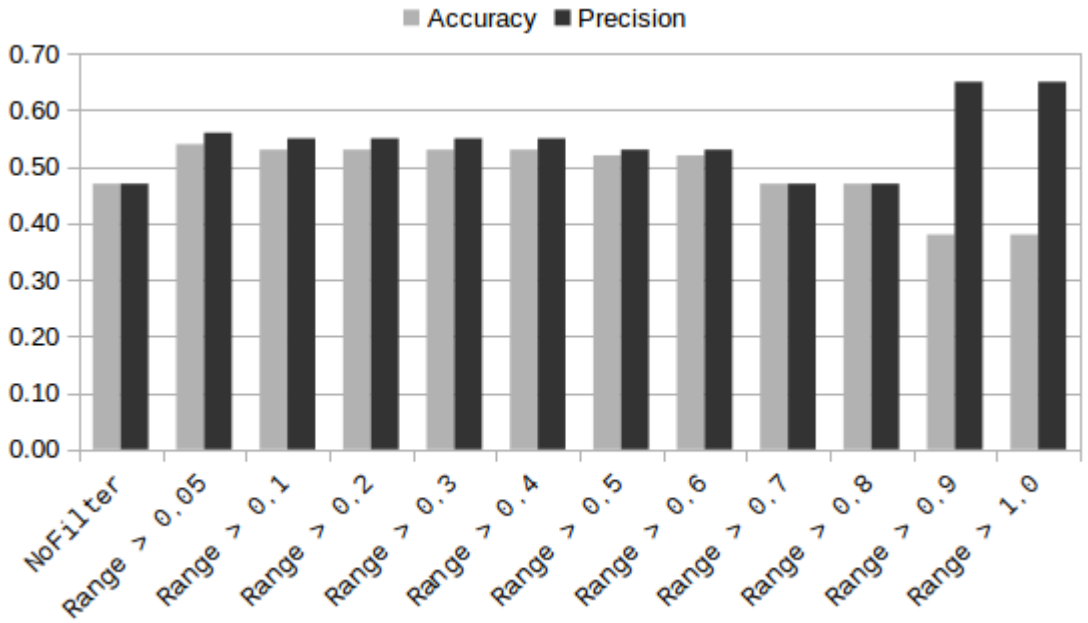

Figure B.3: Results of various Range Filters

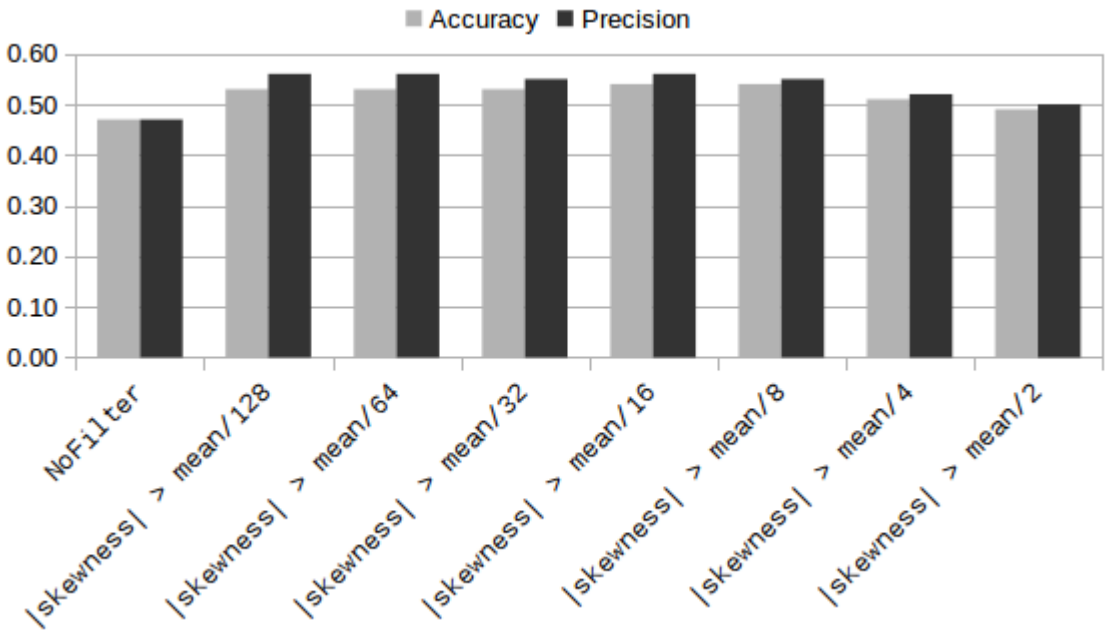

Figure B.4: Results of various Skewness Filters

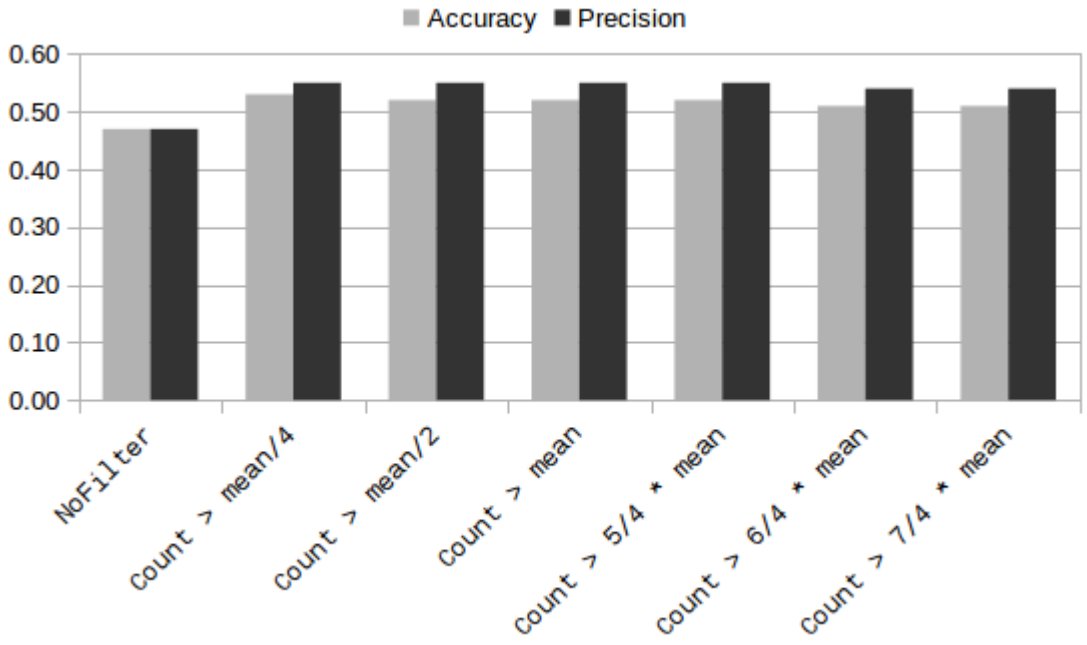

Figure B.5: Results of various Count Filters

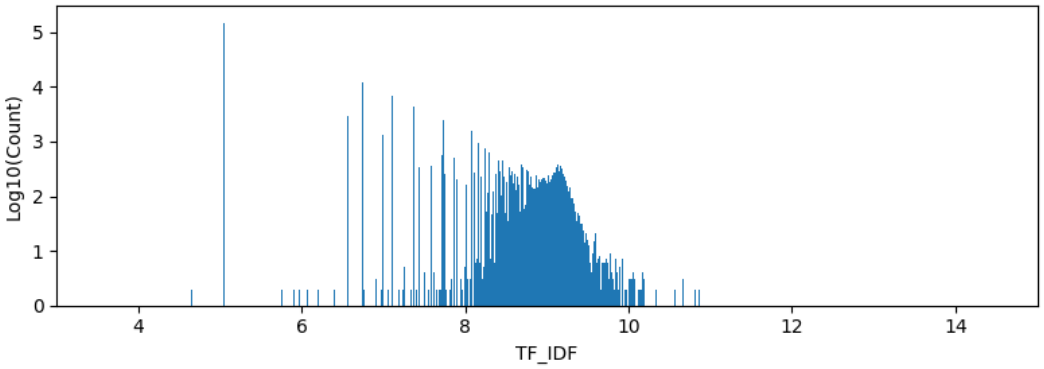

Figure B.6: Word count in log scale vs. Unigram TF-IDF before filters

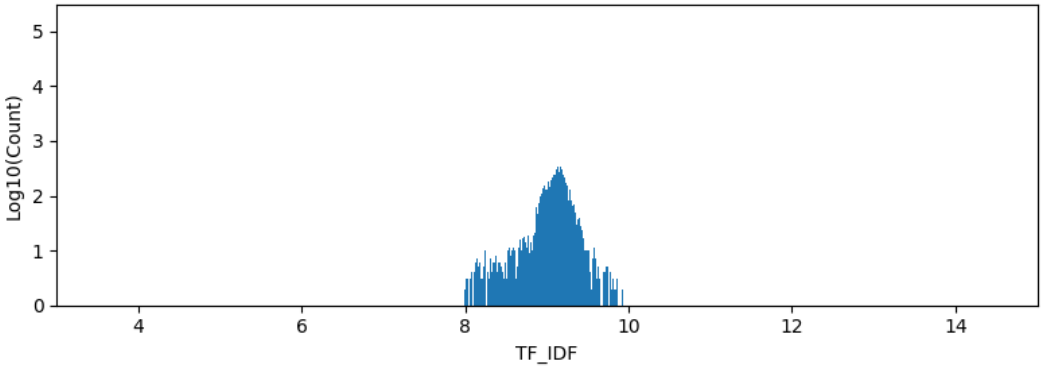

Figure B.7: Word count in log scale vs. Unigram TF-IDF after filters

Appendix C

Miscellaneous

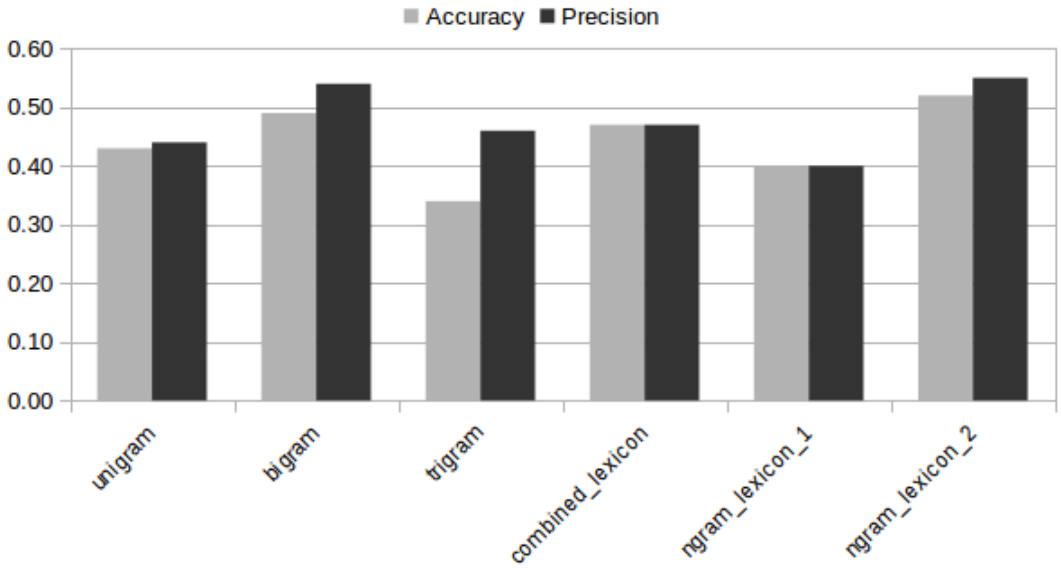

Figure C.1: Effect of various tokenization methods

Appendix D

Test Results

D.1 Confusion Matrices

Table D.1: Confusion matrix of Test data classified using Emoticion Seed Lexicon

		PREDICTED			
	Class	A	B	C	Total
ACTUAL	A	229	70	54	353
	B	49	243	61	353
	C	116	115	122	353
	Total	394	428	237	
	Class	TP	TN	FP	FN
	A	229	541	165	124
	B	243	521	185	110
	C	122	591	115	231
	Class	Precision	Recall	F-score	Accuracy
	A	0.58	0.65	0.61	0.73
	B	0.57	0.69	0.62	0.72
	C	0.51	0.35	0.41	0.67
	Overall Accuracy	0.56	Mean Precision	0.57	

Table D.2: Confusion matrix of Test data classified using 32-word Seed Lexicon

		PREDICTED			
	Class	A	B	C	Total
ACTUAL	A	207	85	61	353
	B	41	251	61	353
	C	107	128	118	353
	Total	355	464	240	
	Class	TP	TN	FP	FN
	A	207	558	148	146
	B	251	493	213	102
	C	118	584	122	235
	Class	Precision	Recall	F-score	Accuracy
	A	0.58	0.59	0.58	0.72
	B	0.54	0.71	0.61	0.7
	C	0.49	0.33	0.4	0.66
	Overall Accuracy	0.54	Mean Precision	0.56	

Table D.3: Confusion matrix of Test data classified using 2-word Seed Lexicon

		PREDICTED			
	Class	A	B	C	Total
ACTUAL	A	200	87	66	353
	B	71	231	51	353
	C	114	126	113	353
	Total	385	444	230	
	Class	TP	TN	FP	FN
	A	200	521	185	153
	B	231	493	213	122
	C	113	589	117	240
	Class	Precision	Recall	F-score	Accuracy
	A	0.52	0.57	0.54	0.68
	B	0.52	0.65	0.58	0.68
	C	0.49	0.32	0.39	0.66
	Overall Accuracy	0.51	Mean Precision	0.52	

D.2 Top 30

Table D.4: Top 30 ngrams generated with Emoticon Seed Lexicon

Unigrams		Bigrams		Trigrams	
Positive	Negative	Positive	Negative	Positive	Negative
लाभको	हाहाकार	मेरो.तर्फबाट	जनता.मार्ने	प्रत्यक्ष.निर्वाचित.कार्यकारी	अरु.केही.होइन
दामि	लाउडा	जित्नु.पर्छ	भो.भनेर	दिने.एक.मात्र	त.हो.नि
उत्तरोत्तर	आउथ्यो	छ.हजुरलाई	देश.बेच्ने	नेकपा.एमाले.जिन्दाबाद	राजा.महेन्द्र.ले
टिमलाई	येत्रो	एउटा.प्रश्न	लाज.लाग्नु	बि.बि.सी	त.ठिकै.हो
हिममत	लगायो	गर्नु.हुने	गु.खाने	के.पी.ओली	लाई.के.थाहा
अभिवादन	खान्छन	थापा.लाई	लाज.सरम	साझा.सवाल.कार्यक्रम	के.हेरेर.बसेको
टीम	औसधि	अर्पण.गर्दछु	अमेरिका.र	धैर्य.धारण.गर्ने	पनि.कानै.चिरेको
अगृम	लेनको	नारायण.जी	सुरु.भयो	बधाई.तथा.सफल	देशका.नदि.नाला
सुखलाई	तिमिहरुलाइ	मन.पराएको	हरु.त	र.जनताको.लागी	हिम्मत.छ.भने
पेजलाई	ज्याला	बि.बि	रै.छ	रेखा.मैसाप.जी	भनेको.यही.हो
क्षमतावान	कौडिको	लागि.हार्दिक	सबै.थोक	तथा.सफल.कार्यकालको	कानमा.तेल.हालेर
सून्दर	ग्यासको	सी.नेपाली	कानै.चिरेको	चिर.शान्तिको.कामना	के.नै.गर्न
शिघ्र	ढुकुटि	छोटो.समयमै	को.गुलामी	बधाई.छ.कमरेड	रगत.र.पसिना
उहाँहरु	गुजारा	लागि.विशेष	छानबिन.गर्ने	धेरै.बधाई.छ	आए.पनि.कानै
हार्दीक	सिदै	नेपाली.साहित्यलाई	भारतलाई.बेचेर	बी.बी.सी	आखामा.छारो.हालेर
सूपार	पीडितका	नेपाली.साहित्य	घुस.खाने	बधाई.तथा.शुभकामना	के.हुन.सक्छ
शान्तीको	बुडी	धेरै.शुभकामना	साला.चोर	धेरै.धेरै.बधैछ	जोगी.आए.पनि
श्रद्धान्जली	बिबरण	चित्र.बहादुर	हिम्मत.छ	यदि.छ.भने	को.नाम.मा
भावपुर्ण	स्मार्ट	निर्वाचित.कार्यकारी	मर्ने.बेला	हार्दिक.बधाई.तथा	नेता.हरु.ले
भावपूर्ण	दरबारमा	दिने.एक	तिरेको.कर	एक.पटक.यो	भनेको.यहि.हो
सफल्ता	रित्तो	सुभकामना.छ	यो.नेता	धेरै.धेरै.शुभकामना	सबैलाई.चेतना.भया
अम्बर	टाउकोमा	धेरै.जनाले	गए.हुन्छ	आत्माको.चिर.शान्तिको	ने.क.पा
लक	लोकल	नमस्ते.जी	पीडितका.लागि	बिकास.र.समृद्धिको	का.नेता.हरु
सवालको	भत्काएर	बिबिसी.नेपाली	जनता.माथि	बाबुराम.भट्टराई.ले	को.हो.र
दिवंगत	खल्ती	लाख.शुभकामना	रगत.को	जय.जय.जय	यस्तै.हो.भने
रोल	िकन	यदि.तपाईले	आखामा.छारो	को.हार्दिक.मंगलमय	के.गर्ने.त
पराएको	कार्बाहि	धेरै.खुशी	के.गर्छौ	यदि.तपाई.लाई	हाम्रो.देश.नेपालमा
गदर्छु	गर्छस	आत्माको.चिर	छ.बा	माया.गर्ने.र	के.यहि.हो
दशमीको	थोक	मृत.आत्माको	साला.कुकुर	मन.छुने.नेपाली	क.पा.एमाले
बदाई	बेचि	लाभको.कामना	का.दलाल	शुभकामना.व्यक्त.गर्दछु	त.होला.नि

Table D.5: Top 30 ngrams generated with 32-word Lexicon

Unigrams		Bigrams		Trigrams	
Positive	Negative	Positive	Negative	Positive	Negative
सामुदायिक	प्रभुलाई	व्यक्त.गर्दछु	सबै.थोक	यो.पेज.लाईक	राजा.महेन्द्र.ले
परिवारजनमा	भारतकै	धेरै.शुभकामना	जानेको.छ	हार्दिक.बधाई.तथा	नेपाल.को.कानुन
लक	धोतिको	जन्म.दिनको	मधेसी.लाई	धैर्य.धारण.गर्ने	रगत.र.पसिना
उत्तरोत्तर	गन्डकी	गर्नु.भएकोमा	मधेश.मा	गरे.कसो.होला	के.हुन.सक्छ
रामप्यारी	भारुको	नायिका.रेखा	पनि.भारत	एक.पटक.यो	लागि.जे.पनि
मैसाप	अंगिकृत	को.सुभकामना	भन्न.बेर	हुन्छ.भने.एक	लाई.के.थाहा
दाईको	पड्काउने	राम्रो.गर्नु	भारतलाई.बेचेर	बधाई.तथा.सफल	कानमा.तेल.हालेर
हमाल	खोजेको	प्रदान.गरुन	जनता.भेडा	धेरै.राम्रो.काम	सबैलाई.चेतना.भया
बदाई	तँलाई	नेपाली.क्रिकेट	बनाउने.होईन	धेरै.धेरै.शुभकामना	नै.छ.र
श्रद्धान्जली	भाउमा	हार्दिक.स्वागत	र.एमालेले	हिजोको.साझा.सवाल	आखामा.छारो.हालेर
मिलोस्	मोदिको	प्रगतिको.कामना	धोती.को	जन्म.दिन.को	नेपाल.को.नेता
अम्बर	भुस्याहा	दशमी.को	गु.खाने	स्वास्थ्य.लाभको.कामना	प्रचण्ड.र.बाबुराम
रुद्रप्रिया	कालापानी	यदि.छ	मा.भारतीय	पेज.लाईक.गरि	त.हो.नि
दामि	बिखण्डनकारी	श्रद्धाञ्जली.अर्पण	तातो.न	चिर.शान्तिको.कामना	ने.क.पा
सुखद	टिकापुर	सफल.होस	मरे.पनि	हुने.थियो.कि	त.के.कुरा
जिवनमा	डराउने	मन.छुने	भ्रष्ट.र	बधाई.तथा.शुभकामना	का.नेता.हरु
हिममत	कौडिको	आत्माको.चिर	राजा.महेन्द्र	दिने.एक.मात्र	के.पी.ओली
साहित्यिक	चिनिया	को.ल	साला.चोर	शुभकामना.व्यक्त.गर्दछु	जे.पनि.गर्न
रुद्र	जरै	मन.मुटुमा	न.छारो	रेखा.जी.को	के.का.लागि
लाभको	केरा	लागि.विशेष	का.दलाल	राम्रो.काम.हो	पर्.यो.नि
वधाई	घुस्याहा	मलाई.मन	आन्दोलन.गर्न	राम्रो.हुन्थ्यो.कि	के.हेरेर.बसेको
सून्दर	लेन्डुप	हजुरलाई.पनि	एकातिर.कुम्लो	भए.राम्रो.हुन्थ्यो	भनेको.यहि.हो
गर्दछु	मद्याएर	नेपाली.चलचित्र	मन्त्री.पद	रेखा.मैसाप.जी	क.पा.एमाले
बिनम्र	रित्तो	माथी.उठेर	रगत.र	प्रत्यक्ष.निर्वाचित.कार्यकारी	को.हात.मा
मुटुलाई	निकाले	पृथ्वी.नारायण	साला.कुकुर	यो.साझा.सवाल	त.होला.नि
पराएको	असुल	शान्तिको.कामना	भ्रष्टाचार.गरेर	बि.बि.सी	केही.फरक.पर्दैन
सुन्दरता	थुईक्क	सफलताको.कामना	र.पसिना	हार्दिक.बधाई.छ	बाहेक.अरु.केही
सूपार	गर्छस	शुभ.यात्रा	सबैलाई.चेतना	धारण.गर्ने.शक्ति	नेपाल.का.नेता
दाइको	आतंकवादी	जय.माता	रगत.को	हार्दिक.मंगलमय.शुभकामना	कांग्रेस.र.एमाले
बनौँ	चुरोट	पनि.शुभकामना	अनि.नेता	यदि.छ.भने	अरु.के.नै

Table D.6: Top 30 ngrams generated with 2-word Lexicon

Unigrams		Bigrams		Trigrams	
Positive	Negative	Positive	Negative	Positive	Negative
सुन्दैछु	बिस्तारबादी	थापा.ले	यिनै.हुन	रेखा.मैसाप.जी	चिर.शान्तिको.कामना
उत्तरोत्तर	लिनुपर्छ	मदन.दाई	लाईक.गरि	मेरो.बिचार.मा	के.नै.गर्न
थिको	समयसम्म	लाख.शुभकामना	जनताको.मत	धेरै.राम्रो.छ	यदि.तपाई.लाई
प्रस्तुती	संम्बिधान	यो.फोटो	मा.भारतीय	लाई.धेरै.धेरै	राजा.महेन्द्र.ले
हिरोइन	हजारको	लागि.हार्दिक	दलाल.र	लाई.बधाई.छ	त.के.कुरा
मालाले	जातियताको	धन्यबाद.नेपाल	सम्झनु.होला	बी.बी.सी	आत्माको.चिर.शान्तिको
बाडने	ढाल्ने	बि.सी	हरु.कै	बि.बि.सी	पेज.लाईक.गरि
हमाल	आतंकवादी	छ.हजुरलाई	राष्ट्रपति.र	को.साझा.सवाल	भनेको.यही.हो
सूपार	काठमाडौमा	श्री.आदरणीय	शोक.सन्तप्त	तथा.सफल.कार्यकालको	हुन्छ.भने.एक
हिममत	लाद्र	धेरै.मन	चुनाव.जित्ने	धेरै.राम्रो.लाग्यो	कुनै.औचित्य.छैन
लक	सम्बीधान	राम्रो.बिचार	सम्भावना.छ	सफल.कार्यकालको.शुभकामना	प्रत्यक्ष.निर्वाचित.कार्यकारी
गोबिन्दे	प्रावधान	नेपाली.चलचित्र	अधिकार.सम्पन्न	रेखा.थापा.को	पनि.कानै.चिरेको
दामी	मोर्चाको	धेरै.शुभकामना	भिन्नै.तरिकाले	हिजोको.साझा.सवाल	धारण.गर्ने.शक्ति
सून्दर	चिनिया	को.ल	को.आन्दोलन	रेखा.जी.को	भने.एक.पटक
खिचेर	राजसंस्था	अग्रिम.बधाई	आदिवासी.जनजाति	भए.राम्रो.हुन्थ्यो	कानमा.तेल.हालेर
रमैलो	कार्यकर्ताहरु	धेरै.खुशी	सभाको.चुनाव	नेपाल.प्रहरी.को	आए.पनि.कानै
लभ	घरलाई	राम्रो.लाग्छ	मृत्यु.दण्ड	राम्रो.काम.गर्नु	प्रचण्ड.र.बाबुराम
हिर्दय	डरलाग्दो	समाचार.सुन्न	हक.र	राम्रो.काम.हो	धैर्य.धारण.गर्ने
बधैछ	भनिने	यो.साझा	आफ्नो.पहिचान	यो.साझा.सवाल	रगत.र.पसिना
मैसाप	कौडिको	धेरै.धन्यबाद	सङ्ग.आबद्ध	राम्रो.हुन्थ्यो.कि	यदि.छ.भने
नगरेकै	करीब	सवाल.ले	माथिल्लो.कर्णाली	धेरै.धेरै.बधैछ	जोगी.आए.पनि
रेख	संबिधानको	राम्रो.लाई	मा.करोड	धेरै.बधाई.छ	लाईक.गरि.भ्रस्टाचार
रि	अंगिकृत	कतार.मा	देश.बेच्ने	धेरै.राम्रो.काम	बाहेक.अरु.केही
दामि	पाता	मैसाप.जी	र.मधेशी	साझा.सवाल.को	लागि.जे.पनि
स्टार	आवश्यक	सवाल.कार्यक्रम	पनि.कानै	को.साथ.मा	एक.पटक.यो
बदाई	सिटामोल	रुद्रप्रिया.फिल्म	स्थानीय.निकायको	बधाई.तथा.सफल	माया.गर्ने.र
दशमी	टुक्राउन	धुर्मुस.सुन्तली	यदि.तपाईले	नेपाल.प्रहरी.ले	दिने.एक.मात्र
मंगलमय	फास्ट	बी.बी	स्थापित.गर्न	नै.धर्म.हो	मन.पर्छ.भने
प्यारी	रटान	धेरै.बधैछ	जन्म.कैद	धेरै.धेरै.शुभकामना	मन.मुटुमा.बस्न
फिलिम	बिखण्डन	बी.सी	राजा.र	धेरै.धेरै.धन्यबाद	मन.छुने.नेपाली

D.3 Word Clouds

D.3.1 Unigrams

Figure D.1: Word Cloud of Negative Unigrams (2-word Primary Lexicon)

Figure D.2: Word Cloud of Positive Unigrams (2-word Primary Lexicon)

Figure D.3: Word Cloud of Negative Uni-grams (32-word Primary Lexicon)

Figure D.4: Word Cloud of Positive Uni-grams (32-word Primary Lexicon)

Figure D.5: Word Cloud of Negative Uni-grams (Emoticon Primary Lexicon)

Figure D.6: Word Cloud of Positive Uni-grams (Emoticon Primary Lexicon)

D.3.2 Bigrams

Figure D.7: Word Cloud of Negative Bigrams (2-word Primary Lexicon)

Figure D.8: Word Cloud of Positive Bigrams (2-word Primary Lexicon)

Figure D.9: Word Cloud of Negative Bi-grams (32-word Primary Lexicon)

Figure D.10: Word Cloud of Positive Bi-grams (32-word Primary Lexicon)

Figure D.11: Word Cloud of Negative Bi-grams (Emoticon Primary Lexicon)

Figure D.12: Word Cloud of Positive Bi-grams (Emoticon Primary Lexicon)

D.3.3 Trigrams

Figure D.13: Word Cloud of Negative Tri-grams (2-word Primary Lexicon)

Figure D.14: Word Cloud of Positive Tri-grams (2-word Primary Lexicon)

Figure D.15: Word Cloud of Negative Tri-grams (32-word Primary Lexicon)

Figure D.16: Word Cloud of Positive Tri-grams (32-word Primary Lexicon)

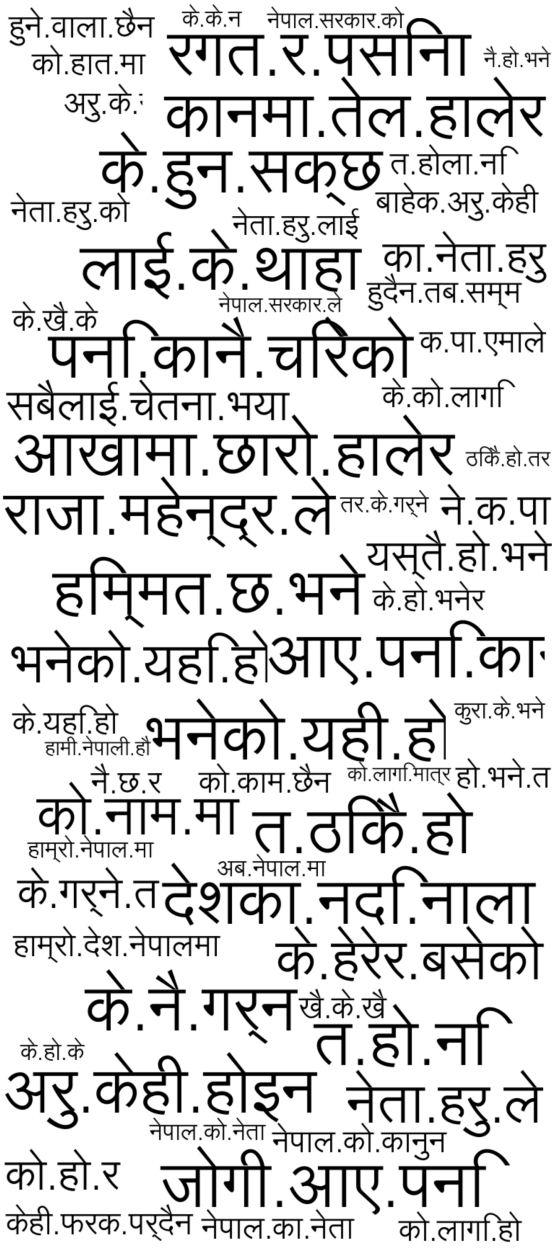

Figure D.17: Word Cloud of Negative Tri-grams (Emoticon Primary Lexicon)

Figure D.18: Word Cloud of Positive Tri-grams (Emoticon Primary Lexicon)