Generating Opinion Lexicon for Nepali using Microblogs

Sawan Vaidya, *Student, NCIT,*

***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.**

***Index Terms*—Opinion Lexicon, Sentiment Analysis, Microblog, Facebook, Classification, Nepali, Nepalese, Emoticon**

I. Introduction

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.

A. 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.

S. Vaidya (sawanvaidya@gmail.com) was a student at Nepal College of Information Technology enrolled in MS in Computer Science

Asst. Prof. Bal Krishna Bal of Kathmandu University was the thesis supervisor

II. Literature Review

There are two mainly used methods to generate Opinion Lexicons[7]. 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.

A. 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[8].

One of the most popular resource used for sentiment analysis in English is SentiWordNet, currently in version 3.0 [1][4]. 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.

[5] 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[5].

B. 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 [7].

One of the most used metrics in corpus based approaches is pointwise mutual information (PMI). [10] 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.

[11] 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.

Another popular method is based on adjectives connected by conjunctions such as 'and' and 'but'. In [6] 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.

C. Study Area

As already mentioned in the Introduction section, resources for Nepali language are rare. [5] 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 [3]. A corpus based approach has to be used.

The approach taken in this thesis will attempt to use a method similar to [11]. 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. [9] 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.

III. RESEARCH METHODOLOGY

This section will describe the features of the OL that will be built during this research. In this research two manually built and one generated OLs will be experimented with. The manually built ones are the 2-word (Table VI) and the 32-word OL (Table VII). The emoticon lexicon (Table VIII) is obtained from a research on sentiment of emoticons [9]. 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 [7]. The probability of a sentence being subjective based on the presence of at least one adjective was found to be 55.8% [2]. 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 interpreted as 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 [9].

A. The Core Insight

Building a 2-word lexicon such as the one in Table VI can be done easily. But to build a larger lexicon requires many hours of manual work. 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.

The intuition behind the OL 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 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.

IV. SYSTEM MODEL AND WORK FLOW

In Figure 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

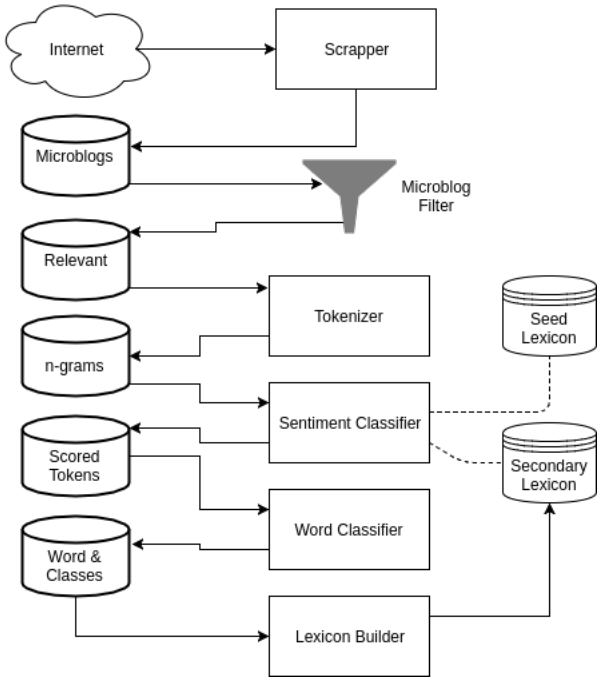

Fig. 1. Block Diagram of the proposed solution

starting from Sentiment Classifier and below repeats for a specified number of iterations.

A. 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.

B. 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.

C. Tokenizer

The tokenizer's responsibility is to break down sentences in microblogs 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. 1 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:

1) Word Count and TF-IDF:

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

$$IDF = log_{10}(\frac{N}{D}) \tag{1b}$$

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 1 is used to compute a Term Frequency - Inverse Document Frequency of each word. These statistics are used in the next step

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.

D. Sentiment Classifier

The sentiment classifier uses the tokens from the previous step to classify the token collections based on the ngrams of known sentiments. Those token collections that do not have any linkages with the supplied lexicons 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 VI
- A devanagari lexicon was manually developed and contained 16 positive and 16 negative words (Table VII)
- Emoticon Lexicon was obtained from [9]. It contained emoticons and their corresponding sentiment scores (Table VIII)
- 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)

E. 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.

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.

F. 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.

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:

1) Skewness Filter: Pearson's median skewness (Eqn. 2) 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. 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 minimum threshold skewness.

$$skewness = \frac{3 * (mean - median)}{standarddeviation} \tag{2a}$$

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). 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

$$range = highestscore - lowestscore \tag{3a}$$

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.

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

V. EXPERIMENTS

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

- **Experiment 1** : Emoticon Lexicon (Fig. VIII)
- **Experiment 2** : Manually build 32 word lexicon (Table VII)
- **Experiment 3** : 2 word lexicon (Table VI)

TABLE I
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 II*

TABLE II
BUILD FILTER PARAMETERS

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

VI. 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.

A. Test Data 1 - Microblog Subset

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.

TABLE III
SAMPLE TEST DATA (MICROBLOG SUBSET)

Text	Sentiment
म मंगल ग्रह बाट सुन्दै छु	0
प्रधानमंत्री को भारत यात्रा ले नेपा 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

- Total Unique 1 grams: 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

TABLE IV
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

- Total Unique 1 grams: 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

B. Test Data 2 - Word Meanings

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 IV

C. 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.

$$Precision = \frac{TP}{TP + FP} \tag{4a}$$

$$Recall = \frac{TP}{TP + FN} \tag{4b}$$

$$Fscore = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4c}$$

$$Accuracy = \frac{TP + FN}{TP + FP + TN + FN} \tag{4d}$$

where,

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

Formula 4 are used for the calculations for each class, whereas, formula 5 are used for overall calculations.

$$OverallAccuracy = \frac{Sumofdiagonals}{OverallSum} \tag{5a}$$

$$MeanPrecision = \sqrt{Precision(A) * Precision(B)} \tag{5b}$$

Formula 5a 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 5b 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(5a) and Mean Precision (5b).

VII. RESULTS AND DISCUSSION

Top 30 positive and negative words obtained after the experiment with emoticon lexicon seed is listed in IX. In the untruncated version, Emoticon approach yielded 4004 total ngrams; 2-word approach yielded 4237 ngrams; and 32-word approach yielded 3924 ngrams.

A. Mistakes

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.

Some positive terms such as उहाँहरु, गद्छु, छ.हजुरलाई, गर्नु.हुने, यदि.तपाईं.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.

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 गद्छु may have been reported as positive because it occurs frequently within ngrams such as शुभकामना.व्यक्त.गर्दछु or श्रद्धान्जली.व्यक्त.गर्दछु.

B. Achievements

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, गद्छु 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.

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

C. Test Results

The following two sections compare the accuracy and precision of results obtained from 3 primary/seed lexicons.

1) Test Data 1: Figure 2 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 4 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.

2) Test Data 2: Figure 3 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 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.

VIII. 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 V 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.

A. 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.

B. 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

BELOW: ACCURACY AND PRECISION OR LEXICON OBTAINED FROM 3 SEED LEXICONS

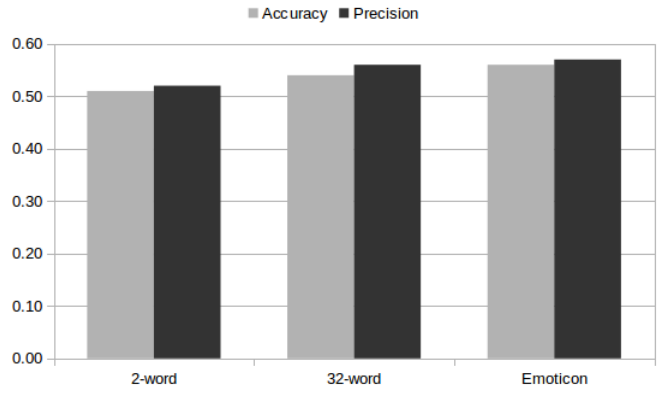

Fig. 2. Tested on Test Data 1

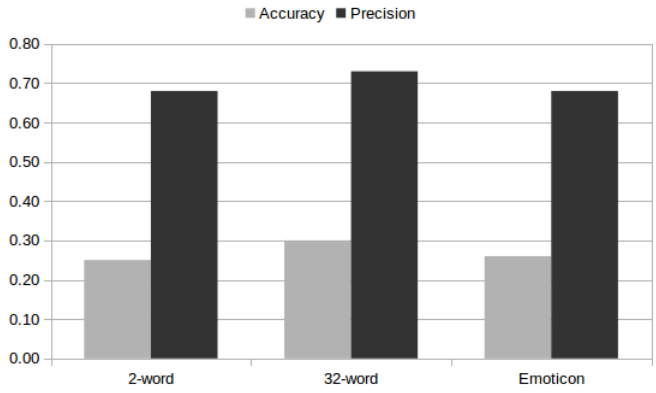

Fig. 3. Tested on Test Data 2

BELOW: PERCENT OF POSITIVE, NEGATIVE AND UNKNOWN CLASSIFICATIONS FROM 3 SEED LEXICONS

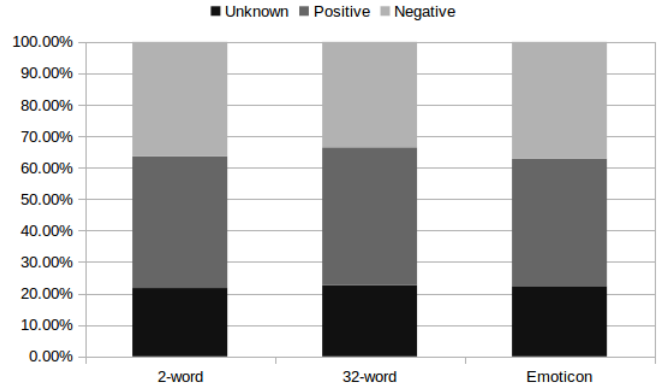

Fig. 4. Using Test Data 1

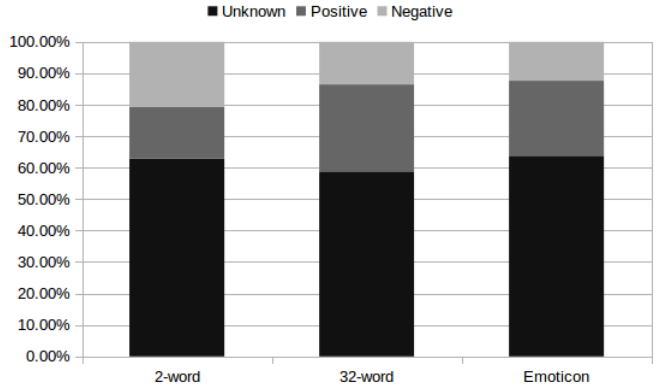

Fig. 5. Using Test Data 2

TABLE V
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	बाझ	-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
सुस्	-0.91	झुर	-0.15	शान्ति	0.91	विश्वसनीय	1

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 [11]) would be more useful that using a purely statistical approach.

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.

REFERENCES

[1] 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.

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

[3] 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.

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

[5] 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.

[6] 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.

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

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

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

[10] 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.

[11] 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.

APPENDIX
SEED LEXICONS

TABLE VI
SIMPLE 2-WORD SEED LEXICON

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

TABLE VII
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	हेचुवा	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

TABLE VIII
COMPLETE EMOTICON SEED LEXICON (SOURCE: EMOJI SENTIMENT RANKINGS V1)

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		0.32	0.31		0.15	0.68
	0.47	0.05		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			

APPENDIX
RESULT SAMPLE
TABLE IX
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.तपाईंले	आखामा.छारो	को.हार्दिक.मंगलमय	के.गर्ने.त
पराएको	कार्बाहि	धेरै.खुशी	के.गर्छौं	यदि.तपाईं.लाई	हाम्रो.देश.नेपालमा
गद्छु	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