Automatic Acquisition of Opinion Words from Myanmar Facebook Movie Comments

Win Win Thant, Kiyoaki Shirai
University of Information Technology (UIT), Japan Advanced Institute of Science and Technology (JAIST)
winwinthant@gmail.com, kshirai@jaist.ac.jp

Abstract

This paper presents a method for the automatic acquisition of new Myanmar opinion words based on syllable n-gram frequencies, since the Myanmar language uses a syllabic writing system. There is a need for the automatic construction of opinion lexicons for any specific domain, and the proposed method makes it possible to construct an opinion lexicon for Myanmar movies using a bootstrapping approach. We use Myanmar movie comments from Facebook, a popular social network, as a corpus, and determine the valence or polarity of the comments (positive or negative) using a small number of opinion words. These opinion words are then extracted from polarity-identified Facebook comments. Our proposed method is based on n-grams of syllables without word segmentation, since Myanmar is an under-resourced language and no public tool exists for word segmentation.

Keywords: Opinion lexicon, Myanmar movie comments, n-gram syllables, bootstrapping

1. Introduction

Textual information can be divided into two main domains: facts and opinions. While facts focus on the transmission of objective data, opinions express the sentiments of their authors. Opinions are also subjective expressions that describe people's emotions, appraisals or feelings toward entities, events and their properties. The quantity of users' comments is increasing daily, since most people enjoy giving their opinions on the Web. The concept of an opinion is very broad, and yet is so important that whenever we need to make a decision we ask for others' opinions. This is not only true for individuals but also for organizations.

All of us tend to infer the meaning of opinion leaders in our way. For some, opinion leaders are great people such as Nelson Mandela, Mother Teresa and Mahatma Gandhi, who paved the way for revolutions and completely changed the faces of entire countries with their opinions and successful actions. In today's changing world, social media platforms have taken on the role of opinion leaders. Advertisers use social media-based and celebrity opinion leaders to carry and "trickle down" their message through social media such as Facebook or Twitter, to influence their target groups.[1]

In Myanmar, most people use social media, and especially Facebook, to express their opinions about specific topics in the Myanmar language. Moreover, most of the popular Myanmar film and movie production companies have set up social media accounts to allow users to express their personal opinions about movies. Customer comments are valuable, and are an important source of data for multiple purposes. The reason for using Facebook movie comments in this study is that they provide good material for analyzing the subjectivity and opinions of users. Due to a lack of domain-dependent opinion lexicons, we propose a method for the automatic extraction of opinion words within the domain of movies. To our knowledge, no prior work has been done exclusively on the automatic acquisition of new opinion words using Facebook movie comments in Myanmar.

[1] https://www.facebook.com/mccollinsmedia/

The remainder of this paper is organized as follows. Section 2 presents related work, focusing on methods of lexicon creation, n-gram statistics and Myanmar syllable segmentation. Section 3 explains the proposed system in detail. The experimental results are described and analyzed in Section 4. We conclude with a summary of the proposed technique and future work in Section 5.

2. Related Work

Our work is inspired by a word-level classification model for automatically generating a Twitter-specific opinion lexicon from a corpus of unlabeled tweets (Bravo-Marquez et al., 2015a). These authors proposed a distributional representation for words by treating them as the centroids of the tweet vectors in which they appear. Bravo-Marquez et al. (2015b) extended a supervised method for expanding an opinion lexicon in the context of emoticon-annotated tweets, by creating a lexicon with disambiguated POS entries and a probability distribution for positive, negative, and neutral classes. In a subsequent paper, these authors proposed a methodology for expanding the NRC word-emotion association lexicon based on a collection of unlabeled tweets (Bravo-Marquez et al., 2016). They compared different word-level features extracted from unlabeled tweets such as unigrams, Brown clusters, POS tags, and word2vec embedding. In contrast to their previous work, this methodology enabled the identification of emotional words from any domain-specific collection using unlabeled tweets.

Our approach is similar to a method proposed by Nagao et al. (1994), who developed a new method and software for n-gram frequency calculation for values of n up to 255; they also automatically calculated n-grams for several large texts in Japanese, including between two and thirty million characters and derived words, compound words and collocations. Here, we use a collection of Facebook comments as a corpus and Myanmar syllables as input for n-grams. Many studies of lexicon creation have been carried out in other languages such as Portuguese and German. Souza et al. (2011) proposed the integration of different linguistic resources to identify opinion-bearing terms, and to create a single opinion lexicon for the Portuguese language. Remus et al. (2010) described the

structure of a publicly available German-language resource for sentiment analysis called SentiWS, three sources including the General Inquirer (GI) lexicon, a co-occurrence analysis of rated product reviews and the German collocation dictionary that was utilized to assemble this, and a semi-supervised method used to weight the strength of the entries.

A final line of related work concerns syllable segmentation. In addition to the methods described in Section 1, which use syllabic input, some methods of syllable segmentation have been read for syllabic input. Zin and Mikami (2008) proposed a rule-based approach to a syllable segmentation algorithm for Myanmar text. They created the segmentation rules based on the characteristics of Myanmar syllable structure, but did not consider the non-Myanmar characters within the script in their approach. Hla and Kavi (2008) described the need and possible techniques for segmentation of Myanmar script. They used a combination of stored lists, suffix removal, morphological analysis and syllable-level n-grams to hypothesize valid words with an accuracy of about 99%. They built a list of 1216 stop words, 4550 syllables and 800,000 words from a variety of sources, including their own corpora. Tin and Mikami (2010) proposed the automatic syllable segmentation of Myanmar text using a finite state transducer, without using step-by-step heuristic rules and an annotated corpus. They proved that this approach could handle both the regular and irregular syllable structures of Myanmar with acceptable performance. Although they did not publish this syllable segmentation software, their segmentation of the Myanmar syllable was an important step.

3. Proposed Method

The overall process of the proposed method is shown in Figure 1.

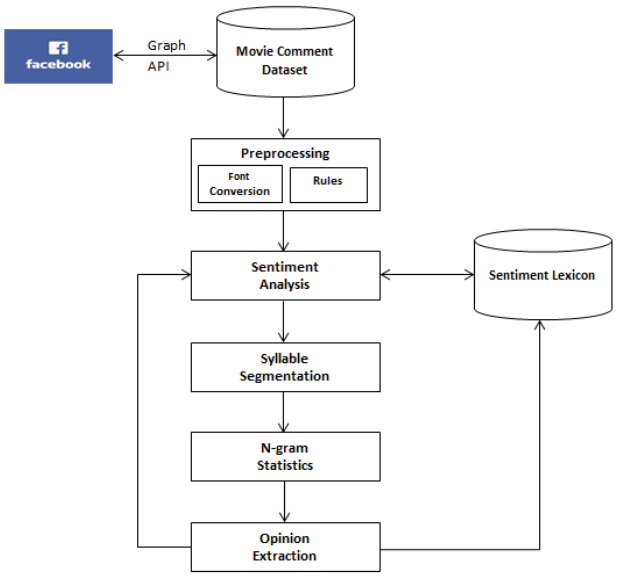

Figure 1: The proposed system

Firstly, an initial sentiment lexicon containing a small number of opinion words is manually created. Facebook comments concerning movies are retrieved and preprocessed. Next, each comment is classified as positive or negative using the initial lexicon. Then, an arbitrary n-gram of syllables is extracted using a publicly available syllable segmentation tool as candidates of the opinion words. Next, the opinion words in the movie domain are chosen from these candidates, based on the statistics of the syllable n-grams, and added to the sentiment lexicon. The procedures for the sentiment analysis of the comments and acquisition of the opinion words are repeated until no opinion word is obtained.

3.1 Data

Before applying the preprocessing step, it is common practice to collect data. Accurate data collection is important, as it helps to ensure the integrity of research. If existing data are not accurate, or do not provide enough information or the right kind of information, the study cannot be validated, misleading other researchers into pursuing fruitless avenues of investigation and distorting findings, resulting in wasted resources.

3.1.1 Collection of Facebook Movie Comments

In the last few years, Facebook has taken the world by storm and has become an important element in the field of communications. The Facebook Myanmar movie comment dataset is a corpus of movie comments used for the automatic acquisition of new opinion words. The original unprocessed Facebook comments used in this research were collected using the Graph API,[2] a programming tool designed to support greater access to conventions on the Facebook social media platform. Statistics for these data are shown in Table 1. In this table, the numbers of positive and negative comments are counted manually; note that the manually annotated polarity of the comment is not used for construction of the sentiment lexicon in the proposed method.

Type	Number
Facebook Movie Web Page Links	14
Posts	394
Comments	12,123
Syllables	92,167
Positive Comments	5,697
Negative Comments	458

Table 1: Statistics of data

3.1.2 Json to csv converter

All comments from movie posts are retrieved and the Json file of extracted comments is converted into a csv file using a converter[3] and then transformed to txt format, turning the separate movie files into a single file.

3.2 Preprocessing

Data preprocessing allows the production of a higher quality of text classification and a reduction in the computational complexity. Our preprocessing procedure includes the following steps.

3.2.1 Font Conversion

There is an issue with fonts in the Myanmar language; most users are very familiar with the Zawgyi font, and use this in Facebook comments, while most applications in the

[2]https://developers.facebook.com/tools-and-support/
[3]https://json-csv.com/

technology field accept only Unicode. Thus, in this research, the collected comments in Zawgyi are converted to Unicode using an online converter.[4]

3.2.2 Rules for Preprocessing

This is the conversion step from unstructured to structured data. Before starting the creation of n-grams, it is necessary to carry out several preprocessing and cleaning steps. One difficulty in the processing of most comments and documents is the presence of certain kinds of textual errors, such as spelling and grammatical errors. An analysis of data that have not been carefully screened can produce misleading results. We therefore investigate the influence of preprocessing rules on the quality of the system. To do this, the system performs the following preprocessing steps:

- Emoticons are replaced with their Myanmar words.
  For example: :) = ပြုံးသည် (smile) :(= မဲ့သည် (sad)
- Several useful English words and loan words are first converted to lower case and then translated to similar Myanmar words, for example: 'good' to ကောင်းသည် , 'like' to ကြိုက်သည် , 'academy' to အကယ်ဒမီ
- English acronyms are translated to similar Myanmar words. For example: 'LOL' to အရမ်းရယ်ရသည်, 'WOW' to အံ့အားသင့်သည်
- English-Myanmar combined words are translated to meaningful Myanmar words. For example: 'Sပေါ' , '$ပေါ' to အရမ်းပေါ (so bad)
- Myanmar words with English pronunciation are translated to Myanmar words, for example: ဂွတ်တယ် (good in English) to ကောင်းတယ်
- Certain Myanmar synonym adverbs like အားကြီး, အကုန်, အသေ are replaced by အရမ်း (very).
- Repeated Myanmar syllables are replaced by three syllables. For example: ခုန်းခုန်းခုန်း (encourage) , ခစ်ခစ်ခစ် (laugh)
- Punctuations and other non-Myanmar words in the comments are deleted, and white space is reduced.
- Words in comparative form are replaced with the basic form. Examples of such words are ပိုတော် (cleverer), အတော်ဆုံး (cleverest)

3.2.3 Syllable Segmentation

This refers to the ability to identify the components of a word, phrase, or sentence. Since Myanmar is a tonal language and has a syllabic writing system, the fundamental building blocks of the language are syllables. Almost every syllable has a meaning in the Myanmar language, and this can also be seen in the work of Hopple (2003). We use publicly available syllable segmentation software[5] to segment the syllables in Facebook comments and apply these segmented syllables in the calculation of n-gram statistics.

In Myanmar, unlike in English, word segmentation is not clearly denoted in sentences. For this type of language, the sentences are usually divided into a sequence of words in preprocessing. However, since Myanmar is an under-resourced language, no public tool for word segmentation is available. Therefore, in this study, the arbitrary n-grams of the syllables are extracted as candidates for the opinion words. Some of them are not genuine words, and irrelevant syllable n-grams are automatically discarded. The acquisition of Myanmar opinion words is more challenging than in other, more richly resourced languages.

3.3 Initial Lexicon

A sentiment lexicon is a list of positive and negative opinion words. Positive opinion words are used to express desired states, while negative opinion words are used to express undesired states. Most of the opinion words are not of one syllable, since a phrase such as "ဒီရုပ်ရှင်က မကောင်းဘူး" (The movie is not good) can be wrongly classified as a positive sentence, due to the syllable ကောင်း (good). In the proposed system, the initial sentiment lexicon contains one positive word "ကြိုက်" (like) and one negative word "ပေါ" (bad). These are the most common words in the comments, and Figures 2 and 3 show the method used to find these initial positive and negative seeds within the comments.

Input: Facebook comments
Output: a positive seed
1. Read all Facebook movie comments.
2. Find unigram syllable count, combine duplicate syllables and sort them in descending order.
3. Select highest 1% of total unigram count and remove stopwords from these.
4. Reduce the unigram count by the bigram count and negation (for example, subtract the count of "not good" from the count of "good").
5. Sort unigram syllables again and manually select the highest frequency positive unigram syllable as an initial positive seed.

Figure 2: Algorithm for initial positive seed selection

Input: Facebook comments
Output: a negative seed
1. Read all Facebook movie comments.
2. Find unigram and bigram syllable counts, combine duplicate syllables and sort these in descending order.
3. Select highest 1% of total unigram count and total bigram count, and remove stopwords from these.
4. Manually select the highest frequency negative unigram or bigram syllable as an initial negative seed (this is often found in both unigram and bigram ("bad" and "not good") in the comments)

Figure 3: Algorithm for initial negative seed selection

3.4 Sentiment Analysis of Facebook Comments

The classification of a text or a sentence according to its semantic orientation or polarity (positive, negative or neutral) can be performed by several methods including

[4] http://www.mcfmyanmar.org/myanmar-unicode-converter/
[5] http://myanmarnlpteam.blogspot.jp/2008/02/syllable-segmentation-software.html

machine learning, lexicon-based methods or hybrid methods. We use a lexicon-based approach to discover sentiments. Prior polarities are defined by the initial lexicon, and the polarity value of a comment is a comparison of the prior polarities of its sentiment words. If the positive (greater than zero) opinion count is greater than the negative count, it is classified as positive; otherwise, it is classified as negative. The algorithm for polarity classification of the comments is shown in Figure 4. After sentiment analysis, we obtain a set of positive and negative sentences, and these are used in the next step.

```
read a comment in dataset
while there is a comment in dataset do
    pcount ←0
    ncount ←0
    read an opinion word in lexicon
    while there is an opinion word in lexicon do
        if the comment contains positive opinion word
        then        pcount ←pcount+1
        if the comment contains negative opinion word
        then        ncount←ncount+1
    end while
    if(pcount>ncount) then
        classify the comment as positive
    else if (ncount>pcount) then
        classify the comment as negative
end while
```

Figure 4: Algorithm for polarity classification of comments

3.5 Opinion Word Extraction

Following data collection and preprocessing, opinion words are extracted from a set of positive and negative sentences, based on the statistics of syllable n-grams.

3.5.1 N-grams

In the fields of computational linguistics, an n-gram is a contiguous sequence of n items from a given sequence of text or speech. The items can be phonemes, syllables, letters, words or base pairs, according to the application. The n-grams are typically collected from a text or speech corpus. Statistics for the combination of letters (n-grams) are very useful, as this offers substantial savings in terms of human effort, and the use of n-gram statistics in opinion acquisition is therefore very attractive. Statistics for high-frequency words tend to be more reliable than those for low-frequency ones. The syllable n-gram representation of the Myanmar word 'အကယ်ဒမီ (academy)' is shown in Table 2.

n=1	n=2	n=3	n=4
အ (a)	အကယ် (aca)	အကယ်ဒ (acade)	အကယ်ဒမီ (academy)
ကယ် (ca)	ကယ်ဒ (cade)	ကယ်ဒမီ (cademy)	
ဒ (de)	ဒမီ (demy)		
မီ (my)			

Table 2: Syllable n-gram representation of the Myanmar word 'အကယ်ဒမီ (academy)'

3.5.2 Generation of N-gram Frequency Statistics

The generation of n-grams starts with the corpus of preprocessed Facebook comments (sentences). From this corpus, all relevant n-grams (1-, 2-, 3-, 4- and 5-grams) are created, and each n-gram is associated with its frequency, i.e. the number of times it occurs in the corpus. It is very difficult to calculate n-grams for a large value of n, due to the memory limitations of computers. Most of the Facebook comments have an average of 14 syllables per sentence; 5-grams and below are enough for movie opinions, and these were therefore calculated for the corpus.

no	1-grams		2-grams		3-grams		4-grams		5-grams	
	word	count	word	count	word	count	word	count	word	count
1	တယ်	3846	အ ရမ်း	1086	ကြည့် ချင် နေ	349	အ ကယ် ဒ မီ	229	အ ရမ်း ကြိုက် နေ တယ်	136
2	အ	3488	ပါ တယ်	891	အား ပေး နေ	344	ကြည့် ချင် နေ တယ်	145	အ ကယ် ဒ မီ ရ	108
3	ပါ	2890	အား ပေး	859	အ ရမ်း ကြိုက်	306	ကယ် ဒ မီ ရ	140	အ ရမ်း အား ပေး နေ	77
4	မ	2326	ကြည့် ချင်	779	ကယ် ဒ မီ	277	အား ပေး နေ ပါ	113	ကယ် ဒ မီ ရ ပါ	75
5	ကို	2043	ကြိုက် တယ်	596	အ ကယ် ဒ	235	ပေး နေ ပါ တယ်	110	ဒ မီ ရ ပါ စေ	58
6	ပေး	1429	ပေး ပါ	472	တင် ပေး ပါ	234	အ ရမ်းကြိုက် တယ်	103	အ မြဲ အား ပေး နေ	39
7	တာ	1383	တင် ပေး	415	သ ရုပ် ဆောင်	232	အား ပေး နေ တယ်	102	အား ပေး ချင် နေ တာ	37
8	ကြိုက်	1339	ပေး နေ	378	ဟိန်း ဝေ ယံ	176	အ မြဲ အား ပေး	101	အ ရမ်း ကြည့် ချင်တယ်	29
9	ကြည့်	1306	ရမ်း ကြိုက်	369	ကြည့် ချင် တယ်	175	ကို မြင့် မြတ် ကို	93	အ ကယ် ဒ မီ ရ	28
10	က	1246	မိုး စက်	355	အ ရမ်း ကောင်း	165	အား ပေး ပါတယ်	86	အား ပေး နေ ပါ တယ်	27
11	ရမ်း	1242	မြင့် မြတ်	352	ဒ မီ ရ	148	ဒ မီ ရ ပါ	81	ကယ် ဒ မီ ရ ပါ	27
12	ရ	1161	အ ကယ်	315	မိုး စက် မေ	145	မီ ရ ပါစေ	77	ဒ မီ ရ ပါ စေ	25
13	ကြ	1146	ချင် တယ်	315	အ ရမ်း ချစ်	140	တင် ပေး ပါလား	75	အ မြဲ အား ပေး နေ	25
14	နေ	1097	ကယ် ဒ	276	နေ ပါ တယ်	134	သ ရုပ် ဆောင် တာ	73	အ ရမ်း ကြည့် ချင် လိုက်	24
15	ကား	1071	ဒ မီ	249	ရ ပါ စေ	127	မြဲ အား ပေး နေ	62	ကြည့် ချင် တာ အ ရမ်း	24

Table 3: Occurrence of the top 15 opinion n-grams from positive comments

3.5.3 Comparing and Ranking N-Gram Frequencies

The scores of n-grams are sorted in descending order and checked as to whether the highly ranked n-grams are valid opinion words. We generate at most 5-grams (n=5) and the highest ranking n-grams are mostly bi-grams (n=2); as we move towards longer n-grams, we generally obtain fewer valid words. It should be noted that these observations apply mostly to shorter comments with an average of about 14 syllables per sentence, such as those from Facebook movie webpage links. If these sentences were longer, a shift from 5-grams to higher n-grams would be seen. The ranking of the top 15 n-gram frequencies for positive comments is shown in Table 3, and n-gram statistics for Myanmar syllables are shown in Figure 5.

3.5.4 Score for Opinion Word Extraction

N-grams of syllables are candidates for opinion words. We calculate ScorePOS(w), the score of the positive orientation of each possible word w (syllable n-gram). First, the probability of generating w in the sets of positive sentences and all sentences, Ppos(w) and Pall(w), are calculated using Equations (1) and (2). Then, ScorePOS(w) is defined using Equation (3), where fre_pos(w) represents the count of w in positive opinion sentences. We assign a higher score to words that appear relatively frequently in positive sentences, and give a lower score to more general words. Frequent words are also preferred as positive opinion words to be added into the sentiment lexicon. The score of the negative orientation, ScoreNEG(w), can be calculated similarly. Finally, the most highly ranked candidates are extracted as positive or negative opinion words.

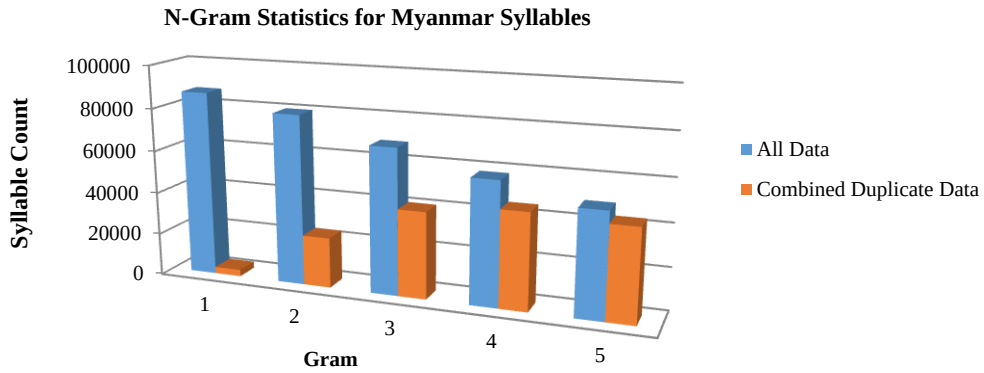

Figure 5: N-gram statistics for Myanmar syllables

$$Ppos(w) = \frac{count\ of\ w\ in\ positive\ opinionated\ sentences}{count\ of\ all\ n-grams\ in\ positive\ opinionated\ sentences} \quad (1)$$

$$Pall(w) = \frac{count\ of\ w\ in\ all\ sentences}{count\ of\ all\ n-grams\ in\ all\ sentences} \quad (2)$$

$$ScorePOS(w) = log\frac{Ppos\ (w)}{Pall\ (w)} * fre_{pos(w)} \quad (3)$$

3.5.5 Extraction of Opinion Words Using Bootstrapping Method

Starting from the initial lexicon with two words, syllable n-grams with scores greater than or equal to five are extracted as potential new opinion words[6]. In order to choose only meaningful words, the potential opinion word is added to the sentiment lexicon if it is a lemma or an inflectional form of a word in the Myanmar word list[7]. Unigrams are neglected to improve performance, since most of these words are meaningless. Furthermore, we do not add combinations of existing opinion words (for example, if an existing opinion word is “like”, its combinations of existing opinion words are “be like”, “like more”, “like most”, “like+stop word”, “other word+like”) and Myanmar stop words such as prepositions/postpositions, particles, inflections and conjunctions, which appear as suffixes of other words. When a word is assigned as both positive and negative, the difference of the polarity scores is considered. If its positive and negative scores are nearly equal, the word is not extracted as an opinion word; if the difference between ScorePOS(w) and ScoreNEG(w) is greater than three, the word is added as either a positive or negative word to the sentiment lexicon. The polarity of the added word is the same as the polarity of the higher score.

[6]The threshold of syllable n-grams is set to five by our intuition.

[7]https://github.com/kanyawtech/myanmar-karen-word-lists/blob/master/burmese-word-list.txt

4. Evaluation and Discussion

This section reports the results of the experiment conducted to evaluate our proposed method. After the first time with one opinion each for positive and negative, 11 and 10 new words were extracted as positive and negative opinion words, respectively. The program was run with 20 positive opinion words and 19 negative opinion words for the second time.

Our experiments showed that six and seven bootstrapping iterations were sufficient to extract 44 positive words and 35 negative words, respectively. These newly acquired words were evaluated in terms of precision. Precision is defined as the ratio between the number of correctly extracted opinion words and the total number of extracted words, as shown in Equation (4).

$$Precision = \frac{number\ of\ correctly\ extracted\ opinion\ words}{total\ number\ of\ extracted\ opinion\ words} \quad (4)$$

We manually checked the correctly extracted opinion words. The results of this evaluation of the lexicon obtained by bootstrapping are shown in Table 4. We found that the precisions of all positive and negative words were 86% and 89% respectively, which are relatively high.

Based on these bootstrap-based experiments, we found that most opinion words were syllable bi-grams and tri-grams, and their statistics act as a good scoring metric. Incorrect opinion words were retrieved; these were the names of actors that appeared many times in the comments, Myanmar adverbs and meaningful sub-words of other opinion words.

Round	Positive		Precision	Negative		Precision
	All extracted words	Correct words		All extracted words	Correct words	
1	11	8	73%	10	8	80%
2	31	26	84%	29	26	90%
3	41	35	85%	32	28	88%
4	43	37	86%	33	29	89%
5	44	38	86%	34	30	88%
6	-	-	-	35	31	89%
7				-	-	-

Table 4: Results of evaluation of the lexicon using bootstrapping

5. Conclusions and Future Work

In this paper, we propose a syllable-based n-gram approach for the automatic extraction of new opinion words from Facebook movie comments. These opinion words can be combined with existing opinion words to increase the accuracy of the opinion words and expand the lexicon. Opinion words are very important in sentiment analysis, and we believe that the proposed system can perform well in any domain to retrieve these words automatically.

A variety of steps can be taken to extend this work:

- We need to develop a larger corpus of movie documents and more data cleaning rules.
- We plan to handle spelling variations and to incorporate our method with a word segmentation tool.
- We intend to experiment further using a wider variety of Facebook comments, which we hope will give rise to more opinion words.
- We plan to perform statistical substring reduction (SSR) on the acquired n-gram statistics.
- We need to handle slang words.
- We need to handle implicit emotional comments, where the sentiment of an author is not literally written.
- We also intend to apply our approach to other domains.

6. Bibliographical References

Bravo-Marquez, F., Frank, E. and Pfahringer, B. (2015a). From unlabelled tweets to Twitter-specific opinion words. In Proceedings of the 38th International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 743-746.

Bravo-Marquez, F., Frank, E. and Pfahringer, B. (2015b). Positive, Negative, or Neutral: Learning an Expanded Opinion Lexicon from Emoticon-annotated Tweets. In Q. Yang & M. Wooldridge (Eds.), Proceedings of the 24th International Joint Conference on Artificial Intelligence, pp. 1229-1235), Buenos Aires, Argentina: AAAI Press.

Bravo-Marquez, F., Frank, E. and Pfahringer, B. (2016). Determining Word-Emotion Associations from Tweets by Multi-label Classification. IEEE/WIC/ACM International Conference on Web Intelligence (WI), Omaha, NE, 2016, pp. 536-539. doi: 10.1109/WI. 2016.0091

Htay, H. H. and Murthy. K. N. (2008). Myanmar Word Segmentation using Syllable Level Longest

Matching. In Proceedings of the 6th Workshop on Asian Language Resources, ACL ID I08-7006.

Nagao, M. and Mori, S. (1994). A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese. In Proceedings of the 15th Conference on Computational linguistics (COLING '94), Volume 1, pp. 611-615.

Souza, M., Vieira, R., Busetti, D., Chishman, R. and Alves, I. M. (2011). Construction of a Portuguese Opinion Lexicon from Multiple Resources. In Proceedings of the 8th Brazilian Symposium on Information and Human Language Technology, STIL, Mato Grosso.

Hopple, P. (2003). The structure of nominalization in Burmese, Ph.D. Thesis.

Remus, R., Quasthoff, U. and Heyer, G. (2010). SentiWS: A Publicly Available German-language Resource for Sentiment Analysis. In Proceedings of the 7th International Language Resources and Evaluation (LREC'10), pp. 1168-1171.

Hlaing T. H. and Mikami. Y. (2013). Automatic Syllable Segmentation of Myanmar Texts Using Finite State Transducer. The International Journal on Advances in ICT for Emerging Regions (ICTer), Volume 6, Number 2.

Maung, Z. M. and Mikami. Y. (2008). A Rule-based Syllable Segmentation of Myanmar Text. In Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pp. 51-58, ACL ID I08-3010.