https://dx.doi.org/10.22161/ijaems.4.5.7 *ISSN: 2454-1311*

Senti-Lexicon and Analysis for Restaurant Reviews of Myanmar Text

Yu Mon Aye, Sint Sint Aung

University of Computer Studies, Mandalay

***Abstract*—** *Social media has just become as an influential with the rapidly growing popularity of online customers reviews available in social sites by using informal languages and emoticons. These reviews are very helpful for new customers and for decision making process. Sentiment analysis is to state the feelings, opinions about people's reviews together with sentiment. Most of researchers applied sentiment analysis for English Language. There is no research efforts have sought to provide sentiment analysis of Myanmar text. To tackle this problem, we propose the resource of Myanmar Language for mining food and restaurants' reviews. This paper aims to build language resource to overcome the language specific problem and opinion word extraction for Myanmar text reviews of consumers. We address dictionary based approach of lexicon-based sentiment analysis for analysis of opinion word extraction in food and restaurants domain. This research assesses the challenges and problem faced in sentiment analysis of Myanmar Language area for future.*

***Keywords*—Dictionary-based, Myanmar Language, Opinion Word Extraction, Senti-Lexicon, Sentiment Analysis**

I. INTRODUCTION

Nowadays, people are thrilled in online communication according to the rapid growth and development of World Wide Web. People express their opinion in social media with contents are usually unstructured texts. In Natural Language Processing, sentiment analysis (opinion mining) is an emerging field of artificial intelligence deals with analyzing opinions, sentiments and emotions articulated in informal data. Sentiment lexicons and opinion words extraction are main part of sentiment classification system.

There are no various resources and tools in sentiment analysis for Myanmar Language such as copora, sentiment lexicons and dictionary. So, we faced challenges and language specific problem. When writing social media texts, there are mainly two style of writing such as formal and informal style Textual reviews may contain sufficient information but it is often complex to work for unstructured review.

There is a problem that inconsistency of customer review between star rating and text reviews. This paper tackles the textual reviews to extract information. This information is vital for customers and organizer.

This paper is structured into seven sections. Section 2 provides state of art for other language. In section 3, the methods in sentiment classification are discussed. The center contribution of this paper is expressed in section 4. Section 5 shows in detail how to extract the opinion words, the way it is used to extract from Myanmar's text reviews. Section 6 gets the experiments to evaluate the work. In section 7, the last section concludes the paper and also presents an evaluation of this method.

II. LITERATURE REVIEW

Many studies have been carried out in the sentiment analysis. Researchers have proposed various approaches and developed different systems to deal with the problem. Most of systems are developed for the English language. In this section, we discuss the related works of sentiment analysis for other languages.

Rehman and Bajwa [10] presented Lexicon-based sentiment analysis for Urdu Language. This research intends at generating an application of Urdu comments on a variety of websites. Convoluted system architecture is conversed in specify with techniques employed; experiment process and establish results of 66% accuracy are premeditated and F-measure is 0.73.

Wu, et.al. [2] Discussed several common sentiment dictionaries into a larger dictionary. They expressed a language independent method of integrating existing sentiment dictionaries with value extrapolated from seed words. They built an evaluation Chinese Sentiment dictionary based on commonsense facts for sentiment classification of song lyrics system. They compared the performance iSentiDictionary with ANEW, SenticNet and SentiWordNet.

Akhtar, Ekbal and Bhattacharyya [5] proposed aspect based sentiment analysis in Hindi for resource construction and assessment. They assessed the dispute of sentiment in Hindi by providing benchmark setup by creating an annotated dataset of high quality with product

https://dx.doi.org/10.22161/ijaems.4.5.7 *ISSN: 2454-1311*

reviews from diverse online resource. This paper used CRF and SVM of classification algorithm for aspect term extraction and sentiment classification. The average F-measure is 41.09% for aspect extraction and 54.05% is the result of sentiment classification.

Santarcangelo, et.al [8] proposed an approach of Italian Language for social opinion mining by considering the state of art. They showed interesting approach based on Adjective, Intensifier and Negation (AIN) approach is built-up for Italian. This approach is based on the use of an Italian Sentiment Thesaurus developed by the writer and presented.

III. METHODOLOGY

Sentiment analysis can be classified into two approaches such as machine learning approach and lexicon based approach. Lexicon based approach handle by searching the sentiment words from the sentence and then compare with seed words. Two branches of this approach are dictionary and corpus based approach.

Machine learning contends with sample review for the sentiment words. There are two approaches in this approach. First, unsupervised approach which compares each word of the text with valued of positive and negative word for ranking. Second, supervised approach which uses equations to obtain the sentiment and various machine learning algorithms are used for sentiment classification. In this paper, we used lexicon based approach in our study.

3.1. Lexicon Based Approach

Sentiment words are used in many tasks of sentiment classification. The lexicon based approach is based on the statement that the appropriate sentiment orientation is the sum of the each sentiment words or phrases. This approach is an unsupervised learning approach since it does not require prior training datasets. Lexicon based approach deals with searching the lexicons such as adjective, adverb, verb, etc. from the sentence and comparing with seed words. Two approaches are: Dictionary Based Approach and Corpus Based Approach [3, 6]. In this paper, we used the dictionary based approach for sentiment analysis of Myanmar Language.

3.1.1. Dictionary Based Approach

Sentiment words are collected manually to form a small list, which is later developed by searching more words from a known corpora wordnet. Wordnet is a corpora which produces synonyms and antonyms for a word. The new words found exclusive of the seed words are included to the list. The process continues until now new words are found from the corpora [3].

3.1.2. Corpus Based Approach

This approach is to resolve the problem of dictionary based approach. Corpus based approach is not as efficient as dictionary based approach because there is a need to make a huge corpus for covering words and this approach is very difficult task [3]. It requires annotated training data to produces accurate semantic word.

1.2. Approach for Creation of Senti-Lexicon

There are two approaches to build the resource of sentiment lexicon are manual and automated.

3.2.1. Manual Creation of Sentiment Lexicons

Opinion lexicons are manually created which involves simply make a decision on the structure of the sentiment lexicon and annotate the list of lexicon with their polarity value. The list can be attained from the corpus and dictionary. As an outcome, no computational or algorithmic complexity is occupied. This is beneficial property for sentiment classification using an accurate resource is bound to execute better. However, the problem of this approach is time consuming.

3.2.2. Automatic Creation of sentiment Lexicons

Automatic methods are to overcome the disadvantage of manual sentiment lexicons creation. One of the most popular of several methods is to create the set of starting seed words with known sentiment orientation but enlarge the seed using an offered lexical resource. The advantages of an autonomous approach to the promise of high coverage are achieved only by compliance with its accuracy dictionary, as the methods used are perfectly excellent [4].

IV. BUILD LANGUAGE RESOURCES

The proposed system creates own dataset and extract the sentiment lexicon from the sentence.

4.1 Corpus Creation

We faced the largest problem which is the lack of available annotated data for Myanmar Language. To overcome this difficulty, we built the resource for our own corpus. We manually collected reviews of restaurants from the social media (Facebook). This corpus contains the objective and subjective reviews such as positive, neutral, negative reviews and mixed by writing with formal and informal style without any segmentation. Customers write the review with different opinion. Some write only positive review or negative review. Some writes the mix opinion both positive and negative reviews. Some reviews are not clear for their opinion.

In this paper, we collect 800 reviews of customers for food and restaurant domain from social media facebook page. These reviews contain opinionative and non-opinionative reviews. Sample of reviews are shown in Table 1.

Table 1. Sample of Restaurants' reviews.

Positive	အရသာကောင်းသော ဟင်းဖွယ်များ (good taste dishes)

https://dx.doi.org/10.22161/ijaems.4.5.7 *ISSN: 2454-1311*

Negative	ဝန်ဆောင်မှုတော်တော်ညံ့ပါသည် (bad customer service)
Neutral	ဈေးနှုန်းကတော့ ပုံမှန်ပါပဲ (price is regular)
Objective	စားချင်တယ် ဘယ်ကလဲ မကွေးကောရှိလား (I would like to eat this food. Where does this restaurant locate? Is this located in Magway city?

1.3. Creation of Myanmar Senti-Lexicon for food and restaurants' domain

Senti-Lexicon is a lexical resource for sentiment classification which is a database of lexical element for a language beside with their sentiment orientations. This section presents the creation of Myanmar Senit-Lexicon for food and restaurant domain. These are no any reference resources to classify the sentiment orientation in Myanmar Language. Therefore, in this thesis a lexicon that includes the sentiment words associated with a restaurant review is constructed by analyzing the restaurant reviews. We collect manually sentiment bearing words from the restaurants' reviews based on our knowledge and grew by searching more antonyms and synonyms for Myanmar Language. We made by using a based dictionary that available from Myanmar Lexicon (Version 2)-Lexique Pro. A small set of sentiment words for food and restaurants' reviews of customers' emotion are collected. And we also collect emoticons which are used to express their feeling.

We assign the polarity of the sentiment words and emoticons such as positive, negative and neutral with their target such as food and taste, place, price, staff, service and common. We change the polarity of each sentiment word into a numeric value to calculate the further computation i.e. positive=1, Negative=-1, Neutral=0. We collected 872 (817 sentiment words and 55 emoticons) which included 425 positive words, 428 negative words, 19 neutral. The sentiment lexicon (L) is made up of a set as

L={Target, Sentiment word, POS, Polarity}

The value that corresponds to target is a subject matter that expresses an emotion. This represents an evaluative attributed such as food and taste, place, price, staff and service that can feel when visiting a restaurant and sentiment lexicon is added to the previous work [12]. When target is not shown explicitly, it is expressed as common. Sentiment word expresses an emotional word. POS expresses a part of speech of the emotive word. A polarity is expressed as positive, negative or neutral of the emotional word. Additionally, a word such as "စား(eat)", which carries little emotive meaning, is eliminated [13].

Table 2. Example of Senti-Lexicon for Food and Restaurant domain

Target	Sentiment word	POS	Polarity
Food & taste	ညှီစော်နံ (smell acrid)	Verb	Negative
Place	ကျဉ်းကျပ် (cramped)	Adj	Negative
Service	ချက်ချင်း (immediately)	Adv	Positive
Staff	ပျူငှာ (be cordial)	Verb	Positive
Price	သက်သာ (be cheap)	Verb	Positive

1.4. Emoticons

Users of social media use a variety of emoticons such as :) :-) :D :-(and :P. Emoticons have been widely used in sentiment analysis as features or as entries of sentiment lexicons. Customers write the reviews with emoticons to express their feeling. This paper collects the 21 category of emoticons to classify the polarity of reviews such as happiness/smile, wink, amused, kiss, thumbs up, etc and included 55 emoticons icon [1].

Table 3. Example of Emoticons.

Emoticons	Category	Polarity
:)	Happiness/smile	Positive
:(	Sadness	Negative
:P	Kidding	Positive
:'(	Cry	Negative

https://dx.doi.org/10.22161/ijaems.4.5.7 *ISSN: 2454-1311*

V. OPINION WORD EXTRACTION OF MYANMAR LANGUAGE

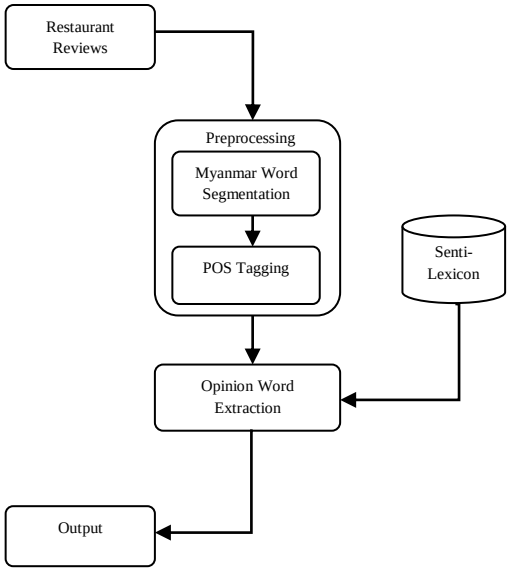

Fig. 1: System Architecture for Opinion Word Extraction

This section describes a method for performing of sentiment classification of restaurants' review by using lexicon based sentiment analysis. Opinion words extraction is also the essential part of the sentiment classification system. In this paper, we propose opinion word extraction for Myanmar restaurant reviews.

5.1. Preprocessing of Myanmar Text

Input texts of sentiment analysis are restaurants' reviews from social media which are Myanmar texts. Myanmar text is a sequence of characters without word boundary delimiters. Texts are written in string from left to right with no explicit word boundary markup. We need preprocessing steps of Myanmar reviews for informal and formal texts.

5.1.1. Segmentation of Myanmar Syllable

A syllable is a fundamental sound or sound unit. A word consists of one or more syllables. A Myanmar syllables has a base character, a post-base, an above based and a below base character. A syllable is formed based or rules that are quite specific and unambiguous in Myanmar text. An approach of rule based heuristic applies for segmentation of Myanmar syllable [11].

The following six syllable segmentation rules were proposed in [7] is used for syllable segmentation.

1. Single character rule
2. Special ending characters rule
3. Second consonant rule
4. Last character rule
5. Next starter rule
6. Miscellaneous rules (Non-Myanmar characters, Numeric characters, Punctuation marks, spaces and similar characters)

5.1.2. Syllable Merging

The next step is to merge the segmented syllables into words. Dictionary based approach with longest matching is used to perform syllable merging. Word segmentation for Myanmar language is an essential part which is prior to natural language processing (NLP). Syllable segmentation and syllable merging are two steps of Myanmar word segmentation [7].

The two methods of word segmentation can be roughly classified into dictionary-based and statistical methods. In dictionary-based methods, only words that are stored in the dictionary can be identified and the performance depends to a large upon the coverage of the dictionary. New words appear constantly and thus, increasing size of the dictionary is a not a solution to the out of vocabulary word (OOV) problem [9].

Although statistical approaches can identify unknown words by utilizing probabilistic and also suffer from some drawbacks. The main issues are: this approach requires large amounts of data and the processing time required. For low-resource languages such as Myanmar, there is no freely available corpus. We faced linguistic specific problem for the lack of resource such as lexicon and corpus for Myanmar sentiment classification. There is no large amount of data reviews to use this approach [9].

5.1.3. Part-of-Speech (POS) tagging

Part-of-Speech tagging is the makeup of assigning the suitable part of speech or lexical type. POS tagging is a primary task in Natural Language Processing (NLP).

5.2. Opinion Word Extraction

Opinion words are extracted from reviews based on Myanmar sentiment lexicon of food and restaurants domain such as Adjective, Verb, Adverb, Noun and emoticons. And the polarity is assigned to each word match with sentiment dictionary. An objective review based on fact information, while a subjective review expresses some personal opinions, beliefs, feelings, or impression. We also extract the opinion word with negation (not such as မ, မ---ဘူး).

Example: ဒီနေ့ မှာစားတာ အရမ်းစားကောင်းတယ် လာပို့တာလဲ မြန်တယ် (Today, very good taste and fast delivery service.)

Opinion word extraction: ကောင်း(good), မြန်(fast)

These reviews express the opinion and feeling. We can extract the sentiment words match with sentiment dictionary.

VI. EXPERIMENT AND RESULT

https://dx.doi.org/10.22161/ijaems.4.5.7 *ISSN: 2454-1311*

Customers write the reviews which contain the opinion words about their feeling, opinion and emotion. These opinion words are important to classify the polarity of sentiment analysis. In this section, we describe the extraction of opinion words from customers' reviews of Myanmar Language. In this paper, we used 500 customers' reviews to extract the opinion words by using the proposed 872 Myanmar Senti-Lexicon of food and restaurant domain.

Table 4. Opinion words extraction of Sample Reviews

Customers' Reviews	Extracted Opinion words from customers' reviews
အလွန်ကောင်းတယ် သန့်ရှင်းတယ် ဝန်ထမ်းတွေလည်း ဖော်ရွေတယ် (Good taste, clean and affable staff)	ကောင်း(good), သန့်ရှင်း(clean), ဖော်ရွေ(affable)
ကြိုက်တယ် မြန်မာမုန့်တွေက သန့်ရှင်းတယ် (I like Myanmar snacks and clean.)	ကြိုက်(like), သန့်ရှင်း(clean)
အရမ်းစားလို့ကောင်းတဲ့ ဆိုင်လေးပါ၊အေးချမ်းပြီး ဝန်ထမ်းတွေလဲ ပျူငှာကြပါတယ် စျေးနှုန်းလဲ သင့်တင့်ပါတယ် (This restaurant is good taste, frosty and cordial staff, decent price)	ကောင်း(good), အေးချမ်း(frosty), ပျူငှာ(cordial), သင့်တင့်(decent)
အရသာနဲ့ ပေးရတဲ့စျေးမဆိုးပါဘူး (taste and price are not bad)	မဆိုးဘူး(not bad)
ဝန်ထမ်း ဆက်ဆံရေးညံ့ (employee relation is bad)	ညံ့ (bad)
နေရာလည်းကောင်း အင်တာနက်လည်းကောင်း :) (good place and internet is good connection)	ကောင်း(good), ကောင်း(good), :) (emoticon)
တော်တော်လေး စိတ်ပျက်မိတယ် :'((somewhat disappoint)	စိတ်ပျက်(disappoint), :'((emoticon)
Casher က တစ်ခုတည်းနဲ့တန်းစီစောင့်ရတာ လုံးဝအဆင်မပြေပါဘူး (It is not convenience to wait with only one casher.)	စောင့်(wait), အဆင်မပြေ(inconvenience)
စားကြည့်ချင်တယ် ကွေ့တီယိုက ဘာလဲ မသိလို့နော် (I would like to eat "kwe ti yo", and what is it? I don't know.)	-
မနေ့ညက ကြက်သားအာလူး ကတ်သလိပ် ၁ပွဲ ၂၅၀၀နဲ့ ဝယ်စားတယ် (Yesterday, I ate "chicken potato" with 2500 kyats.)	-

Evaluation is used to calculate the overall performance of the proposed system with opinion word identification, error of extracted opinion words and not extracted opinion words. We compared with manually extracted opinion words 1113 which contains 38 emoticons icons from 500 reviews which are chosen randomly from 800 reviews. In this experiment, we cannot extract 66 opinion words from the review and extract 98 opinion words incorrectly. We can extract all of 38 emoticons icons and 977 opinion words contain in the customers' reviews correctly.

Accuracy = no. of correct opinion word extraction

Table 5. Analysis of Opinion Words Extraction

Description	Result
Accuracy of Opinion word extraction (contains 3% of emoticon icon extraction)	85%
Error of Opinion words extraction (cannot extract the opinion word (6%) + incorrectly extracted opinion word (9%))	15%

https://dx.doi.org/10.22161/ijaems.4.5.7 *ISSN: 2454-1311*

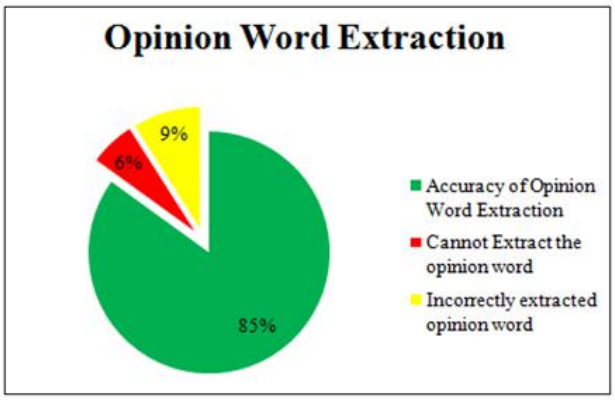

Fig. 2: Result of Opinion Word Extraction

We can extract 85% accuracy for the opinion words from the customers' reviews. This is important in sentiment analysis for food and restaurant domain and aims to classify the polarity by using this sentiment words for the sentiment analysis of the customers' reviews to develop the business.

VII. CONCLUSION

This paper built the Senti-Lexicon using manual approach to over the challenges and language specific problem. We used lexicon based approach to extract the sentiment word extraction. This system tested with 500 customers' review randomly without unseen reviews and simple. In this paper, we proposed the resources for food and restaurant domain of Myanmar Language and analysis of opinion word extraction. We can extract opinion words 85% correctly with proposed lexicon. This lexicon does not contain the informal opinion words and cannot extract the informal opinion words from informal reviews. We extracted the opinion words incorrectly from comparison reviews and due to spelling error. For future work, we need to improve the performance of opinion words extraction, to cover both formal and informal reviews and to classify the subjectivity classification contain sentiment analysis such as subjective review (positive, negative or neutral) reviews and objective reviews.

REFERENCES

[1] Vashisht, G. and Thankur, S. (2014). Facebook as a Corpus for Emotions-Based Sentiment Analysis. In *IJETAE*, 904-908.

[2] Wu, H.H., Tsai, A.C.R., Tsai, R.T.H. and Hsu, J.Y.J. (2013). Building a Graded Chinese Sentiment Dictionary Based on Commonsense Knowledge for Sentiment Analysis of Song Lyrics. In *Journal of Information Science and Engineering*, 647-662.

[3] Haseena Rahmath, P., and Ahmad, T.(2014). Sentiment Analysis Techniques - A Comparative Study. In *IJCEM International Journal of Computational Engineering & Management*, Vol. 17, Issue 4, 25-29.

[4] Korayem, M., Crandall, D., Abdul-Mageed, M. (2012, December). Subjectivity and Sentiment Analysis of Arabic: A Survey. In *International Conference on Advanced Machine Learning Technologies and Applications*, 128-139.

[5] Akhatar, M. S., Ekbal, A. and Bhattacharyya, P. (2016). Aspect based Sentiment Analysis in Hindi: Resource Creation and Evaluation, In *LREC*, 2703-2709.

[6] Rohini,V and Thomas, M. (2015). "Comparison of Lexicon based and Naïve Bayes Classifier in Sentiment Analysis", In *IJSRD - International Journal for Scientific Research & Development*, Vol. 3, Issue 04, 1265-1269.

[7] Thet, T.T, Na, J.C and Ko, W.K. (2008). Word segmentation for the Myanmar language. In *Journal of Information Science*, 34 (5), 688-704.

[8] Santarcangelo, V., Oddo, M. Pilato, G., Valenti, F. and Fornaro, C. (2015). Social Opinion Mining: an approach for Italian language. In *International Conference on Future Internet of Things and Cloud*, IEEE, 693-697.

[9] Teahan, W. J., Wen, Y., McNab, R., and Witten. I. H. (2000). A compression-based algorithm for Chinese word segmentation. *Computational Linguistics*, 26(3), 375-393.

[10] Rehman. Z.U. and Bajwa. I.S. (2016). Lexicon-based Sentiment Analysis for Urdu Language. in *Sixth International Conference on Innovative Computing Technology, IEEE*, 497-501.

[11] Myanmar lexicon analyzer – Sorting and Segmentation. Retrieved from https://github.com/ minthanthtoo/myanmar-collation-stats.

[12] Aye, Y.M. and Aung, S.S. (2017). Sentiment analysis for reviews of restaurant in Myanmar Text, In *18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, Kanazawa, Japan, 321-326.

[13] Kang, H., Yoo, S. J., and Han, D. (2012). Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. In *Expert Systems with Applications*, 39(5), 6000-6010.