Sentiment Analysis for Reviews of Movie in Myanmar Text

Wai Phyoe Zaw
University of Computer Studies (Magway)
Magway, Myanmar
cumgy.wpz.2019@gmail
waiphyozaw13@ucsmgy.edu.mm

Zar Chi Su Su Hlaing
Faculty of Information Science
University of Computer Studies (Magway)
Magway, Myanmar
zarchisusuhlaing@ucsmgy.edu.mm

***Abstract*—Nowadays, a lot of information is valuable for people who want to predict and decide the market or economic or political or many different areas. Because of the information age, customers are also no longer interested in trying it out for themselves. They are judging a product or an item based on the opinions of other customers. It is valuable in social media monitoring because several reviews that were written by a lot of users in social media mainly include forecasting to judge whether it is good or not. Many researchers have done sentiment analysis to obtain a summarized report for an opinion about different contexts. Moreover, researchers want to build a lexicon with their language. Some researchers in Myanmar have also implemented sentiment analysis to obtain Myanmar lexicon in some areas such as restaurants, hotels, movies. This system focuses on building a lexicon with Myanmar Text in Movie comments and predicts Movie comments on Facebook whether these are positive or negative or neutral by using a lexicon-based approach.** Researchers from Myanmar have difficulties with any annotated **data for Myanmar Language in sentiment analysis. So, this system also completely meets a challenge with no annotated data. This system uses word segmentation, removal of punctuations, and correction of some words in pre-processing steps. In the evaluation steps, four metrics are calculated accuracy, precision, recall, and f-measures to know the overall performance and accuracy of the proposed system.**

Keywords— sentiment analysis; sentiment lexicon; lexicon-based; Myanmar Text

I. INTRODUCTION

Sentiment analysis is the classification of texts, emotions or reviews to get results for economic or other fields and gives the recommendations users' reviews whether these are good or not. There are a lot of websites which are selling or marketing products on the internet. Later economical contrivers have known the importance of reviews from online users and understood the critical comments from online are the valuable data to use in the marketing of their businesses. Many researchers are emphasizing sentiment analysis or opinion mining with a variety of areas to predict future results with user's comments or reviews or ratings. So, there are many sentiment analysis implementations or researches in movies, restaurants, hotels with users' reviews on the internet's sites such as Facebook, Twitter, and Amazon. Among them, sentiment lexicons have mostly created with the English language by researchers.

So, this system mainly focuses on achieving sentiment lexicon with Myanmar Text for Movies on Facebook and concentrates on determining reviews of movies that are good or not. Today users can't see all comments because of the enormous mass of comments. Sentiment analysis helps the conclusion of polarity for users to judge these movies have weakness or advantages and for producers to know by observing users' reviews whether users are pleasant or not their movies. So, they can improve or modify their movie's qualities with sentiment analysis's research easily.

In our knowledge, no implementations with sentiment analysis on Facebook Movies with Lexicon based. Lack of annotated data with Myanmar Language solves by collecting data manually.

II. RELATED WORKS

Sentiment Analysis is a popular topic for determining a topic or product is like or not from customers. There are several implementations for sentiment analysis techniques. We present the related works for sentiment analysis topics in this sector.

In [1], the authors described the sentiment analysis method based on lexicon-based for foods and restaurants in Myanmar Language. For language challenges, they collected sentiment words from social media and built a lexicon with 766 sentiment words. They focused on the extraction of the relevant pairs of words from the users' reviews on food and restaurant and classified as positive, negative, and neutral. Moreover, they also concentrated on generating content-rules to calculate sentiment polarity. The accuracy is 96% with 500 customers' reviews without considering unseen and confuses objective and subjective sentences in Myanmar Language.

In [2], the authors have presented the implementation with aspect level opinion tasks for Hotel with Myanmar Language. The author describes the complex reviews and opinions for each level of Myanmar Language as positive, negative, and neutral. It also presents a different perspective if it does not contain complex reviews. The system determined the challenges of opinion words from compound sentences or complex sentences in the Myanmar Language.

In [3], the authors intended to solve language-specific challenges for Myanmar Language. So, they focused objective words or subjective words or intensifier words in an informal style for the food and restaurant domain. They did their best to get better accuracy and classification than without considering of intensifiers and objective words in sentiment analysis for Myanmar language-specific challenges.

In [4], the authors implemented to solve language-specific challenges with lexicon-based sentiment analysis for Persian text. For the lexicon-based sentiment analysis

technique, they needed a wide range of Persian vocabulary entries and their sentiment. They have tried to cover both formal and informal cases in their lexicon. An evaluation of the developed GATE pipeline shows an encouraging overall accuracy of up to 69%.

In [5], the authors proposed the assigning of polarity scores for Facebook Movies in Myanmar Text. They calculated the polarity scores of each positive and negative word in the movie domain-specific polarity lexicon. And then the polarity scores to each comment of the plain text movie corpus was assigned. The system archived 89% and 85% accuracy on positive and negative opinion words respectively in the evaluation of polarity score lexicon. They also made the comment polarity for 3-class evaluation and 5-class evaluation based on the scores of comments.

III. METHODS IN OPINION ANALYSIS

Two types of approaches are machine learning and lexicon-based in sentiment analysis. Although it is a needed training dataset in machine learning, it is not lexicon-based. In machine learning, text reviews use by supervised or unsupervised methods to calculate polarity as positive or negative. On the other hand, rule-based methods use a sentiment dictionary to decide an item or product is good or bad.

A. Lexicon based approach

- **Dictionary-based approach**: One of the simple techniques in this approach bases on bootstrapping using a small set of seed opinion words and an online dictionary. The strategy is to first collect a small set of opinion words manually with known orientations and then to grow this set by searching in the WordNet for their synonyms and antonyms. The newly found words add to the seed list. The next iteration starts. The iterative process stops when no more new words found. The opinion words collected from it have one major weakness. The approach is unable to find opinion words with field and context-specific orientations, which is quite common.
- **The corpus-based approach**: It can help deal with the problem of a dictionary-based approach. The methods in the corpus-based approach rely on syntactic or co-occurrence patterns and also a seed list of opinion words to find other opinion words in a large corpus. It can help to find the domain and context-specific opinion words and their orientations using a domain corpus. [6]

IV. BUILDING LEXICON

This section builds sentiment words for Myanmar movies dataset to be used in the system, and then extracts the sentiment words and calculates the polarity. Although other languages are rich in the lexicon, the lack of annotated data regarding the Myanmar language is the main problem we faced. Because the system is a lexicon-based technique implementation, it is needed to have a lexicon.

Because there is any annotated dataset for Myanmar language, we create a dataset manually from the Facebook page with a collection of positive words, negative words, compound words, and inflection words regarding the target of Movies and Actors or Directors. Some reviews are unclear reviews that cannot define in any opinion. There were 200 positive words, 100 negative words, and intensifier 32 words for a total of 332 sentiment words. Intensifier words are assigned to the percentage point by [8]. The lexicon includes opinion words for negative, positive, negation, and intensifier as shown in Table I.

TABLE I. SAMPLE OF SENTIMENT LEXICON OF MOVIE REVIEWS

No	Target	Sentiment Word	Polarity
1	Movies	ကောင်း (good)	positive
2	Movies	ကြည်နူး (enjoy)	positive
3	Movies	ပျင်းစရာ (bore)	negative
4	Movies	ပေါကား (silly)	negative
5	Movies	ညံ့ (bad)	negative
6	Actor	ချစ်စရာကောင်း (sweet)	positive
7	Director	လေးစား (respect)	positive
8	Common	အရမ်း (very)	25%
9	Common	တအား (special)	50%
10	Common	နည်းနည်း (little)	-20%

Intensifiers are mostly adverbs, but sometimes noun phrases and prepositional phrases. These are used to show an increase or decrease in the power of verbs, adjectives, and adverbs. Amplifiers have strengthened the meaning of other expressions and show emphasis such as Downtowners are words or phrases such as which reduce the power of another phrase or word.

Some Myanmar intensifier words which is particle suffixed to verbs or adjectives to convey the meaning of excessive. The polarity of intensifiers based on 100%. The sentiment dictionary is the polarities of the words in the dictionary are set according to a specific domain such as restaurants' reviews. The same word in different domains can have different meanings; the dictionary used in this approach is made for the movie review domain. The dictionary also contains the polarity of every word. [1]

V. SENTIMENT ANALYSIS OF MOVIE REVIEWS

The system focuses on the calculation of performances based on lexicon-based. The diagram for the proposed system describes in Figure 1.

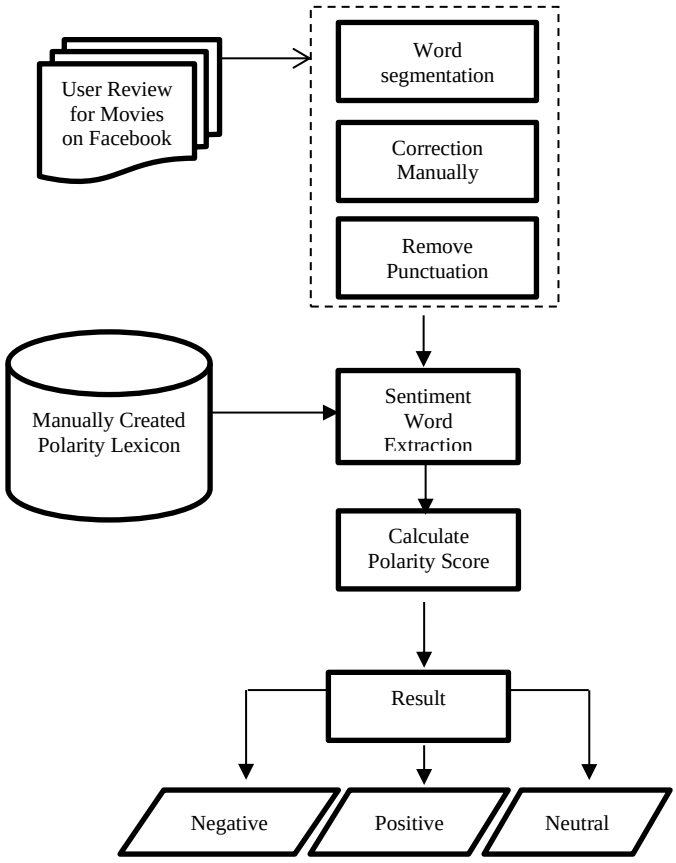

Fig. 1. Flowchart of the lexicon-based sentiment analysis of Movie Reviews

A. Data processing

Input texts of sentiment analysis are movies' 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 pre-processing steps of Myanmar reviews for informal and formal texts.

1) Word Segmentation of Myanmar Text: A word consists of one or more syllables. A syllable is a fundamental sound or sound unit. A Myanmar syllable has a base character, a post base, an above base 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. 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].

2) Spelling checking: IN this system, we collected movie' reviews in Myanmar Text from Facebook which have been written by different users. Some reviews are vague and misspelled. Such misspelled words need to be corrected.

B. Sentiment word extraction and calculate polarity score

Lexicon-based sentiment analysis is mainly to extract aspect words in a sentence with a sentiment dictionary. The system takes every word that has an opinion word meaning for adverbs, verb, and noun phrase. And the system converts those opinion words into numeric such as positive is 1, negative is -1, and neutral is 0.

C. Reviews Sentiment calculation

We presented the estimating of review as a polarity in the sentence-level calculation as following.

$$P(S) = P(OW) \quad (1)$$

$$P(S) = [\ 100\% + P(I)\ ] * P(OW\ ) \quad (2)$$

$$P(S) = [\ 100\% + P(I)\ ] * [\ 100\% + P(I)\ ] * P(OW) \quad (3)$$

$$P(S) = (-1) * P(OW) \quad (4)$$

$$P(T) = \sum_{i=1}^{n} P(S)_i \quad (5)$$

$$P(R) = P(T)/N \quad (6)$$

Where,
P(OW) is opinion words,
P(I) is intensifier of Myanmar words,
P(T) is sum of all polarity,
P(S) means score of opinion words,
P(R) is ratio of score,
N is number of opinion words,

Above the overall score of each equation consider positive is equal or greater than +0.3, negative is equal or less than -0.3, and neutral is between +0.3 and -0.3.

If the sentence has only opinion word, the system uses Equation (I).
Example 1:
ဒီဇာတ်ကားကကောင်းတယ်(This movie is nice.)

Extract word: ကောင်း(nice)

P(S) =1
The polarity score is greater than +0.3, so this sentence is positive.

Example 2:
ဒီဇာတ်ကားကဆိုးတယ်(This movie is bad.)

Extract word: ဆိုး(bad)

P(S) = -1
The polarity score is less than -0.3, so this sentence is negative.

If the sentence has opinion word and one intensifier word, the system uses Equation (II).

Example 3:
ဒီဇာတ်ကားကအရမ်းကောင်းတယ်(This movie is very good)

Extract word: အရမ်းကောင်း(very good)

P(S) = [100% + 25%] * 1 = 1.25
The polarity score is greater than +0.3, so this sentence is positive.

If the sentence has opinion word and two intensifier word, the system uses Equation (III)

Example 4:
ဒီဇာတ်ကားကအရမ်းအရမ်းကောင်းတယ်။
(This movie is very very nice)

Extract word: အရမ်းအရမ်းကောင်း (very very nice)

P(S) = [100% + 50%] * [100% +50%] * 1 =2.25
The polarity score is more than +0.3, which means as positive.
If the sentence has opinion word and one negation, the system uses Equation (4).

Example 5:
ဒီဇာတ်ကားကမကောင်းဘူး။(This movie is not good).

Extract word: မကောင်း

P(S) = (−1) * P(OW)
P(S) = −1
The polarity score is less than -0.3, so this sentence is negative.

If the sentence has mix opinion words, the system considers intensifier words.

Example 6:
ဒီဇာတ်ကားကတော်တော်ကောင်းတယ်၊မင်းသမီးကလည်းအရ
မ်းတော်တယ်။(This movie is pretty good actor is very good)

Extract word: တော်တော်ကောင်း

အရမ်းတော်

တော်တော်ကောင်း = [100% - 20%] * (1) = 1.2

အရမ်းတော် = [100% + 50%] * 1 = 1.5

The system uses Equation (5).
The system uses Equation (6).
P(T) =2.5/2=1.25
The polarity score is greater than +0.3, so this sentence is positive.

We do not consider conjunctions. And some sentences cannot describe any opinion if we do not consider intensifier.

Example 7:
ဒီဇာတ်ကားကနည်းနည်းဆိုးပေမဲ့မင်းသမီးကအရမ်းကောင်းတ
ယ်။ (This movie is little good but actress is very good)

Extract word: ဆိုး, ကောင်း

ဆိုး= -1

ကောင်း= 1

The system uses Equation (5).
P(T) = 1-1=0
And then system uses Equation (6).
P(R) =0
This sentence is neutral because the score is between -0.3 and +0.3.

If we consider the intensifier,
Extract word: နည်းနည်းဆိုး
အရမ်းကောင်း

နည်းနည်းဆိုး= [100% - 50%] * (-1) = -0.5

အရမ်းကောင်း= [100% + 50%] * 1 = 1.5

The system uses Equation (5).
P(T) = 1.5+(-0.5)=1
And then system uses (6).
P(R) =1/2=0.5
This sentence is positive due to greater than +0.3.

VI. Performance Evalution

We collected 500 reviews regarding movie comments from Facebook as 200 reviews for positive, 200 reviews for negative, and 100 reviews for neutral. We used F-measure, Precision, Recall, and Accuracy to analyze the performance of reviews. F-measure is a value between the balances of Precision and Recall. Accuracy focuses on the aggregates of exact prophecies. Precision and Recall are effective ways to evaluate the correction of classes.

PR represents precision,

RE is Recall,

FM is F-measure,

TP is True Positive,

FP is False Positive,

TN is True Negative,

FN is False Negative,

ACC is Accuracy Respectively,

$$PR=\frac{TP}{TP+FP}$$

$$RE=\frac{TP}{FN+FP}$$

$$FM=2\frac{PR.RE}{TP+FP}$$

$$ACC=\frac{TP+TN}{TP+FP+TN+FN}$$

TABLE II. FOUR MATRIX CALCULATION RESULTS

	Accuracy	**Precision**	**Recall**	**F-Measure**
Positive	95%	95%	96%	95%
Negative	96%	93%	93%	93%
Neutral	99%	82%	94%	88%

These evaluation results show 95% for positive, 96% for negative, 99% for neutral, and 94% for total accuracy in sentiment calculation in Table II.

VII. CONCLUSION

The system presented the lexicon-based sentiment analysis for movies with Myanmar Language and collected 500 reviews from Facebook. We calculated the sentiment polarity by rules and accuracy is over 94% because the reviews are uncomplicated comments, not including any confusing sentences, subjective sentences, or objective sentences, conjunction sentences. Myanmar Language has many challenges about opinion. It has difficulties for building a lexicon in Myanmar Language. A word in Myanmar Text has different meanings and opinions. A word in Myanmar Text has different meanings and opinions. Without thinking about these facts, we avoided such sentiment words, mainly focused on creating a lexicon for Movies with Myanmar Language. So, this system is very pure and comes with clear reviews. If there are only ambiguous reviews, the accuracy will be lower the system will be better. In the next plan, we wish to concentrate on including confusion reviews and aspect-based opinion. We will try to try it with more reviews. Moreover, we will continue to work on the system with perspectives that will help the Myanmar language.

ACKNOWLEDGMENT

First of all, the author expresses to parents who have supported finically to study and to research in life. The author also would like to thank Dr. Zar Chi Su Su Hlaing, Professor, and Head of the Faculty of Information Science, for her valuable suggestions and guidelines throughout my works.

REFERENCES

[1] Yu Mon Aye, Sint Sint Aung ,"Sentiment Analysis for Reviews of Restaurant in Myanmar Text" 2018,pp(322-323)

[2] Cho Cho Hnin, Naw Naw, Aung Win, "Aspect Level Opinion Mining for Hotel Reviews in Myanmar Language" in 2018 IEEE International Conference on Agents (ICA), pp (132-135)

[3] Yu Mon Aye, Sint Sint Aung, "Enhanced Sentiment Classification for Informal Myanmar Text of Restaurant Reviews" pp(31-36).

[4] Fatemeh Amiri , Simon Scerri , Mohammad H. Khodashahi, "Lexicon-based Sentiment Analysis for Persian Text", pp(9-16)

[5] Win Win Thant , Nyein Thwet Thwet Aung , Su Su Htay, "Assigning Polarity Scores to Facebook Myanmar Movie Comments", International Journal of Computer Application (0975-8887), Volume 177- No 6, November 2017, pp(24-29)

[6] M.J. Carey, S Ceri, P Bernstein, U. Dayal, C Faloutsos, J.CFreytag,G Gardarin, W Jonker V, Krishnamurthy, M.-A Neimat, P Valduriez, G Weikum, K.-Y Whang, J. Widom, "Web Data Mining",pp(478-480)

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

[8] M.Z. Asghar, A. Khan, S. Ahmad, M. Qasim and I.A. Khan, "Lexicon enhanced sentiment analysis framework using rule-based classification scheme", in PloS one 12, no. 2 (2017): e0171649,pp(8-9)