Aspect based Sentiment Analysis for travel and tourism in Myanmar Language using LSTM

Soe Yu Maw, May Aye Khine
University of Computer Studies, Yangon
soeyumaw@ucsy.edu.mm, mayayekhine@ucsy.edu.mm

Abstract

Big social data analytics is an important tool which can be used to reveal the important insight of the information from the social user. It is an approach which combines various statistical methods, sentiment analysis, multimedia management and social media analytics for forecasting and predicting people and analyzing trends. In Myanmar, most of people use social media, especially Facebook, to express their opinion about specific topic in Myanmar language. Customer's comment and reviews are valuable, and are important source of data for multiple purposes. There are various method were introduced for performing sentiment analysis, still there are not efficient in extracting the sentiment features from a given context of text. In this paper, aspect based sentiment analysis of hotels' and restaurants' reviews using Long Short-Term Memory (LSTM) is proposed.

Keywords: *Sentiment Analysis, Long Short-Term Memory, Big social data analysis*

1. Introduction

Sentiment analysis is the analyzing the user's opinions, feelings and attitudes on the specific products and services. It is the process of transforming information for useful business intelligent information from unstructured text data. It is generally classified the expression as positive, negative or neutral.

People have been expressed their travel experiences, feelings and attitudes on social media as a status post or reviews about hotels, views, places, restaurants. In this paper, we collected hotels and restaurant's reviews on each individual Facebook pages and status and its comments from hotel and restaurant review page. Sentiment Analysis can be categorized as document level sentiment analysis, sentence level sentiment analysis and aspect-based sentiment analysis. A negative opinionated document on a particular object doesn't mean that the opinion holder has negative opinion on all aspect or features of the object. A positive opinionated document doesn't mean that the user like everything. In this situation, document level and sentence level classification fail such information. Therefore, aspect level sentiment classification is proposed to obtain such detail information on each aspect. This paper proposes a model for aspect based sentiment analysis for hotel's and restaurant's reviews and comments written in Myanmar language using LSTM neural network. Aspect based sentiment analysis are divided into three sub-tasks, prediction of how many aspects in each review contain, extraction those aspects and classification of sentiment polarity of those aspects.

This paper organized as follows. The related pervious work on sentiment analysis discuss in section 2. In section 3, type of sentiment analysis and methods in sentiment analysis are discussed in section 4. In section 5, presents the architecture of the proposed system. Finally in section 6, we discuss the conclusion and our future work.

2. Related Works

Aspect based sentiment analysis aims to detect an aspect (features) in a given text and then perform sentiment analysis of the text with respect to that aspect. [1] They performed aspect based sentiment analysis on the micro-blogs and headlines of financial domain. They proposed two neural network, Bidirectional LSTM use for aspect extraction from given text and multi-channel CNN for sentiment prediction.

Venu Dave and DhwaniShah, presented that performs the classification of customer reviews of three places by means of a sentiment analysis using Naïve Bayes classifier and R tool. It analyze dataset of reviews and classify them into three categories – Positive, Negative and Neutral. The strategy steps are review extraction with csv format from trip-advisor and goibibo, text preprocessing, transformation/build

a term-document matrix, classify the input using Naïve Bayes classifier and generate the Summary, accuracy. [2]

Poria S., Cambria E., used a 7-layer deep convolutional neural network to tag each word in the review data as aspect word and non-aspect word. They have also comprehended a small set of grammar constructs to use in combination with neural nets to improve the accuracy. Here, features are word embedding along with their part of speech tags. This work shows that a deep CNN is more efficient than existing approaches for aspect extraction. [3]

Muhammad Afzaal aimed to study the sentiment analysis on Twitter data written by Thai language using deep learning techniques. [4] They compared two deep learning techniques – Long Short-Term Memory (LSTM) and Dynamic Convolutional Neural Network (DCNN) to other bag of word model. The result showed that both deep learning techniques have higher accuracy compare to traditional method: Naïve Bays and SVM.

Jalaj S. Modha proposeed an approach to handle objective as well as subjective sentences and find opinion from them. In their proposed system they followed following steps: (i) Firstly they classified sentences as opinionated and non-opinionated. (ii) Then, they classified opinionated sentences into subjective or objective. (iii) Third step is to classify subjective sentences into negative, positive or neutral category. (iv) Then, classify objective sentences into positive, negative or neutral. They provided context or sentiment orientation as and when needed. [6]

The system [7] outlines a large-scale distributed system for real-time sentiment analysis on Hadoop. There are two components in their system: a sentiment classifier and lexicon builder. These components are implemented using a map-reduce framework and distributed database. A method is introduced to combine sentiment lexicon with machine learning algorithm and improvement in accuracy is observed.

Win Win Thant, Kiyoaki Shirai presented approach of Myanmar language for lexicon based method. They construct Myanmar movie lexicon using bootstrapping approach. The proposed method is based on n-grams of syllables without word segmentation. [5]

3. Type of Sentiment Analysis Levels

Sentiment analysis is the analyzing the user's opinions and convert unstructured data to structure data. There are 3 levels of sentiment analysis that has been studied; document, sentence, and entity or aspect level.

Document level: Overall sentiment of a complete document is decided in this level. For instance, if review of product is given, the task is to decide whether it convey an overall negative or positive opinion regarding the product. The job is to verify whether the whole document is negative, positive or neutral.

Sentence level: The job at this phase is limited to sentences and test if each sentence conveyed a negative, positive or neutral opinion Firstly, sentence is classified as objective or subjective and then sentences which are subjective are categorized as positive, negative or neutral.

Aspect and entity level: This level is more challenging than the other two. Aspect level analyses the opinions instead of analyzing paragraphs, documents, phrases or sentences. It provides finer-grained analysis for each aspect. Opinion is a phrase which consists of a target/subject and sentiment on the target. This idea is used in entity level analysis. This helps in understanding sentiment analysis problem better.

In this paper, we focus on aspect based sentiment classification. Document and sentence level classification classify overall polarity in its level. A sentence may contain different aspects and opinion. For example, food is decent and also price is fair but service is so bad, for aspects food and price is positive polarity while service is negative polarity. Therefore, it is worthwhile to explore the connection between an aspect and the content of a sentence. We extract all aspects containing in the review and calculate polarity for each aspect.

4. Methodology

Methods of sentiment analysis can be categorized into three approaches such as lexicon based approach, machine learning and deep learning based approach and hybrid approach (combine machine learning and lexicon based approach). In this paper, we use deep learning approach in our study.

Lexicon based approach search lexicons from the sentence and compare with the seed words. Lexicon consists of list of words and expressions.

This approach is based on count the number of positive words and negative word from a given text. Two approaches are corpus based approach and dictionary based approach. Machine learning algorithms such as SVM, Naive Bayes can be addressed as a combination of methods to automatically detect the available pattern in the given set of data. There are supervised and unsupervised approaches in this approach.

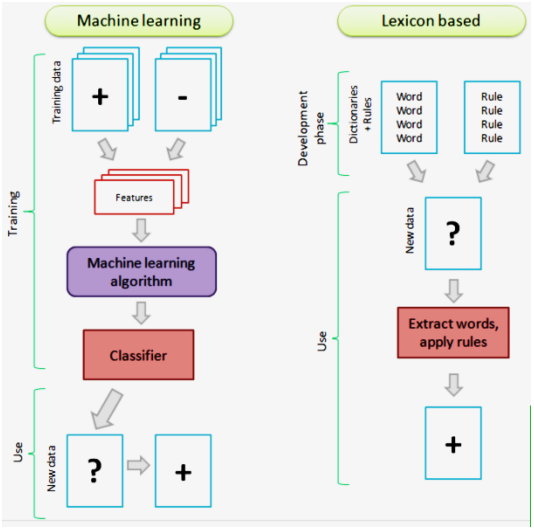

Figure 1. Machine Learning Vs Lexicon based approach

4.1 Recurrent neural networks (RNN)

Recurrent neural networks (RNNs) are designed specifically to learn sequences of data and are mainly used for textual data classification. Inputs from earlier data points in a sequence still have an influence on later iterations, which closely resembles the work process of the human memory on storing information. RNN occur vanishing gradient problem when handling long sequence of data.

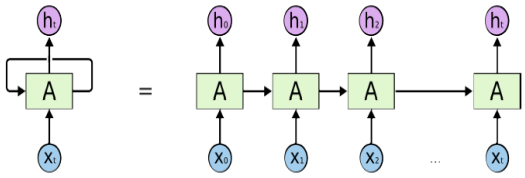

Figure 2: An RNN takes an input xt at every time step t and produces an output ht

4.1.1 Long Short-Term Memory (LSTM)

The typical sequence modeling method that reveals the sequential information from the beginning to the end of the sentence. A typical LSTM cell contains three gates: forget gate, input gate and output gate. These gates determine the information to o_w in and o_w out at the current time step. The cell is denoted as below:

$$f_i = \sigma(W_f[x_i, h_{i-1}] + b_f)$$
$$I_i = \sigma(W_I[x_i, h_{i-1}] + b_I)$$
$$\widetilde{C}_i = tanh(W_C[x_i, h_{i-1}] + b_C)$$
$$C_i = f_i * C_{i-1} + I_i * \widetilde{C}_i$$
$$o_i = \sigma(W_o[x_i, h_{i-1}] + b_o)$$
$$h_i = o_i * tanh(C_i)$$

where f_i, I_i and o_i are the forget gate, input gate and output gate respectively. W_f , W_I , W_o , b_f , b_I and b_o are the weight matrix and bias scalar for each gate. C_i is the cell state and h_i is the hidden output. A single LSTM typically encodes the sequence from only one direction.

4.1.2 Bidirectional LSTM

Two LSTMs can also be stacked to be used as a bidirectional encoder, referred to as bidirectional LSTM. For a sentence s = {w_1,w_2, _ _ _ ,w_L}, Bi-directional LSTM produces a sequence of hidden outputs,

$$H = [h_1, h_2 ... h_L] = \begin{bmatrix} \overrightarrow{h_1} & \overrightarrow{h_2} & \cdots & \overrightarrow{h_L} \\ \overleftarrow{h_1} & \overleftarrow{h_2} & \cdots & \overleftarrow{h_L} \end{bmatrix}$$

where each element of H is a concatenation of the corresponding hidden outputs of both forward and backward LSTM cells. It add reverse sequential learning step to LSTM, Bi-LSTM models both begin to end and end to begin sequential information.

5. Proposed System

5.1 Data Collection

We collected reviews, status posts and comments from Facebook pages. We only collected Myanmar language and data has written with other language are removed. Opinions and feelings are expressed in different way, with different vocabulary, context of writing, usage of short forms and slang, makes data huge and disorganized. The text data contains positive, negative, neutral reviews and mixed by

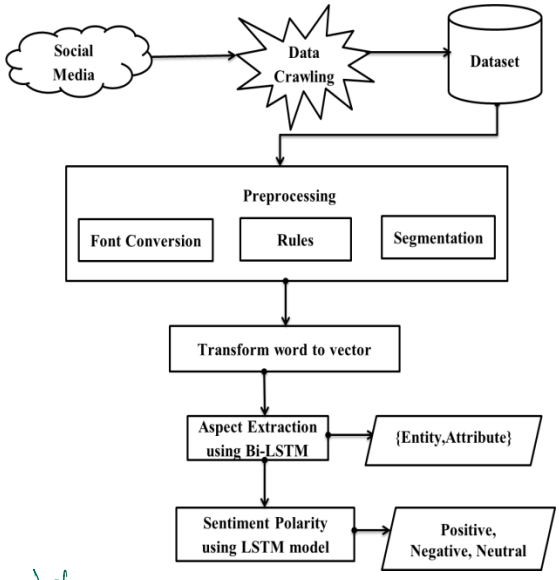

Figure 3: Proposed System

writing with formal and informal writing style without segmentation. Customers express positive, negative, neutral and sometimes both positive and negative opinion in the review. In this paper, we collect about one thousand reviews of customer for hotel and restaurant domain from social media Facebook page.

Table 1. Sample of Restaurants' reviews

ဟိုတယ်ပါရမီကသန့်ရှင်းသပ်ရပ်ပြီး ရှုခင်းတွေလဲလှတယ်။ (Hotel Parami Yangon is nice, clean and the best for seeing beautiful view, Yangon)	positive
သန့်ရှင်းတယ်၀န်ဆောင်မှုလည်းကောင်းတာကောင်းတယ် နောက်တစ်ခါလာတည်းအုံးမယ်။ (clean, good service and I will be back next time)	positive
အရမ်းကိုဆိုး၀ါးတဲ့ ဟိုတယ်ပါ အခန်းတိုင်းမှာလဲ wifi မရဘူး ၀န်ထမ်းတွေလဲအတွေ့အကြုံမရှိ ၀န်ဆောင်မှုလဲ အင်မတန်ညံ့ပါတယ် (So so awful hotel, wifi not available in every room. Staff all are also very bad service and not experience.)	negative

5.2 Preprocessing Steps

5.2.1 Font Conversion

Most of user in Myanmar use Zawgyi font on social media and application in technology field use Unicode. In this research, we collect text data both Zawgyi and Unicode format and then convert to Unicode only using online Zawgyi-Unicode converter.

5.2.2 Rules for preprocessing

This is the conversion step from unstructured to structured data. One difficultly of preprocessing of the text is containing the textual errors such as spelling and grammatical errors. Most of reviews and comments also include emoticons as well as text data. Emoticons are considered to be reliable indicators of sentiment and hence could be used either to automatically generate a training corpus or to act as evidence feature to enhance sentiment classification.

Preprocessing steps:

1 Emoticons: The emoticons symbolic representation is converted in to words at this stage. Emoticons are replaced with their Myanmar Words

Eg ☺ :) =ပြုံးသည် ☹ : (=မဲ့သည်

2 Useful English words and loan word are first convert to lowercase and translated to similar Myanmar words

Eg like=ကြိုက်သည် good=ကောင်းသည်

3 Myanmar word with English pronunciation are translated to Myanmar words

Eg ၀ွတ်တယ်=ကောင်းတယ်

4 certain Myanmar synonym adverb like

အားကြီး ၊ အကုန် ၊ အသေ -> အရမ်း

5 Word in comparative form are replaced with the basic form

Eg ပိုကောင်း ၊ အကောင်းဆုံး

5.2.3 Syllable Segmentation

Myanmar word segmentation is an essential step for natural language processing (NLP) in Myanmar Language because Myanmar text is a string of characters without explicit word boundary delimiters. A Myanmar syllable has a base character

and may also have per-based character, post-base character, above-based character and below-based character. We need the preprocessing steps of Myanmar formal and informal texts. Word segmentation contains two phases syllable segmentation and syllable merging. Segmented syllables are merged into words in segment segmentation. We use word segmentation tool for segment word from the review.

Table 2. Example of Segmentation the reviews

ဟိုတယ်ပါရမီကသန်.ရှင်းသပ်ရပ်ပြီး ရှုခင်းတွေလဲလှတယ်။ (Hotel Parami Yangon is nice, clean and the best for seeing beautiful view, Yangon) ဟိုတယ်_ ပါရမီ_ က_ သန့်ရှင်း_ ပြီး_ ရှုခင်း_ တွေ_ လဲ_ လှတယ်_ ။	positive
သန့်ရှင်းတယ်ဝန်ဆောင်မှုလည်းကောင်းတယ် နော် နောက်တစ်ခါလာတည်းအုံးမယ်။ (clean, good service and I will be back next time) သန့်ရှင်း_ တယ်_ ဝန်ဆောင်မှု_ လည်း_ ကောင်းတယ်_ နောက်တစ်ခါ_ လာ_ တည်း_ အုံး_ မယ်_ ။	positive
အရမ်းကိုဆိုးဝါးတဲ့ ဟိုတယ်ပါ အခန်းတိုင်းမှာလဲ wifi မရဘူး ဝန်ထမ်းတွေလဲအဖအတွေ.အကြိုမရှိ ဝန်ဆောင်မှုလဲ အင်မတန်ညံ့ပါတယ် (So so awful hotel, wifi not available in every room. Staff all are also very bad service and not experience.) အရမ်း_ ကို_ ဆိုးရွားတဲ့_ ဟိုတယ်_ ပါ_ အခန်း_ တိုင်း_ မှာလဲ_ wifi_ မရဘူး_ ဝန်ထမ်း_ တွေ_ လဲ_ အတွေ့အကြုံ_ မရှိ_ ဝန်ဆောင်မှု_ လဲ_ အင်မတန်_ ညံ့_ ပါတယ်_ ။	negative

5.3 Text processing using word2vec

Word vectors are word representations in form of high-dimensional vectors of real numbers. The vectors of words which share a close relationship, because they often appear together in the text corpus, are clustered together in their vector representation as well. This way it is possible to, for example, obtain similar words or synonyms for a word simply by retrieving words with a close vector representation to a given word. After preprocessing step, words are transformed into vector using word2vec tool. It takes the text data as input and produces the word vector as output. The size of corpus can be affected the performance of word2vec.

5.4 Aspect term extraction and polarity classification

The ABSA classification task can be divided into three sub-tasks, namely:

1. Predicting how many aspects a sentence contains
2. Extraction of those aspects
3. Prediction of the sentiment based on each of the sentence's aspects. Since each sentence can potentially contain more than one aspect, the classifier of sub-task 2 had to return a probability distribution of aspects in a sentence sub-task 1 introduce to first define how many aspects the second classifier should use from the values it returns. Sub-task 3 then uses the found aspects to predict what sentiment is applied to each aspect inside a review sentence. Three LSTM networks were used to take care of each of the three classification sub-tasks defined above. The first network takes a review sentence, the pre-trained word vectors to predict how many aspects the sentence contains. The second network (bi-directional LSTM) takes the same data as the first one and returns a prediction of which aspects are most likely contained in the sentence and uses the result from the first network to return the correct number of aspects. The third neural network again takes the same data as the second one; however, an aspect label is also fed into it, to determine the polarity for a specific aspect within the sentence.

Eg. ဒီဆိုင်က မြန်မာမုန့်တွေက အရမ်းစားလို့ကောင်းပါတယ် ပါတယ် ဈေးပိုကြီးတယ်

In this example review we extract the aspect မြန်မာမုန့် (food) and ဈေးနှုန်း (price) and its polarity sentiment words are အရမ်းကောင်း (very good) and ဈေးပိုကြီး (expensive) respectively.

Table 3. Example of aspect and its polarity

Target Entity	Sentiment Word	Sentiment Polarity
ဝန်ဆောင်မှု (service)	အရမ်းညံ့ (very bad)	negative
ဝန်ထမ်း (staff)	ပျူငှာ (be cordial)	positive
အရသာ (food and taste)	တော်တော်ကောင်း (very good)	positive

6. Conclusion

The demand of sentiment analysis is raised due to the requirement of analyzing and structuring hidden information, extracted from social media in the form of unstructured data. Unstructured data helps understanding of customer sentiment through opinion mining. The sentiment analysis is being implementing through deep learning techniques. There was still problem that Bi-LSTM doesn't clearly classify the aspect term with context words on aspect based sentiment analysis. Sentiment analysis requires to extract and analyze hidden information from social media in form of unstructured data. The ongoing research will be described sentiment analysis using deep learning techniques, a hybrid system combing both lexicon based approach and deep learning approach and efficiently cover business insight from unstructured data.

References

[1] Ananchai Muangon1, Sotarat Thammaboosadee, Choochart Haruechaiyasak, "A Lexiconizing Framework of eature-based Opinion Mining in Tourism Industry" ISBN: 978-1-4799-3724-0/14/$31.00 ©2014 IEEE.

[2] Venu Dave, DhwaniShah, DikshiSuthar,Bhagirath Prajapati,Priyanka Puvar "Sentiment Analysis of Tourists Opinions of Amusement, Historical and Pilgrimage Places: A Machine Learning Approach", ISSN: 2231-2803 International Journal of Computer Trends and Technology (IJCTT) – Volume 46 Number 2 – April 2017.

[3] Poria S., Cambria E., & Gelbukh A. (2016), "Aspect extraction for opinion mining with a deep convolutional neural network," Knowledge-Based Systems, 108, 42-49.

[4] Peerapon Vateekul, Thanabhat Koomsubha, "A Study of Sentiment Analysis Using Deep Learning Techniques on Thai Twitter Data", 978-1-5090-2033-1/16/$31.00 ©2016 IEEE.

[5] Win Win Thant, Kiyoaki Shirai "Automatic Acquisition of Opinion Words from Myanmar Facebook Movie Comments".

[6] Jalaj S. Modha, gayatri S. Pandi and Sandip j. Modha, "Automatic Sentiment Analysis for Unstructured Data", International Journal of Advanced Research in Computer Science and Software Engineering.

[7] V. N. Khuc, C. Shivade, R. Ramnath, and J. Ramanathan, "Towards building large-scale distributed systems for Twitter sentiment analysis".

[8] 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.

[9] 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.