Domain-Specific Sentiment Lexicon for Classification

Thet Thet Zin, Kay Thi Yar, Su Su Htay, Khine Khine Htwe,
Nyein Thwet Thwet Aung, Win Win Thant
University of Information Technology, Yangon, Myanmar
thetthetzin@uit.edu.mm, kaythiyar@uit.edu.mm, suhtay@uit.edu.mm, khinekhine@uit.edu.mm
nyeinthwet@uit.edu.mm, winwinthant@uit.edu.mm

Abstract

Nowadays people express their opinions about products, government policies, schemes and programs over social media sites using web or mobile. At the present time, in our country, government changes policies in every sector and people follow with the eyes or the mind on these policies and express their opinion by writing comments on social media especially using Facebook news media pages. Therefore, our research group intends to do sentiment analysis on new articles. Domain-specific sentiment lexicon has played an important role in opinion mining system. Due to the ubiquitous domain diversity and absence of domain-specific prior knowledge, construction of domain-specific lexicon has become a challenging research topic in recent year. In this paper, lexicon construction for sentiment analysis is described. In this work, there are two main steps: (1) pre-processing on raw data comments that are extracted from Facebook news media pages and (2) constructing lexicon for coming classification work. The word correlation and chi-square statistic are applied to construct lexicon as desired. Experimental results on comments datasets demonstrate that proposed approach is suitable for construction the domain-specific lexicon.

Keywords- social media, sentiment analysis, lexicon

1. Introduction

The World Wide Web and the online media provide a forum through which an individual's process of decision making may be influenced by the opinions of others. Online sites such as rottentomatoes.com, allow movie lovers to leave reviews for movies they have seen. Online sites, such as Facebook and blogs, allow users to leave opinions and comments. Other online sites, such as cnn.com and globeandmail.com, allow readers to leave comments. These kinds of online media have resulted in large quantities of textual data containing opinion and facts. Researchers have long measured people's opinion using carefully designed survey questions, which are given to a small number of volunteers. The maturation of social media offers alternative measurement approaches. Social media, now used regularly by more than 1 billion of the world's 7 billion people, contains billions of such communications [1]. Researchers have begun providing these data for a wide range of applications including predicting the stock market and understanding sentiment about products or people. At the present time, Myanmar people express their opinion, feeling and daily activities using Facebook social network by updating status or writing comments on posts using Myanmar language. Therefore, our research group intended to mark people's opinion on important news articles. This is one of the supporting facts for the success of government policies. Sentiment analysis on news articles, especially 21st century Panglong Conference, is ongoing research. In this paper, lexicon construction for sentiment analysis using people comments on news articles is presented.

Feature extraction from data resources is an important step and essential process in sentiment analysis. There are several approaches to extract features from sentences: extraction based on frequency count, n-grams, extraction by exploiting opinion and target relations, extraction using supervised or unsupervised learning, and using topic modeling. According to class information availability in data, there are supervised feature selection approaches [2] as well as unsupervised feature selection approaches. [3] Extracted unigrams, bigrams and combination of unigrams and bigrams is this three different feature vectors applied to different classifiers. In most of the research work, unigram and bigram are highly recommended for tweeter sentiment analysis. In [4] Popescu and Etzioni introduced an unsupervised information extraction system which mines reviews in order to build a model of important product features, the evaluation by reviewers and the relative quality across products. In [5], researchers proposed within class popularity (WCP) feature selection mechanism and then the performance of WCP is then compared with the performance of the most commonly used measures- mutual information, information gain and chi-square. Most of the current approaches study the adaptation or sentiment transfer learning of a trained classifier supervised techniques or lexicon unsupervised techniques from one domain to another which involves having a general lexicon to start with, but very few works actually focus on techniques that build specific domain lexicons without requiring a priori knowledge. Whilst supervised

ISBN 978-99971-0-381-9 © 2017 ICAIT

sentiment classifier performs very well for the domain in which they were trained for, they usually perform very poorly when adapted or transferred to another domain [6].

In this paper, the focus is on feature selection by construction lexicon using n-gram, word correlation and chi-square statistic methods. These feature selection methods are applied to Facebook users' comments on news media pages written in Myanmar language. Myanmar texts are sequence of characters without word boundaries. Therefore, user comments need to be parsed and tokenized into individual words first. In this work, Myanmar syllable segmenter for Unicode Myanmar is used for syllable segmentation. The paper is organized as follows. In the next section data collection process is briefly described. Section 3 contains pre-processing on collected unstructured data. Section 4 provides feature filtering and selecting by applying word correlation and chi-square method. Lexicon Coverage Analysis and Experiments and Results are described in section 5 and 6 respectively. The final section contains a discussion of the obtained results, some remarks and issues that remain to be addressed and that we intend to investigate in future work.

2. Extract Users' comments from Facebook

The internet is a resourceful place with respect to sentiment information. From a user's view, people are able to post their own content through various social media. From a researcher's view, many social media sites release their API, allowing data collection and analysis by researchers and developers. Facebook also releases API for data collection. But it has privacy issues for the researcher. However, news media pages focus on public and not intended to personally. According to my knowledge, news media pages post their articles with public setting. Thus, privacy is not issue for this work. In this paper, the comments are obtained from the Facebook news pages using the Facebook graph application programming interfaces (Graph API) tool on special news article, the Union Peace Conference-21st century Panglong Conference. This conference is very important for our country and people have various opinions on this. First time of conference began on 31 August 2016 and second time began on 24 May 2017. It is planned to the conference in every six month until the agreement is reached and negotiations and political dialogue will be continued. Occasionally, the government holds the panel discussion and revised previous conferences and prepares for the coming conference. People express their opinions on these conferences and events that are related to conferences by writing comments on Facebook news pages. We extract people comments related to peace conferences on news pages. Myanmar people use 10 popular news media pages and Information Committee page according to Myanmar Facebook page statistics by socialbakers.com. Therefore, users' comments are collected from these news media pages using graph API. When data are collected, some of the articles are overlaps on news pages but the users are different. The following figure shows that number of posts related to this specific article and number of extracted comments from these media pages. There are 27337 comments are extracted from 11 news pages. Facebook posts can be as long as 5,000 characters and comments have a maximum of 8,000 characters. Some comments are too long but don't include any opinion words. Some people write just only opinion words. Average length of the comments in collected data is 21 words. The average number of words in collected data is 238532 words and the format of prepared lexicon is "word #polarity".

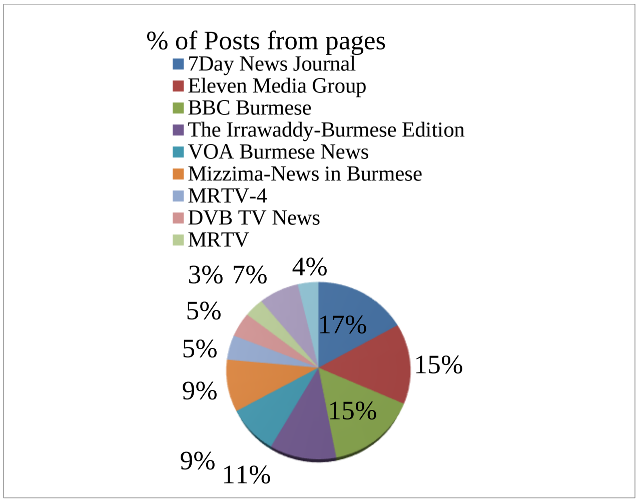


Figure1.Percentage of posts from Facebook pages

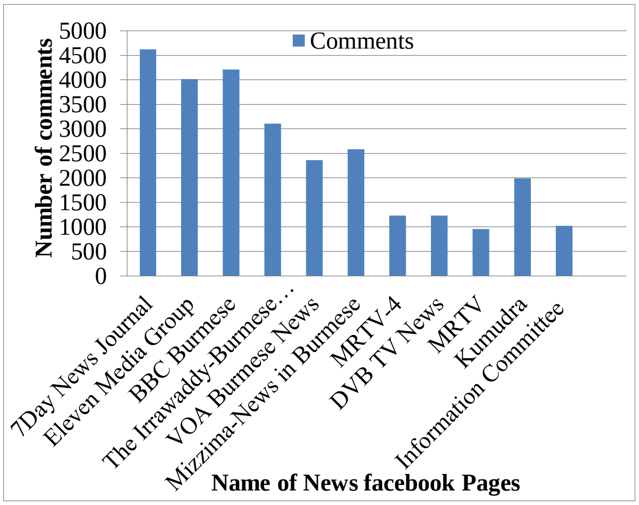


Figure2. Number of comments from news pages

3. Pre-processing

Online data have several flaws that potentially hinder the process of sentiment analysis. The general techniques for data collection from the web are loosely controlled. Therefore, the resultant datasets consist of irrelevant and redundant information. Several pre-processing steps are applied on the available dataset to optimize it for further experimentations. The proposed flow diagram for constructing sentiment lexicon is shown in figure 3.

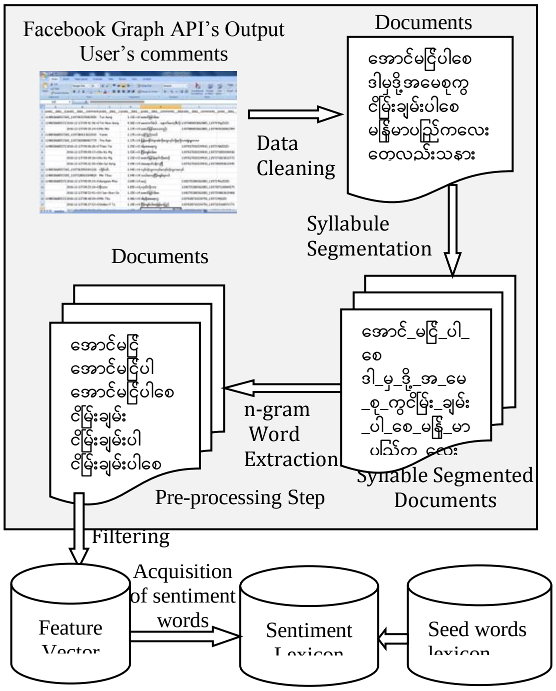


Figure3. Sentiment Lexicon Construction

3.1 Data cleaning process

Data must be preprocessed in order to perform any data mining functionality. Data preprocessing is done to eliminate the incomplete, noisy and inconsistent data. The biggest advantage of looking at words and phrases directly rather than using dictionaries is that one gets much more comprehensive coverage of the language that may be significant. Many of the "words" would not show up in any standard dictionary, either because they are emoticons or variant spellings, or because they are multiword expressions. Abbreviations are often used, orthographic mistakes are made on purpose and hashtags as well as emoticons are present in order to communicate the message of the author in a few words. Further, we find enthusiasm and extroversion are indicated by "အောင်မြင်ပါစေ (be success)" or "ထောက်ခံတယ်!!!! (Okay)" that is simply misspelling words indicates low conscientiousness, or novel spellings signal multicultural backgrounds. New types of words such as emoticons (:) and <3), hashtags("#Bieber"), URLs ("http://www.aapss.org"), photos or stickers comments are presented in social media data. The aforementioned flaws have been somewhat overcome in the following ways:

Step 1: In Myanmar language there are two types of fonts; Unicode based fonts and ad hoc font (Zawgyi). Facebook supports two types of these fonts. But most of the Myanmar people write comments using ad hoc fonts because firstly Facebook supports only this font. This font is not obeying Unicode encoding rules and different to process for other works such as searching, segmentation, etc. Unicode is the standard for character encoding in Myanmar. Thus, all comments extracted from Facebook are changed to Unicode fonts. There are many code conversion tools (Kanaung converter, Myanmar NLP Unicode converter, ThanLwinSoft converter, Burmese Font converter and etc.) for Myanmar language. In this work, reliable Burmese Font converter [15] is used for this purpose. In this font converter website, users give ratings about converter. This converter is 99% convertible now. Moreover, 1000 comments written in ad hoc font change into Unicode font using this converter and we check manually the output results. Converter can change 1000 comments correctly. But it fails in changing in some Myanmar Pali (Buddhism) words. These words are rarely found in people comments.

Step2: Eliminate all words except Myanmar Unicode characters. By doing like this, white spaces between characters, emoji comments, myanglish (written mixing with Myanmar and English words) words, number characters, blank lines, symbol characters and punctuation marks are already eliminated. Orthographic mistakes and multiword expressions still exist in comments dataset.

3.2 Syllable segmentation process

As mention above, Myanmar language does not have boundary word markers. Syllable segmentation is done on cleaning dataset by using syllable segmentation tool. This tool is developed by University of Information Technology (UIT) students. The accuracy of this tool is 100% for our test data and available for research purpose. Test data contains sentences from Myanmar middle school textbook. But if sentences do not match with syllable units; it may be fail in segmentation. However, at the present time, this tool can fully support for our work based on testing. In the later, we will test and compare with other syllable segmentation tools for this work. In this paper, word segmenter is not used. Myanmar language has resource spare problem. Thus, large-scale general purpose lexicons are not publicly available at the present time. For this issue, this paper presented construction of specific domain-specific lexicon.

3.3 Words Extraction process

Syllable segmented dataset is the input for the entity extraction module. The sentence will have some valuable information about its sentiment and the rest of the words will not give any clue regarding the sentiment. Such words should be removed from sentiment lexicon. Syllable segmented dataset is tokenized applying the n gram method by setting the minimum and maximum grams. According to our analysis result on 2000 comments, bigram is set as minimum and 6-gram as maximum gram. As the same time, words extraction process also calculated frequency count for each n-gram word.

4. Filtering process

After pre-processing on the dataset, domain dependent ngram words are already extracted. Many duplicated words and meaningless words also appeared. The more training data, the more duplicated and unnecessary word such as verb suffixes for sentiment analysis are came out. The performance of the lexicon can increase by filtering some unnecessary non-opinion words from ngram lexicon features. There are two steps to perform this purpose.

Step1: From ngram words, the spare words or features are removed based on frequency counts. In order to avoid data sparseness problem frequency counts of the word is less than 5 times; this word is removed from lexicon bag of words list.

Step2: The N-gram words list also contains Myanmar stop words, verb suffixes eg. ("ပါတယ⊞"), conjunction words, preposition words and numbering suffixes words eg. ("တစ်ခု"). These words are also removed from the list. Myanmar stop words list, verb suffixes list and numbering suffixes list are created in my previous research work. Some stop words are manually corrected based on 1000 comments. There are totally 603 stop words for this work.

5. Generating sentiment lexicon

This step is for the acquisition of sentiment words in lexicon. The words in the lexicon have positive or negative polarity. Some of the words may be neutral. This paper only focuses on positive and negative polarity classes. Therefore, it is needed to decide polarity of every word in the lexicon and filtering neutral word from lexicon as much as we can. The main purpose is to increase the system accuracy, and to decrease the computational cost because of the overuse data. The selection process is conducted by selecting every relevant feature that is for the input feature having a correlation to the output from the system. Firstly, seed words lexicon is created from labeled comments. Opinion words are extracted from these comments to create seed lexicons by manually. Positive seed words and negative seed words are manually extracted from positive and negative comment sentences. There are 305 positive words and 201 negative words. There are two steps in building sentiment lexicon.

Step1: words in lexicon are labeled the class (positive or negative) by using seed words lexicon and calculate the probability

$$w_i^C = \log \frac{p(w_i, w_c)}{p(w_i)\,p(w_c)}$$

And then calculate class probability of correction word labeled and unlabeled word. Selected maximum probability of class polarity and label this word with this class. If this combination can determine positive or negative, we label this combination words as maximum probability of polarity class. Repeat this steps again until no more combination words appear. After this process, some words remain as unlabeled words in lexicon.

Step2: Statistic approach is one effective way to do a feature selection process within the data. The word scores of the words are tested based on chi-square method. It also creates a list of all positive and negative words. There are two events, the observed count "O" and the expected count "E". Chi-square measures how much the expected count and observed count derivate from each other. The two events are occurrence of the word/feature and occurrence of the class. When the two events are independent, the observed count is close to the expected count, thus a small chi square score. The higher value of the X^2 score, the more likelihood the feature is correlated with the class, thus it has to select for sentiment lexicon.

$$X^2 = \frac{N(AN - MP)^2}{PM(N-P)(N-M)}$$

A: the total number of positive words that contain feature X

M: the total number of words that contain feature X

N: the total number of words

P: the total number of positive words

After applying step1 and step2, unlabeled remaining words are assumed as neutral and removed from lexicon.

6. Lexicon Coverage Analysis

Sentiment lexicon represents how a word is distributed among the set of all opinion words. If the classification result is not optimally distributed across the space of unique words, it might be better to greedily increase the word coverage from the perspective of the sentiment lexicon extraction. To approach the initial dataset problem from the lexicon coverage view, we test lexicon coverage analysis for finding maximum lexicon size for classification. Firstly, we put longest 1000 sentences in the system and extracted opinion words. We iteratively add the next 1000 sentence that has the minimum cosine similarity between the words that have been covered. According to analysis, the sentiment lexicon for news domain is stable (no new opinion words appeared) over 8000 training sentences.

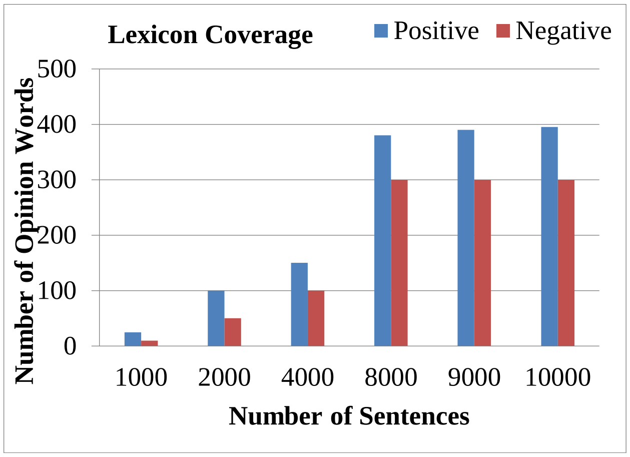
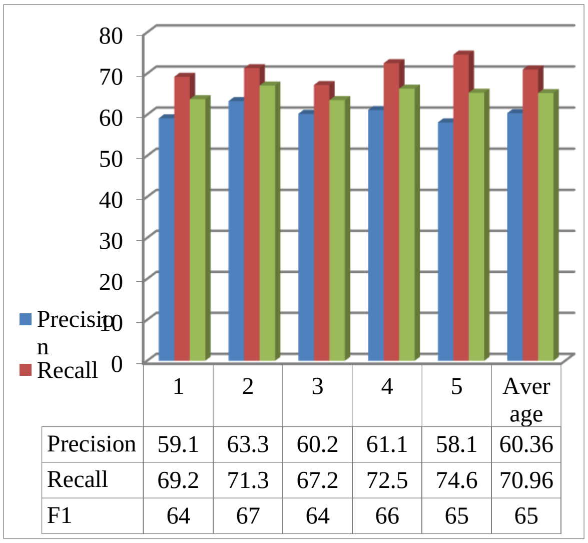


Figure4. Lexicon Coverage Analysis Results

7. Experimental Results

Above section we analyzed coverage of lexicon. We test how effect this sentiment lexicon on news domain in sentiment labeling task. To evaluate the accuracy, we followed a 5-fold cross validation process on 17337 sentences: each dataset was randomly split into five different non overlapping training and test sets. Each fold has 3467 sentences. There are 27337 comment sentences. In this work, 10000 sentences are already used in sentiment lexicon construction process. Therefore, remaining 17337 sentences are used to test created lexicon. Precision measures how many sentiment words are correctly identified in their classes. Recall expresses how many words, in the whole set, have been correctly recognized: a low recall means that many relevant comments are left unidentified. We assumed that a word not found in the lexicons has a neutral polarity. Five-fold cross validation results precision is 60.4%. Recall on testing dataset is 71%. Five folds cross validation results are shown in Figure 5.

Error Analysis is carrying on every fold results. People have complex ways of expressing opinions. We manually performed error analysis on dataset. Error analysis revealed that most of the errors are revealed to neutral words. Some of the positive polarity words come out as neutral. This fact leads to decrease performance of the lexicon. There are two type of strong errors are recognized. First error is negative prefix words. Some of the positive words have negative polarity with negative prefix of Myanmar language. But the system cannot label such words correctly. Another is spelling missing, spelling mistake and dialect. The system cannot determine their classes. But they have respective polarity class.



	1	2	3	4	5	Aver age
Precision	59.1	63.3	60.2	61.1	58.1	60.36
Recall	69.2	71.3	67.2	72.5	74.6	70.96
F1	64	67	64	66	65	65

Figure5. Cross Validation Result

8. Conclusion and Future Work

In Myanmar, the intensive use of Internet, especially social media, to express opinion or view on certain matter, marks the opportunity to further develop the research in this field. In a sentiment analysis task choosing lexicons from the appropriate domain is important. There is a need for a method which can create domain-specific lexicons because there are no lexicons for every domain and creating them manually is expensive and requires an expert in that domain. The input of these methods is a small seed lexicon (semi-automatically) and unlabeled domain-specific texts. The best results were given by the manually assembled lexicon. But they are much more expensive and it is hard to create one for all domains, thus automatic methods are needed. The result shows that the proposed method is useful for increasing the performance of sentiment analysis systems in all domains.

As to future work, we intend to combine unsupervised feature selection method for lexicon expansion. Moreover, another work is needed on improving the accuracy of the sentiment classification on huge amount of information dataset.

9. References

[1] H. Andrew Schwartz and Lyle H. Ungar, “Data-Driven Content Analysis of Social Media: A Systematic

Overview of Automated Methods”, ANNALS, AAPSS, 659, May 2015.

[2] J. Novakovic, P. Strbac, and D. Bulatovic, “Toward Optimal Feature Selection using Ranking Methods and Classification Algorithms”, Yugoslav Journal of Operations Research, DOI: 10.2298/YJOR1101119N, 21(2011), Number 1, pp.119-135.

[3] A. Go, R. Bhayani, and L. Huang, “Twitter sentiment classification using distant supervision,” CS224N Project Report, Stanford, vol. 1, 2009.

[4] Ana-Maria Popesu and Oren Etzoni, “Extracting Product Features and Opinion from Reviews”, Proceeding of Human Language Technology Conference and Conference on Empirical Methods in Natural Language, ACL, Vancouver, October 2005, pp. 336-339.

[5] S. R. Singh, H. A. Murthy and T. A. Gonsalves, “Feature Selection for Text Classification Based on Gini Coefficient of Inequality”, JMLR: Workshop and Conference Proceedings 10, the Fourth Workshop on Feature Selection in Data Mining, pp.76-85, 2010.

[6] Songbo Tan, Xueqi Cheng, Yuefen Wang, and HongboXu, “Adapting naive bayes to domain adaptation for sentiment analysis”, In European Conference on Information Retrieval. Springer, 2009, pp. 337–349.

[7]V. Jijkoun, M.de Rijke, and W. Weerkamp, “Generating Focused Topic-specific Sentiment Lexicons”, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Uppsala, Sweden, 11-16 July 2010, pp. 585-594.

[8] N. Kurian and S. Asokan, “Summarizing User Opinions: A Method for Labeled-Data Scarce Product Domains”, International Conference on Information and Communication Technologies (ICICT 2014), ScienceDirect, 2015, pp. 93-100.

[9] N. Kaji and M. Kitsuregawa, “Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents”, Proceeding of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics, Prague, June 2007, pp. 1075-1083.

[10] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu and B. Liu, “Combining Lexicon-based and Learning –based Methods for Twitter Sentiment Analysis”, HP Laboratories, 2011.

[11] F. Del Vigna, A. Cimino, F. Dell’Orletta, M. Petrocchi, and M. Tesconi, “Hate me, hate me not: Hate speech detection on Facebook”, 2015.

[12]V. Hangya, “Automatic Construction of Domain Specific Sentiment Lexicons for Hungarian”, 2013.

[13] A. Putra N and H. Sujaini, “Analysis of Extended Word Similarity Clustering based Algorithm on Cognate Language”, International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol.4 Issue 11, November 2015.

[14] K. Labille, S. Gauch and S. Alfarhood, “Creating Domain-Specific Sentiment Lexicons via Text Mining”, Proceedings of workshop on issues of sentiment discovery and opinion mining, Halifax, Canada, August 2017.

[15]http://burglish.mymm.org/latest/trunk/web/fontconv.htm