Tamil English Language Sentiment Analysis System

R. Thilagavathi
Computer Science and Engineering
A.V.C. College of Engineering
Mannampandal, Mayiladuthurai

Mrs. K. Krishnakumari, M.E., (Ph.D.,)
Computer Science and Engineering
A.V.C.College of Engineering
Mannampandal, Mayiladuthurai

***Abstract*— Sentiment analysis is a hard problem, while multilingual sentiment analysis is even harder due to the different expression styles in different languages. Although many methods for multilingual sentiment analysis have been developed in the open literature, most of them suffer from two major problems. The first is their excessive dependence on external tools or resources (e.g., machine translation systems or bilingual dictionaries), which may not be readily obtained, especially for minority languages; The second is conflictive sentiments, i.e., the sentiment polarity of some parts of a text is inconsistent with its overall sentiment polarity. It is observed that in a product or service review there usually exist a few sentences which play a more important role in determining its sentiment polarity, as compared to others. Therefore, differentiating key sentences from trivial ones may be helpful to improve sentiment analysis. Inspired by this observation in this paper we propose a novel framework to estimate the sentiment polarity of reviews by virtue of opinion lexica and key sentences automatically extracted from unlabeled data. This framework cannot only overcome the problem of excessive dependence on external resources, but also is able to capture the overall sentiment polarity of reviews. Experimental results on realistic review datasets demonstrate that the proposed framework is effective and competitive with the representative baselines.**

Keywords*— Bilingual dictionary; opinion mining; conflictive)

I. INTRODUCTION

Multilingual Sentiment analysis aims to automatically identify the sentiment polarity of given texts in multi languages, which has broad applications, including recommendation systems, sentiment summarization, opinion retrieval, and so on. Given the explosively growing number of online reviews in different languages, multilingual sentiment analysis has recently attracted a great deal of attention from both academia and industries. According to the resources employed, existing methods for multilingual sentiment analysis can basically be categorized into two types, namely, machine-translation-based methods and bilingual-dictionary-based methods. In this project we are using Tamil language reviews this reviews are translated into English language by using Google translate and then these translated reviews are classification, clustering by using some algorithms and techniques.

According to the training mode, the existing methods for sentiment analysis can be roughly categorized into two types, namely, supervised methods and unsupervised methods. Supervised methods usually regard polarity identification as a classification task and use a labeled corpus to train a sentiment classifier. conducted polarity classification of reviews using, Naive Bayes demonstrated that using large feature vectors in combination with feature reduction, high accuracy can be achieved in even very noisy data of customer feedback. Used neutral reviews to help improve the classification of positive and negative reviews. Unsupervised methods usually conduct sentiment classification using a sentiment lexicon. Then finally compare the both supervised and unsupervised method's strengths.

After the sentiment analysis product aspect ranking is also done here. Numerous consumer reviews of products are now available on the Internet. Consumer reviews contain rich and valuable knowledge for both firms and users. However, the reviews are often disorganized, leading to difficulties in information navigation and knowledge acquisition. This article proposes a product aspect ranking framework, which automatically identifies the important aspects of products from online consumer reviews, aiming at improving the usability of the numerous reviews. The important product aspects are identified based on two observations: 1) the important aspects are usually commented on by a large number of consumers and 2) consumer opinions on the important aspects greatly influence their overall opinions on the product. In particular, given the consumer reviews of a product, we first identify product aspects by a shallow dependency parser and determine consumer opinions on these aspects via a sentiment classifier. We then develop a probabilistic aspect ranking algorithm to infer the importance of aspects by simultaneously considering aspect frequency and the influence of consumer opinions given to each aspect over their overall opinions.

II. PROPOSED SYSTEM

Propose a novel multilingual sentiment analysis framework. In the proposed framework, no manually labelled corpus is needed and all extracted information is domain-dependent. In general, the contributions of this study can be summarized as follows:

- Propose a statistical method for opinion lexicon extraction based on a few seed words, which can be easily transplanted to almost any language and does not need to refer to synonyms and antonyms dictionaries;

- On the basis of the extracted opinion lexicon, propose a key sentence extraction method for capturing the overall opinion of reviews, which solves the problem of conflictive sentiments;
- Propose a Self-Supervised Learning (SSL) method for sentiment classification, which combines unsupervised and supervised techniques together by virtue of the above extracted opinion lexicon and key sentences;
- Finally, extensive experiments on multilingual datasets in different domains well demonstrate the effectiveness of the proposed methods.

ADVANTAGES

- Cost-efficient framework for multilingual sentiment analysis.
- Improve the performance of sentiment classification.
- The concept of key sentences is reasonable and key sentences are useful for dealing with conflictive sentiments.
- The key sentence extraction method is effective and language-independent Abbreviations and Acronyms

III. SYSTEM ARCHITECTURE

The unlabeled Tamil dataset are first translated by using the Google translator. Opinion words are first extracted from unlabeled data based on only a few seed words. Next, the sentiment polarities of all extracted opinion words are initialized using a K-Nearest-Neighbor (KNN) method and further refined by a voting mechanism among multiple domains An opinion lexicon is a list of opinion words, such as *good* and *poor*, along with sentiment polarities.

Next, we need to identify the sentiment polarity of each extracted opinion word. In this project, the process follows two steps. First, for a given domain, an initial polarity is assigned to each opinion word according to the polarity information of its K Nearest Neighbors (KNN). Since only a few seed words are used, we apply an iterative KNN strategy to propagate their polarity to all the other words.

As we mentioned before, one challenge for document-level sentiment analysis is that not every part of a review is equally informative for inferring its sentiment polarity. Generally, the polarity of a review mainly depends on the overall evaluation of the reviewer rather than the details about some specific aspects. Therefore, for a given review, we hope to extract the most important sentences that express the overall sentiment. First of all, from the position perspective, we have observed that key sentences are usually located in the beginning or the end of a review; Second, from the content perspective, key sentences often have stronger sentiment polarity than trivial sentences; Finally, from the representation style perspective, key sentences may contain some conclusive words or phrases, such as, *"overall"* and *"in general"*. or attitude of the reviewer. supervised methods usually regard sentiment polarity identification as a classification problem and use a labelled corpus to train a sentiment classifier. In order to reduce the burden of manually labelling data, we propose a Self-Supervised Learning (SSL) method for sentiment classification by combining the strengths of both unsupervised and supervised techniques. In more detail, an unsupervised technique is first employed to label a portion of reviews from the given corpus, where the key sentences of all reviews have been clearly marked; Next, some informative sample reviews with clear sentiment polarities are selected, based on which we train a supervised classifier for multilingual sentiment classification.

In this SSL method, informative sample reviews are selected according to the clarity of their sentiment polarities using the extracted opinion lexicon. It is obvious that, for a review *d*, the larger the difference between the number of positive words, denoted as *TP* (*d*), and the number of negative words, denoted as *TN*(*d*), the more confident we are about its sentiment polarity. However, the difference between *TP* (*d*) and *TN*(*d*) is significantly affected by the length of the review. It is often the case that the longer the review, the larger its *TP* (*d*) or *TN*(*d*), and thus the larger the difference between its *TP* (*d*) and *TN*(*d*). Therefore, we adopt the normalized absolute difference between *TP* (*d*) or *TN*(*d*) as the criterion function for selecting sample reviews in order to avoid the effect of the length of different reviews.

$$\text{InfRev(d)} = \frac{|TN(d) - TP(d)|}{TP(d) + TN(d)} \quad (1)$$

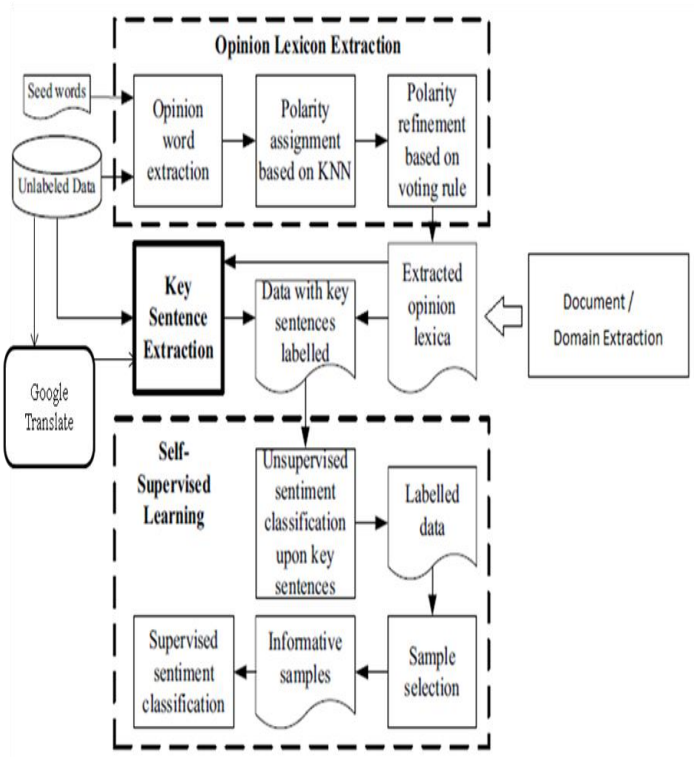

Fig.1. System Architecture

IV. MODULES DESCRIPTION

We propose framework for sentiment analysis that consists of five basic modules,

1. Google Translate
2. Opinion Lexicon Extraction Module
3. Key Sentence Extraction Module

4. Supervised Learning Module
5. Result Analysis Module

The project is primarily concerned with extracts the opinions from review data set (available online) for the Hotel Domain. The project uses a larger collection of reviews for both training and testing data. The review data sets are stored in a format of text file.

1. GOOGLE TRANSLATE

In this module we translate the tamil comments into english by using google translate. all comments are translated into English.

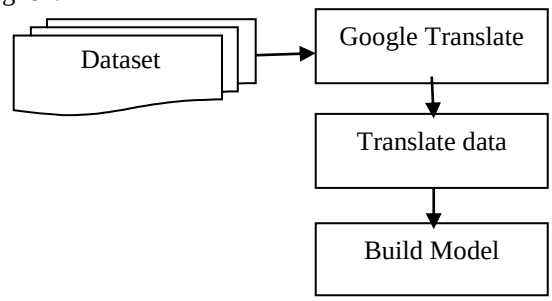

Fig 2. Google translate

2. OPINION LEXICON EXTRACTION MODULE

An opinion lexicon is a list of opinion words, such as good and poor, along with sentiment polarities. Traditional methods for extracting opinion lexica rely on language-dependent NLP tools (e.g., POS tagger), external resources (e.g., Word Net), or labeled data. In this module, to will extract opinion lexica from unlabelled data using heuristic information among different languages. An initial polarity is assigned to each opinion word according to the polarity information of its K Nearest Neighbors (KNN).

3. KEY SENTENCE EXTRACTION MODULE

To extract the most important sentences that express the overall sentiment or attitude of the reviewer. It should be pointed out that key sentence extraction in this module is different from sentiment summarization in two aspects: First, sentiment summarization aims to generate a summary that well represents the overall sentiment of a set of documents, while key sentence extraction in this paper intends to select sentences from a single document that well express the overall sentiment polarity of the author; Second, a sentiment summary is subject to predefined length constraints. But, there is no length constraint for key sentence extraction.

Key sentences may contain some conclusive words or phrases, such as, *"overall"* and *"in general"*. Consequently, in the key sentence extraction algorithm, to carefully take these features into account by designing three feature functions, respectively. The overall score of any sentence is the sum of the values calculated based on the three feature functions. to present the three feature functions.

- Position Feature Function
- Content Feature Function
- Representation Style Feature Function

3.1 POSITION FEATURE FUNCTION

It is observed that a sentence at the beginning or the end of a review is more likely to be a key one than those in the middle. Therefore, the position feature function should reward sentences at both ends of the document. Intuitively, the curve of a Gaussian probability density function is bell-shaped, and its negative form may fit with the characteristic of the position function.

$$\phi 1(\sigma) = -\frac{1}{2\pi q} e^{-\frac{(s-\mu)^2}{2\partial^2}} \quad (1\ \delta\ \sigma\ \delta\ \lambda\varepsilon\nu) \qquad (2)$$

Where, μ is the mean (location of the peak),
σ is the standard deviation, *len* is the length (that is the number of sentences) of a review.

3.2 CONTENT FEATURE FUNCTION

As key sentences should have strong and clear sentiment polarity, the content feature function is defined as

$$\phi 2(\sigma) = \frac{\sum_{t\varepsilon S} opinion\ lexicon(t)}{\sum_{t\varepsilon S} |opinion\ lexicon(t)|} \qquad (3)$$

where *opinion lexicon*(*t*) denotes the sentiment polarity of word *t* in sentence *s*,
if it is an opinion one. Specifically,
if *t* is a positive word, *opinion lexicon*(*t*) = 1;
if *t* is a negative word, *opinion lexicon*(*t*) = −1;
otherwise, *opinion lexicon*(*t*) = 0.

3.3 REPRESENTATION STYLE FEATURE FUNCTION

conclusive expressions(t) denotes if t is a conclusive snippet in a sentence. Here, a snippet is a single word or a string of words that is split by punctuations in a sentence. For example, the sentence, "Overall, I love this film!" has two snippets, "overall" and "I love this film

3.4. SUPERVISED LEARNING MODULE

Supervised methods usually regard sentiment polarity identification as a classification problem and use a labelled corpus to train a sentiment classifier. In order to reduce the burden of manually labelling data, a Self-Supervised Learning (SSL) method for sentiment classification by combining the strengthes of both unsupervised and supervised techniques. An unsupervised technique is first employed to label a portion of reviews from the given corpus, where the key sentences of all reviews have been clearly marked;

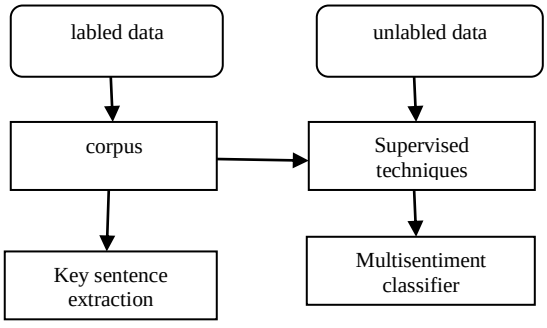

Fig 3 Supervised Learning Module

4. RESULT ANALYSIS MODULE

In this module use the multilingual sentiment corpora, including English, Tamil, and any one Language. In order to highlight the domain-specific nature of opinion words, to collect reviews not only from different languages, but also from different domains, namely, books, DVD, and music. The multilingual sentiment a Naive Bayesian classifier for supervised learning and select top 30 % informatively predicted reviews for training, with which the classifier exhibits the best performance.

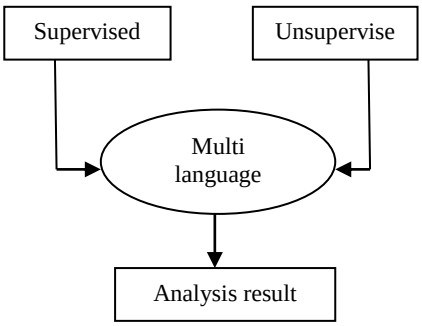

Fig .4 Result analysis module

5.1 KNN Algorithm

K-Nearest Neighbors (KNN) classification divides data into a test set and a training set. For each row of the test set, the K nearest (in Euclidean distance) training set objects are found, and the classification is determined by majority vote with ties broken at random. If there are ties for the Kth nearest vector, all candidates are included in the vote.

5.2 KNN steps

Step 1: Determine parameter K = number of nearest neighbors
Step 2: Calculate the distance between query- instance and all the training samples.
Step 3: Sort the distance and determine the nearest neighbors based on the k-th minimum distance.
Step 4: Gather the category Y of the nearest neighbors.
Step 5: Use simple majority of the category of nearest neighbors as the prediction value of the query instance

5.3 Naïve Bayes Classifier

A Naive bayes classifier is a simple probabilistic model based on the Bayes rule along with a strong independence assumption. The Naïve Bayes model involves a simplifying conditional independence assumption. That is given a class (positive or negative), the words are conditionally independent of each other. This assumption does not affect the accuracy in text classification by much but makes really fast classification algorithms applicable for the problem. In our case, the maximum likelihood probability of a word belonging to a particular class is given by the expression:

$$P(x_i|\, c) = \frac{Count\ of\ x_i\ in\ documents\ of\ class\ c}{Total\ no\ of\ words\ in\ documents\ of\ class\ c}$$

The frequency counts of the words are stored in hash tables during the training phase. According to the Bayes Rule, the probability of a particular document belonging to a class is given by

$$P(c_i|d) = \frac{P(d\,|c_i) * P(c_i)}{P(d)}$$

The classifier outputs the class with the maximum posterior probability. It also removes duplicate words from the document, they don't add any additional information; this type of naïve bayes algorithm is called Bernoulli Naïve Bayes. Including just the presence of a word instead of the count has been found to improve performance marginally, when there is a large number of training examples.

6. ACCURACY DETECTION

Accuracy for this knowledge base can be detected using the Confusion Matrix. Confusion matrix is a tool for analyzing how well classifier can recognize tuples of different classes. Using Confusion matrix predict the accuracy. A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix.

		Predicted Class: Yes	Predicted Class: No
Actual Class	Yes	TP	FN
	No	FP	TN

Accuracy detects the percentage of predictions that are correct using formula

$$\text{Accuracy Detection} = \frac{TP+TN}{(TP+TN+FP+FN)} \qquad (4)$$

Precision is the percentage of positive predictions that are correct.

$$\text{Precision} = \frac{TP}{(TP+FP)} \qquad (5)$$

Recall is the percentage of positive labeled instances that were predicted as positive.

$$\text{Recall} = \frac{TP}{(TP+FN)} \qquad (6)$$

REFERENCES

[1] Esuli and F. Sebastiani. Pageranking wordnet synsets: An application to opinion mining. 45(1):424, 2007.
[2] Hassan and D. Radev. Identifying text polarity using random walks. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 395–403, 2010.
[3] Liu and L. Zhang. A survey of opinion mining and sentiment analysis. In *Mining Text Data*, pages 415–463. Springer, 2012.

[4] Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of* the ACL-02 conference on Empirical methods in natural language -*gprocessing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[5] Banea, R. Mihalcea, and J. Wiebe. Multilingual subjectivity: are more languages better? In Proceedings of the 23rd International Conference on Computational Linguistics, pages 28–36, 2010.