Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 7, July -2017

Sentiment Analysis Approach for Resource-Scarce Languages

Kanika Garg[1],Goonjan Jain[2]

[1]School of Computer & Systems Sciences, JNU, Delhi
[2]Department of Applied Mathematics, DTU, Delhi

ABSTRACT:- *Sentiment Analysis (also known as Opinion Mining) is taking an important role in the modern Web Services. This technique is derived from the most popular language processing known as "Natural Language Processing". Largenumbers of researches are done which provided tools and techniques to extract features from text.The main focus since was on English language. Several techniques regarding sentiment analysis like different classification techniques and machine learning techniques are discussed in this paper in detail.Previous work done in this field provided methods and experimental analysis of reviews mentioned in this paper. This paper is mainly emphasized on less popular languages like Chinese, Urdu, French, and German.*

Keywords: *Natural Language Processing, Sentiment Analysis, Lexicon, Corpora*

1. INTRODUCTION

In Modern world of Technology, time is considered as main factor for people. With the increasing of web information, people are more dependent of web services that provide relevant information inlesser time. Due toincreasing in Internet demand, analysis in web technology is required to producebetter results and accurate. Some of the tools and their methods are already defined by some authors. The Internet or Web Services are now become a business instead of only purchasing. People have become producers also instead of buyers. They can promote their business online. Reviews are one of the major factors to use the web services. As, several people fetch the knowledge about any products and their characteristics by reading their reviews. These reviews can be written by experienced people. So, this technique of providing information about the customer reviews and deciding feeling and emotions regarding the product provides the new concept known as"**Sentiment Analysis**".

Sentiment analysis is a sub-field of Natural Language Processing. It deals with processing of text and provides that text in human readable form. NLP finds its applications in many areas as inK. Garg[1], discussed about importance and basic concept of NLP in education.In their work they discussed about the classical techniques of NLP that can improve education system. In [2] they discussed about its importance in business strategic planning.

Sentiment Theory is also known as Opinion Mining[3]. This technique use reviews based system which tells about positive and negative thoughts for anyproducts. Sentiments analysis consists of classification technique in order to achieve extraction of text. Some levels can also be defined in this analysis like sentence level, document level and aspect level. The opining mining helps to provide opinions for different people from their personal experiences. Several techniques regarding sentiment analysis like classification, lexical analysis, independent language extraction is discussed in this paper. The analysis for several languages other than English like Arabic, Urdu, French, Chinese, etc. is mentioned below.

This paper is main focus on some sentiment analysis techniques. This paper provides the information and work done by several authors in various fields and languages of sentiment analysis. Classification technique and independent language analysis are discussed in this paper. Classification method used machine learning approach that solves the complex problem using the automation system. Types of learning with their advantages and dis-advantages are discussed with their sub techniques.

2. RELATED WORK

In the sentiment analysis, the main focus is on reviews. It deals with feelings and opinions towards the products. Due to increase in technology, the social networking sites have become a part of people for communication. These social sites and blogging provide a user to get mined and accurate information [4]. The knowledge sharing is very specific and accurate several times. Some authors provides methods for sentiment analysis like support vector, machine acceptable,etc. and some discuss about the independent languages or methodologies and tools used in sentiment analysis.A.Guizzardi [5] provides the technique for Classification which can help in tourism information for real time forecasting in 2015.They provided the Business sentimental indicator which improves the real time processing and goodness of fit as well. They calculated the sentiments data and features using the sentimental techniques. In [6] T.Wilson performed phrase-level sentiment analysis. Contextual polarity has been calculated on phrase-level.Online tools and API's for sentiment analysis are analysed in [7]. Amazon reviews are taken in product domain to analyse these tools. Several free web tools that were discussed were Opendover, Sentistrength, Lexinalytics, and Sentimetrix.

3. SENTIMENT ANALYSIS

The word sentiment is related with feelings. It is a combination of Opinions, Emotions and Attitude. It is based on subjective impressions rather than facts. It may consist of opposition binary opinions like for/against, like/dislike, positive/negative, etc [8]. It is a process to extract or identify characteristics or features of a piece of text using some technologies like Natural Language Processing, Artificial Intelligence, and Machine Learning. It is also known by other name as "Opining Mining" as the main focus is to extract or analyze the important features. Some of the techniques used to analyze the sentiments are discussed below with their sub categories.

4. SENTIMENT CLASSIFICATION TECHNIQUES

This is one of the techniques to extract features in sentiment analysis. This classification technique helps to classify the features or set of data into groups using machine learning.Broadly, this classification technique can be dividedinto three categories known as lexicon based approach, machine learning and hybrid technique [9]. Let us discuss each approach with their methods.

4.1 MACHINE LEARNING APPROACH

This technique uses linguistic concept and famous algorithms of ML. They are based on ML that helps in solving text classification problems in sentiment analysis. The problem of text was overcome by linguistic and syntactic features.

4.1.1 Text Classification Problem-

Let us consider the set of records in which every record can be considered as a label for class. Records can be represented as:

T={A1, A2 …An}

The classification models have features which help to underlying record for class labels. For any unknown class at a given instance, prediction for class label is done using this model. There can be possibility of a "hard classification or soft classification". Soft classification means when any approx value at any instance is assigned for a label. Hard classification refers to problems in which at a given instance, only one label is assigned to it. Going further for deep study, we can classify the machine learning approach as a supervised or unsupervised based learning approach.

4.1.2 Supervised Learning

In this technique, set of input and outputs are known and no new class labels are formed. This uses formal approach to solve the problems and samples are trained according to the requirements. This learning provide the accurate output and samples can be given as a training using the machine learning algorithm. This learning provides the better result as compared to unsupervised learning.Some approaches of the machine based supervised learning are discussed below.

4.1.2.1 Probabilistic classifiers:

This is a mixture of classification technique. It means each class use mix models for analysis. Each component of mixture uses general model which provides to find sample for any particular component in class. These classifiers can also be refers as generative classifiers. Some methods of probabilistic approach to solve the supervised learning problems are given below.

4.1.2.2 Naïve Bayer's Classifiers

This is most common and simple method to solve probabilistic problems. It is based on theory of mathematical probability which uses posterior probability method. The probability can be computed on basis of words distributed in any document. Features set can be extracted using this theorem to classify the set of labels for a particular class. This technique requires linear number of variables as a perimeter and scalability is very high[10]. Mathematical equation of this classifier in terms of conditional probability can be represented as:

$$P(\text{outcome}|\text{evidence}) = \frac{P(\text{Likelihood of Evidence}) * \text{Prior probability of outcome}}{P(\text{Evidence})}$$

4.1.2.3 Bayesian Network

This is based on fully dependent approach for all the features. This concept is opposite to Naïve Based which uses feature independent. The use of dependent features in this classifier gives rise to a new concept known Directed Graph Acyclic that consists of nodes and edges to interlink all features with each other. This technique is very expensive in order to extraction of text, so it is not commonly use. This network was introduced to represent relation between diseases and their symptoms. They are basically DAG whose edges are used to find dependencies and act as a linkage between other nodes and nodes represent variables which can be used as a hypothesis in order to find solution. There also exists an advance version of BN known as multi-dimensional Bayesian network[11]which includes relation of different target variables within same classification problems. They were used in semi-supervised technique in order to find classification in sentiment analysis.

4.1.2.4 Neural Network

Neural network is a part of Artificial Intelligence which is based on the artificial intelligence concept. Neural is a smallest unit and several neurons are combined to form a network to show relation between each other. This process is known as Neural Network. It consists of basically three parameters. Firstly, input parameter is defined using some vectors which are used to perform several operations. Secondly, Weight parameter is added toinputs which use some

activation functions and compute mathematical operation in order to produce output. Third, output parameter which gives result and relation between the inputs.In Linear combination, Neural Network can be represented as:
Pi=A.Xi
There is also exists multi layer Neuron Network in which a hidden layer is associated in process. Multi layer network consists of more than one layer in processing stage in which an input is passed to several processes in order to produce outputs.Formally, all neurons are classifies in several layers and each layer uses some transformation methods in order to extract outputs on the basis of given inputs. Artificial Neural Network is mainly used to work as human brain does. It was implemented to make machine as intelligence as human. Neural Networks has very wide applications like Social Networking, Artificial Learning, Machine Learning, Voice Recognition, Pattern Recognition, Face Detection, etc[12].We can represent the linear neural network as the following form.

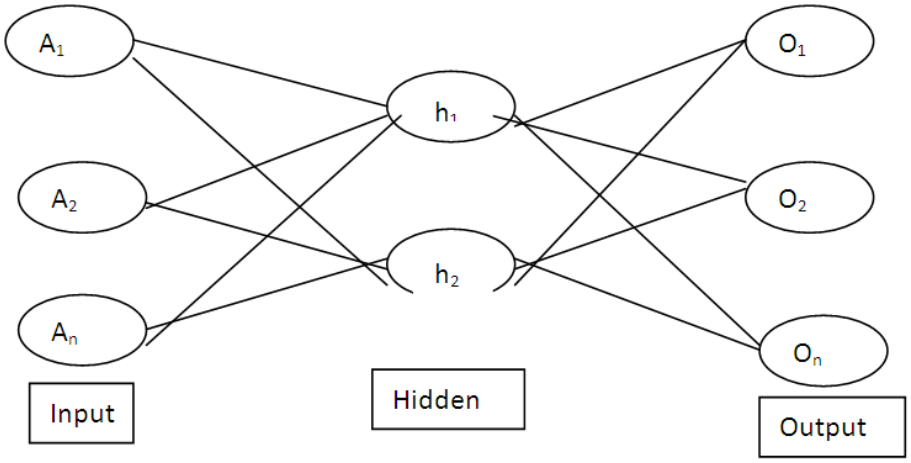

Figure1. Single Layer neural Network

4.1.3 Unsupervised Learning
This technique was introduced by the Guo and Xianghua[13] to discuss about social reviews in Chinese. This technique is just opposite to the supervised learning. In this technique, the training samples are not known and new output class labels can be defined in output set. Output cannot be predictable and less accurate as compared to supervised learning. This technique used approximation technique.
In text classification problems, categories need to be defined which requires a large training samples labelled in supervised learning. Sometimes, task for labelling becomes very difficult and complex. In order to solve those problems, this technique was introduced. This technique helps to classify the big document into several sentences in which each sentence uses list of keywords in order to present same sentence category.This approach provides better outcomes and better partitioning of classes.

4.2 LEXICON APPROACH [CORPA FORMATION]

This is another technique of classification helps in sentimentanalysis. It is based on inbuilt term known as **Lexicon** which is defined as collections of pre-defined and known components in sentiments. The main work is to find opinion lexicon which helps in analyzing text in classification process.
Opinion is one of major component that is used in sentiment techniques. Opinions are based on the practical used or implementation and can varies depend on the personal experience. Like good opinion represent the positivity of product and bad or dislike opinion refers to negativity of product. Opinion lexicon is combines to male opinion phrases and idioms. There are three approaches[14]in lexicon technique to get list of collection words for compilation process.
4.2.1 Manual Approach:
As name suggests, this approach requires man power in order to process any work. This type of technique is very time consuming and does not provide accurate results as well. It also requires large number of workers to operate and processing which leads to wastage of human resources as well.
4.2.2 Dictionary Based Approach
In this approach, orientation is considered as a major component. Small amount of words are collected using manual task with orientation. The set is gradual increase by searching in popularly known Thesaurus or Corpora Word net is used to find antonyms and synonyms for list of collected words. Next iteration begins after adding new words which are found in seed list.Iteration carried out continuously until no words are found in list. Correction and errors can be found manually at end by completion of iteration. This technique does not provide solutions for context specific and domain specific orientations.

4.2.3 Corpus Based Approach

This approach was mainly used to remove limitations of dictionary based approach. It means this approach helps to search for context specific orientation words. This approach is used pattern recognitionthat provides list of opinion words which occur together in a large set of corpus. This method as proposed by McKeown and Hatzivasilogolu[15]with a seed of opinion adjectives. They used these adjectives in linguistic constraints to find additional constraints with orientation. They provide connected constraints which are used to connect two or more words with same orientation. Some connectives are AND, OR BUT, EITHER, etc. This is known as sentiment consistency which is not practically advisable. Some opinion words for reactions are also rises like BUT. It is not much effective as compared to dictionary based approach as it requires large skills and time to prepare list of English words from the set of large corpus. It is only used in order to find domain and specific orientation. This approach used statistical and sentiment approach in order to achieve orientation of words.

5. MONOLINGUAL SUBJECTIVITY AND SENTIMENTAL METHODS

5.1 CHINESE

Zhang et al.[16] was pioneer of the sentiment analysis in Chinese language. This system used the methodology of rule-based system. It was introduces in Chinese articles without any cost for multiple domains. They used system of syntactic structure and sentimental lexicon which is discussed above. The process for this system can be divided into two steps:

- Computing the sentence sentiments.
- Combine these sentiments or aggregate them to find or obtain score for sentiment document.

Document sentiment can be represented in mathematical equation as:

$$R_D = \sum_{i}^{n} p(R_i) * A_i$$

Where D represents document of sentiment
Ai refers to sentence importance and
P(Ri) tells about polarity of sentence.

In above process, subjective sentences are only considered and objective sentences are eliminated. Polarity can be computated by researchers which dependsupon modified polarity in sentences of the words. Basically, polarity in sentence for words can be categorized into three types.

- Prior polarity refers to provide general polarity of words in sentiment.
- Dynamic polarity is used to tell about environment or contextual priority. This type of polarity is depend on domain and topic.
- Modified polarity provides modified words which are update by environment or surroundings such as degree adverbs.

Hownet [17], an Chinese English Lexicon, tells about Chinese subjective words dictionary which consists of 3116 negative words, 3730 were positive words, 1254 were affected negative words (like sad), 836 were positive affected words(like good) and degree adverbs were 219.

Zhang et al[18]. also gives heuristic approach which was based on linguistic approach and can be divided in to two factors for use of linguistic rules.

- Dependency relation between children and words.
- The negation type of children to find modified polarity in each word.

Another method used by Zhang for Chinese sentimental analysis uses SVM which is used to classify review in Chinese. They used some training features like “appraisal phrases and bag of words”.

Here, appraisal phase is refers to emotions and feelings for any object. They are extracted using Hownet lexicon method.Amazon VN review method is used to evaluate it. With the SVM method and kernel, some other techniques were also introduces like decision tree and Multinomial Naïve Bays. From the above all methods, the best technique and accuracy rate was given by SVM classifier.

5.2 URDU

This system has very special features for language itself as this language is written from right to left. This language is based on context sensitive and space cannot provide boundaries for the words. Two different words can be written without using the space and even sometimes a single word may consist of several spaces too. Due to existence of derivations, duplications and inflections, this language has very complex morphology. Like there are several ways for plural determination in this language.

Sayed et.al[19] provides this system language for sentiment analysis which is based on concept of SentiUnits lexicon specifically used for Urdu. This SentiUnits consists of two types of adjectives namely single and multiple phrases adjectives. Every SentiUnits can have five attributes. The names of these attributes are Intensity, Polarity, Orientation, Modifier and Adjectives.Adjectives can be classified into two parts, one for describing quantity and quality and other is

used to describe people.Modifiers can be divides into absolute, superlative and comparative degree. The system performs the three major steps known as pre-processing, classification and Shallow Parsing. The feature of each step is defines as:

- Pre-processing uses HTML to make text and then segmentation technique is applied on this produces text.
- Classification technique is done by comparing semi-unit obtain in shallow parsing step.
- Shallow Parsing provides negation and sun entities known as senti-units.

5.3 FRENCH

It is another supervision system used in sentiment analysis. It is used in French for reviews in movies. Ghorbel[20] introduce this system and uses SVM classifier in order to provide the outputs. Classifiers uses features which can be categorized into three parts namely morpho-syntactics, sentiment features and lexical.

Lexical is first feature which uses Unigrams methodology. A list of stop words improves performance or efficiency of unigrams. A lot of stop words used in French like la, me, re, ja. All infected words are used to decrease features of unigrams. Post tags were introduces to raise unigrams with relation of morpho-syntactic information to provide the negation handling and solve ambiguity problem for future use.

An external resource features method is used known as "SentiWordNet". This technique was used to compute polarity of words which is calculated by converting the French words to English language. This above three steps provides the accuracy rate of 93.2% for feature in sentiment analysis. Some errors was classified which were occurred due to misspell of words, translation errors or system errors.

5.4 SPANISH

Brooke[21]introduces this system in sentiment analysis using concept of lexical dictionaries. This dictionary provides score for each word present in range of -5 to 5. Sentiment Orientation Calculator or SO-CAL is designed to analyse this language in sentiment analysis. They handle negation by shifting the words.

Words can be shifted by 4 places in order to remove negation. Every intensifier achieves a new value. Score of each words for intensifies is multiply by value of intensifiers to produce sentimental score. According to the observations, bias resultfor negative words in lexical classifiers. To overcome this bias problem, a constant value is added to negative expression.

There are basically three different ways which forms dictionary of Spanish:

- Manually build dictionary from scratch or initial point.
- Using bi-lingual dictionary which provides automated translation of words.
- Manually update lists from the bi-lingual dictionaries.
- The first type of dictionary includes informal words which can be built depend upon the spoken.

Automated machine dictionaries consist of formal words which are act as universal words.Due to above reason, manual dictionary is preferred as it provides flexibility and higher understanding of words which can be used in daily life. After review, SO-CAL concept link with SVM classifierfor training on unigrams. Translation of resources leads to loss in information and considered as good baseline. They tell that Language specific resources and domains understanding provide best method in sentimental analysis process.

6. CONCLUSION

In this paper, we discussed about Sentiment Analysis. Sentiment Analysis used in web services in order to provide reviews for the customer. It also promotes the business and industrial applications which are discuss in the mentioned papers in references. Paper can be divided into two parts. First part provides technique used in sentiment analysis for feature extraction. Classification technique is most popular which also consists of machine learning methods are discussed. Some learning methods like supervised and unsupervised learning techniques are also discussed in details and their application with their significance. Lexicon based approach which gives rise to unigrams is also focus in above paper. The main focus since was given on English in independent linguistic sentiment. we have discussed some of the Asianlanguages like French, Urdu, Chinese and their role in sentiment analysis This paper briefly share the knowledge of tools and sub techniques that are used for sentiment analysis extraction.

ACKNOWLEDGEMENT

This work was supported in part by the Council of Scientific and Industrial Research (CSIR), New Delhi with file no. 09/263(1049)/2015-EMR-I.

REFERENCES

[1] K. Garg and G.Jain "Influence of Natural Language Processing in Education." International Journal of Advance Engineering and Research Development, Volume 4 Issue 6, 2017.

[2] K. Garg and G. Jain "Natural Language Processing in Business Strategic Planning." in International Conference on "Computing for Sustainable Global Development, Delhi, March 2017.

[3] http://www.sciencedirect.com/science/article/pii/S2090447914000550.

[4] B.Liu,“ Sentiment Analysis and Subjectivity”, Handbook of Natural Language Processing, Second edition, 2010.
[5] A. Guizzardi, A. Stacchini – “Real-time forcasting regional tourism with business sentiment surveys”; Tourism Management, Vol. 47, pp. 213-223, 2015
[6] T. Wilson, J. Wiebe and P. Hoffmann, “Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis” in Proceedings of Human Language Technology and Conference on Empirical Methods in Natural Language Processing, October, 2005, pp. 347-354.
[7] K. Garg and G. Jain, "Comparative Study of Opinion Mining Tools." International Journal of Advance Engineering and Research Development, 4(2), June 2017.
[8]https://lct-master.org/files/MullenSentimentCourseSlides.pdf.
[9]http://www.sciencedirect.com/science/article/pii/S2090447914000550.
[10] https://en.wikipedia.org/wiki/Naive_Bayes_classifier.
[11]https://en.wikipedia.org/wiki/Bayesian_network.
[12]https://en.wikipedia.org/wiki/Artificial_neural_network.
[13] Fu Xianghua, Liu Guo, Guo Yanyan, Wang ZhiqiangMulti-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon Knowl-Based Syst, 37 (2013), pp. 186-195.
[14] http://www.sciencedirect.com/science/article/pii/S2090447914000550.
[15] Hatzivassiloglou V, McKeown K. Predicting the semantic orientation of adjectives. In: Proceedings of annual meeting of the Association for Computational Linguistics (ACL’97); 1997.
[16] Korayem, Mohammed, Khalifeh Aljadda, and David Crandall. "Sentiment/subjectivity analysis survey for languages other than English." arXiv preprint arXiv:1601.00087 (2016).
[17]C. Zhang, D. Zeng, J. Li, F. Wang, and W. Zuo. Sentiment analysis of chinese documents: From sentence to document level. *Journal of the American Society for Information Science and Technology*, 60(12):2474–2487, 2009
[18] C. Zhang, W. Zuo, T. Peng, and F. He. Sentiment classification for chinese reviews using machine learning methods based on string kernel. In *Convergence and Hybrid Information Technology, 2008. ICCIT’08. Third International Conference on*, volume 2, pages 909–914. Ieee, 2008.
[19] A. Syed, M. Aslam, and A. Martinez-Enriquez. Lexicon based sentiment analysis of urdu text using sentiunits. *Advances in Artificial Intelligence*, pages 32–43, 2010.
[20]H. Ghorbel and D. Jacot. Sentiment analysis of french movie reviews. *Advances in Distributed Agent-Based Retrieval Tools*, pages 97–108, 2011.
[21] J. Brooke,M. Tofiloski, andM. Taboada. Cross-linguistic sentiment analysis: From English to spanish. In *Proceedings of the 7th International Conference on Recent Advances inNatural Language Processing, Borovets, Bulgaria*, pages 50–54, 2009.

