A Comparative Analysis of Opinion Mining and Sentiment Classification in non-English Languages

Nishantha Medagoda[#1], Subana Shanmuganathan[#2], Jacqueline Whalley[#3]
Auckland University of Technology

[1]nmedagoda@yahoo.com, [2]subana.shanmuganathan@aut.ac.nz, [3]jacqueline.whalley@aut.ac.nz

Abstract*— In the past decade many opinion mining and sentiment classification studies have been carried out for opinions in English. However, the amount of work done for non-English text opinions is very limited. In this review, we investigate opinion mining and sentiment classification studies in three non-English languages to find the classification methods and the efficiency of each algorithm used in these methods. It is found that most of the research conducted for non-English has followed the methods used in the English language with only limited usage of language specific properties, such as morphological variations. The application domains seem to be restricted to particular fields and significantly less research has been conducted in cross domains.

***Keywords*— Natural Language processing, Text mining, Machine Learning.**

I. INTRODUCTION

Text mining is a sub area of the study of natural language processing that relates to understanding and generating the human languages such as English, French, Japanese, and Hindi etc. The understanding of a given language is not only of the spoken language but in the use of written scripts as well. Text mining is more suited to the written text of documents including the textual information about, facts and opinions. Opinions are subjective expressions of human thoughts, emotions and feelings. The research area of analyzing the opinions contained in texts is popularly known as opinion mining and it is basically about two methods that are run in a sequence [1]. The first identifies the subjectivity of the opinionated sentence or clause of the sentence and the second classifies the opinionated text as positive or negative. The former method is known as subjectivity classification and the latter one is referred to as sentiment analysis.

It is apparent that presently more people, especially web surfers, are expressing their views, opinions or experience on politics, products, services and many other things on the web more than ever before. This has been increased by the introduction of social networks in the later part of the 20thcentury. Mining opinions is a systematic approach that reveals precise information hidden in these views [2]. If customers are interested in finding specific information on a certain product or service, the opinion mining system helps them enormously in their investigation. The relevant information can be gathered using opinion mining tools without having to read the verbal comments of the clients who use the same product or service. Governments as well as political parties benefit enormously from the use of data mining to predict election results based from comments given by the public using the social networks as the means for voicing their opinions..

Manufacturers or merchants could also be benefited by using opinion mining systems. They might be interested in determining the success of a new version of a product or service based on its popularity or identifying the demographics of those who like or dislike the special features of the commodity before launching a new advertising campaign. Identifying this kind of information systematically by using opinion mining tools saves time and money by comparison with the use of time consuming surveys or market research. In addition the results are likely to be more accurate and reliable since the data has been created by real customers in ideal situations without forcing responses from them.

Traditionally the opinion or comments are collected using questions where the researcher has allocated some space in the questionnaire to write views. These questions are defined as the open-ended questions and information other than that of a quantitative nature is included in the responses. [3]. Today the most popular resource of collecting such information, also from a freely available source, is the World Wide Web. Blogs, review sites and micro blogs, all provide a good understanding of the perceptions of customers of the products and services [4]. The social networks are the newly invented repositories for customer comments. Social networks especially contain a wealth of human feelings and expressions on a vast number of topics such as politics, products, services and actions taken by the governments or countries [2].The content of the social webs is dynamic and rapidly changing to reflect the societal and sentimental fluctuations of contributors as well as the use of language. Even though contents of the social media are messy and highly heterogeneous they represent a valuable source of information of attitudes, interests and expectations of the citizens and consumers everywhere [5]. Another popular source of expressions of one's personal opinions for any given theme is the blogs. With the dramatic increase of internet usage the blog pages are also growing rapidly. Unlike the comments or opinions in social webs the blogger content tends to be longer and the language is more structured. Blogs are used as a source of opinions in many of the studies related to sentiment analysis [6]. Review sites are useful for any consumer who is looking for the others' comments on a certain product or services. A large and growing body of user generated reviews is available on the internet [4].

The remainder of this paper is organized as follows: Section II provides a brief description of widely used methods

in sentient classification followed by an evaluation of methods of opinion classification in section III. A detailed review of sentiment classification in Hindi is presented in section IV. Section V explains two studies in the Russian Language and a comparison of Chinese sentiment classification is reported in section VI. Finally, section VII concludes the review task and discusses some possible directions for future research.

II. WIDELY USED METHODS IN SENTIMENT CLASSIFICATION

The core of the sentiment analysis is first identifying the subjectivity of a sentence containing an opinion followed by judging the polarity (sentiment) of the view expressed. These tasks are carried out using machine learning algorithms. The choice of the specific learning algorithm used is a critical step. The methods of determining the semantic orientation used for identifying the polarity of the sentence are categorized into two approaches: supervised and unsupervised classification techniques. The evidence of initial attempts on the application of an unsupervised approach by Pang et al in 2002 and supervised method by Peter Turney can be seen in Kobayasi et al. [7].Supervised classification algorithm is one of the learning algorithms most frequently used in text classification systems [8]. In supervised classification two sets of opinion sets are required namely, training and testing data sets. The training data set is used to train the classifier to learn the variation of the characteristics of the sentence or document and the test data is used to validate the performance of the classification algorithm. The supervised machine learning techniques, such as Naïve Bayes, support vector machines (SVM) and maximum entropy, are the most popular ones and they have been proven to be the most successful in sentiment classification [4].

Naïve Bayes algorithm is the most widely used and it is a simple but effective supervised classification method [9]. The basic idea of the method is to estimate the probabilities of sentiment (either positive or negative) for the given opinion using the joint probabilities of a set of words in a given category. The method is totally dependent on the naïve assumption of word independence. Support vector machine is the best binary classification method [9] proposed by Vladimir Vapnik. SVM is a non-probabilistic classification technique that looks for a hyper plane with the maximum margin between positive and negative examples of the training opinions. In additional to the above, K-nearest neighbor classification (KNN) is based on the assumption that the classification of an instance is most similar to classification of other instances that are nearby in the vector space. In comparison to the other text classification methods such as Naïve Bayes, KNN does not rely on prior probabilities and it is computationally efficient [10].

Besides the above machine learning classification methods some combined techniques of rule based and supervised learning approaches have been proposed in other studies [4]. In these studies hybrid classification is carried out ,i.e. if one classifier fails to classify an opinion or document, the classifier passes the document to the next classifier until the opinion is classified or no other classifier exists. In the rule based approach, rules are to be defined which will contain antecedents and its associated consequent that have if-else relationships. In this methodology certain rules are to be formed and then the sentiment analyzed using these rules [11].

III. EVALUATION OF METHODS OF SENTIMENT CLASSIFICATION.

The final step of any classification work is to carry out the evaluation of the proposed algorithm using standard techniques. The purpose of evaluation in any machine learning based classification algorithm is to determine the usefulness of the learning classifier on various collections of data sets. The properties of any learning method such as accuracy, comprehensibility and conciseness aim to measure the process of evaluation [12]. The performance measures of a learning algorithm are summarized in various ways. The most commonly used performance metrics are Accuracy, Precision/Recall and Receiver Operating Characteristic (Roc) analysis.

In this research paper we present a comprehensive literature review of opinion mining and sentiment classification of non-English texts. The critical review aims to carry out a sentiment classification study on several languages, by investigating the domain of the study, the classification methods employed, evaluation methods deployed for each language specific work and the level of accuracy achieved by the proposed method. Due to the high volume of the studies carried out in the past decades, we selected a few studies in Hindi, Chinese, and Russian languages. Two papers of each language were considered in the study to compare and contrast the methodologies used and the accuracy.

IV. SENTIMENT ANALYSIS STUDIES IN HINDI

Hindi, one of official languages in India, has a speaker population of 258 million [13]. Two sentiment analysis studies, “A fall-back strategy for Sentiment Analysis in Hindi” by Joshi et.al, [14] and “A lexican resource for Hind Polarity Classification” by Bakliwal et.al, [15]are reviewed in this paper to identify the scale and the depth of sentiment analysis in Hindi. The two studies are summerised in table I.

Joshi et, al, [14] used a SVM classifier in order to determine the polarity of the opinion, in their first approach which is named as in-language sentiment analysis. Google translation was used to translate the corpus in Hindi to English in a machine translation (MT) based method. The translated corpus is input into a classifier. In the third approach the *synset* corresponds to the English in *Senti* Word Net (SWN) projected to the corresponding to *synset* in Hindi to build the *senti* word net (H-SWN) for Hindi.

In this method they managed to create the H-SWN of 16253 *synsets* which consists of Adjective, Adverb, Noun and Verb. Classification under the resource based method was conducted with varying the different structural features, such as changing the n-grams and with stemming and without stemming. It is stated in the paper that the poor performance by the MT based approach was caused by translation errors. The conclusion mentioned that an annotated corpus is an

essential resource for sentiment analysis in languages like Hindi. In addition to the above limitations the research was based on two main assumptions. The assumption of the sentiment of a synset retained across English and Hindi, is very critical for the accuracy of the method. Similarly the assumption of preservation of the sentiment of a document in translation is crucial in the algorithm.

The research of a lexican resource for Hindi Polarity Classification by A. Bakliwal et al, [15] started with building a subjective lexicon. The lexicon was built using a seed list of 45 adjectives and 75 adverbs.

The results revealed that the scoring method outperforms the unigram presence method. The authors claimed that the *wordnet* inability to perform the word sense disambiguation is a major limitation of the proposed algorithm. The morphological variation of Hindi language would affect the results as the stem of a given word accounted for polarity score calculation.

Study	**Appication Domain**	**Classification Method**	**Features**	**Accuracy**
[14]	Movie Reviews: A corpus of 250 reviews	In-Language using SVM	Term Frequency	74.57
			Term Presence	72.57
			TF-IDF	78.14
		Machine Translation Based using SVM	TF-IDF	65.96
		Resource-based using Senti Word list	Most Common Sense	56.35
			All Sense	60.31
[15]	Product reviw data set translated from English to Hindi	Using subjective Lexican constructed by the otrher	Unigram Presence	77.34
			Simple Scoring Method	79.03

Table I : Example of Sentiment Analysis in Hindi

In the above studies the important quality of Asian languages such as, the morphological variation, has not been considered in the analysis. Even though. Bakliwal et all [15] mentioned the effect of the morphological variation of words in their final results as an attempt of incorporating the morphological aspect in algorithm, it is missing.

V. SENTIMENT ANALYSIS WORKS IN RUSSIAN

The Russian language is the most widely spoken native language in Europe with 144 million speakers in Russia, Ukraine and Belarus [16]. The written Russian language uses the Cyrillic alphabet. Yussupova et.al, [17] used a Machine learning approach in their study of sentiment analysis of Russian text and then compared and contrasted this with the approach attempted by Pak et.al,[18].The summary of the two studies including the methods used and accuracy achieved with each approach are presented in Table II.

The goal of their research was to discover how lemmatization affects the accuracy of sentiment classification [17]. The classification of opinions into two, three and five classes was the main aim of the study conducted by Pak et, al. who used a language independent approach. They used an SVM classifier which was totally dependent on feature based attributes such as, n-gram, pos-tags and dependency parsing. In addition to n-grams, the authors proposed a new feature which was similar to n-gram called d-grams. In the research by Yussupova et, al. the "Bagging algorithm" was integrated into a Naïve Bayes classifier to improve the accuracy of the classification.

The training and evaluation of the developed algorithm was carried out using reviews of Russian bank loans. One of the drawbacks of the study was the unbalanced sample used. The authors analysed only 304 positive reviews but 850 negative reviews. Moreover valuable information may have been lost due to lemmatization of the key words.

Study	**Appication Domain**	**Classification Method**	**Features**	**Accuracy**
[17]	Bank Cutomer Reviews	Naïve Bayes	Bagging Multinomial Model	87.69
		SVM	Length>2	88.21
[18]	Product reviews (Books, Movies, Camaras)	SVM	Ngram+tfiDf	90.4
			Digram+tfiDf	91.3

Table II: Example of Sentiment Analysis in Russian

In the study by Pak et, al. [18], "D-grams" are constructed from a dependency parser tree, where words are linked by syntactic relations [18]. In order to avoid the domain–adaption in the classification the proposed system was tested in all combined (Books, Movies and Cameras) reviews. The final results revealed that the proposed system was the most accurate out of the all combinations of options when classifying all the reviews together. The developed algorithm also runs on unseen data in different tracks. Tracks are defined by varying mode (number of classes), features, weights and training set. 2-class track consists of 6 systems of binary class with different d-grams and weights. 3rd and 4th tracks are multiclass and different training sets. According to the performances, in 3-class track the experiment with movie reviews show highest accuracy while in 4th class track reviews on cameras achieved the highest.

VI. SENTIMENT ANALYSIS WORKS IN CHINESE

The Chinese language consists of one shared common written language and a number of spoken dialects. The seven

main dialects are Mandarin, Cantonese, Hakka, Min, Xiangand Gan.The Chinese writing system is based on meaning instead of on phonetics. Two Chinese language sentiment analysis studies are considered in this paper.

The first one is "A Morpheme-based Method to Chinese Sentence-Level Sentiment Classification" by Xing et, al. [19]. In this approach the morphological variations of a set of sentiment bearing words were integrated into the algorithm by extracting the morpheme and then inferring the semantic orientation of the words. These morphemes were of two types namely, positive and negative. According to the authors the Chinese sentiment words can be categorized into static and dynamic sentiment words. These static and dynamic sentiment words contain a key morpheme that determines their emotional tendency. The morphological productivity of positive and negative morphemes contained in the words in the Chinese lexicon used was calculated before determining the polarity of a review. Then the opinionated sentence was first segmented in to four types of sentiment phrases. Using the morpheme productivity score the average polar intensity of the review was estimated to decide the semantic orientation. A set of predefined thresholds was used to determine the semantic orientation i.e. whether the given opinion was positive, negative or neutral. The proposed system has been tested at different levels of granularity namely at morpheme-level, word-level and phrase-level. As per the results presented in the paper, the phrase level, classification outperforms others with respect to F- value. Also the authors compared the proposed system with other morpheme based systems for Chinese languages and concluded that the proposed system performed better than others even though the F-score was slightly worse. In their proposed method the complexity of classification was very high when compared with the other methods. Some rules were included in the method to establish thresholds but without giving justification for the resulting values.

The second study, "Chinese Sentence–Level Sentiment Classification Based on Fuzzy Sets" by Guohong Fu and Xin Wang [20] was aimed at comparing the Chinese sentiment analysis studies. In this study, as in the previous paper, the sentiment morphemes were extracted from a sentiment lexicon and then an opinion score was calculated using chi-square techniques. Word and Phrase level polarities were then each identified using a set of rules. The word level polarity was determined by a key morpheme contained in either static or dynamic polar words. Then the final sentiment intensities of an opinionated sentence were calculated by summing the opinion scores of all phrases within the sentence. To handle the intrinsic fuzziness in sentiment polarity such as "positive", "neutral" and "negative" the authors applied fuzzy set theory to sentiment classification. The fuzzy sets for each category of positive, neutral and negative sentiments were defined by three different membership functions based on semi-trapezoid distribution. The upward rise in the semi-trapezoid distribution for three cases with different parameters was used to determine the polarity by maximizing the membership.

The proposed method had been carried out in three modules namely, lexicon analysis module, subjectivity detection and sentiment classification. A sentiment density based naïve Bayesian classifier was also embedded into the second module to perform the opinion detection in the sentence. The opinions saved in a standard Chinese opinion corpus were tested in the experiment. The 843 documents with 62% of opinion sentences were included in the test data set. The analysis phrases outperformed analysis at the other two of levels of granularity studied, morpheme and word. In the comparison of best system for Chinese opinion, the proposed system gave a higher F-score and it was concluded that the fuzzy based system as the best current model.

In a study by Guohong Fu and Xin Wang [20] one of the limitations is usage of an existing sentiment lexicon to extract the sentiment morphemes. This may limit the integration of all sentiment morphemes used in the Chinese language. But the measurement used in this study to calculate the degree of association between sentiment morphemes and sentiment words is more efficient and reliable than with the same use in the study by Xing et, al. [19].

VII. CONCLUSION

In this paper we investigated the opinion mining and sentiment classification methods in non English languages. Three languages each with two studies were used in the investigation and comparisons were conducted within and among the languages. Most of the research based on statistical methods looked at the lexicon of sentiment words as the primary source of determining the polarity of the word as the basic unit of analysis. Then the bag of words method was employed to measure the polarity of the sentence. As it appears, the Bayesian techniques are the most popular classification algorithms in these statistical based approaches. So far no attempt has been made to classify polarity using language specific features, such as morphemes based methods, except in Chinese, even though some other non-English languages are morphologically rich. Interestingly, some attempts have been made in Chinese where the concept of *senti* morpheme has been introduced to determine the sentiment orientation of a word. The performances of sentiment classification algorithms are higher in domain specific experiments than in domain independent experiments. The task of opinion mining and sentiment classification is complex and further research is needed into the development of efficient algorithms and into their application to various languages.

REFERENCES

[1] Bing, L.Sentiment Analysis and Subjectivity. To appear in Hand Book of Natural Language Processing.(editors: N-Indurkshya and F.J. Damerau)2010 (pp. 3-8)

[2] Budhhika H. Kasthuriarachchy, Kasun De Zoysa and H. L Premarathne. A Review of Domain Adaption for Opinion detection and Sentiment Classification. The international Conference on Advances in ICT for Emerging Regions, (pp. 209 - 213). Colombo. -2012

[3] Medagoda N, Weerasinghe R. An application of Document clustering for Categorizing open-ended Survey Responses. Colombo,(pp.102-110).-2011

[4] G. Vinodhini and RM. Chandrasekaran. Sentiment Analysis and Opinion Mining: A survey. International Journal of Advanced Research in Computer Science and Software Engineering , (pp.282 – 291).-2012

[5] C. Rodriguez, J.Grivolla and J.Codina. A hybrid framework for Scalable Opinion Mining in Social media,(pp. 46-52)-2012

[6] Teng-kai, Chia-Hui Chang. Blogger-centric contextual advertising. 8th ACM conference on Information and knowledge management, (pp. 1803-1806).-2009

[7] Kobayasi N.,K.Inui and Y. Matsumoto. Opinion mining from web documents: Extraction and structurization. Transaction of the Japanese Society for Artificial Intelligence, (pp. 227-238).-2007

[8] S.B.Kotsianta, I.D. Zaharakis, P.E. Pintelas. Machine Learning: a review of classification and combining techniques. Springer,(pp.159-190).-2006

[9] RuiXia,Chenquing Zond and Shoushan Li. Ensemble of feature sets and classification algorithms for sentiment classification. Information Sciences ,(pp.1138-1152).-2011

[10] Eui-Hong (Sam) Han, George Karypis and VipinKumar.Text Categorization Using Weight Adjusted k-Nearest Neighbour Classification. 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), (pp. 53-65).-2001

[11] kawathikar S. A. and Manali S. kashigar. (2012). Sentiment analysis using hybrid approach involving Rule-Based and Support Vector Machine. JOSR Journal for Engineering, (pp.55-58).-2012

[12] Japkowize, N. Why Question Machine Learning Evaluation Methods. The 1st AAAI-06 Workshop on Evaluation Methods for Machine Learning. Ontario (pp.6-11).-2006

[13] Kan, D. Rule based approach to sentiment analysis at ROMIP 2011.(Availablefrom:http://www.dialog-21.ru/digests/dialog2012/materials/pdf/Kan.pdf)

[14] Aditya Joshi, Balamurali A. R. and Pushpak Bhattacharyya. A Fall-back Strategy for Sentiment Analysis in Hindi: a Case Study. International Conference on Natural language Processing (ICON), Karagpur.-2010. (Availablefrom: http://www.cse.iitb.ac.in/~balamurali/papers/ICON%20229.pdf)

[15] Akshat Bakliwal, Piyush Arora and Vasudeva Varma. A lexicon resource for Hind Polarity Classification. The eighth international conference on Language Resources and Evaluation (LREC). Hyderabad.-2012

[16] Russian language. (Retrieved 05 13, 2013, from Wikipedia: http://en.wikipedia.org/wiki/Russian_language

[17] Yussupova N. and Diana Bogdanova. Applying of Sentiment Analysis for Texts in Russian Based on Machine Learning Approach. The Second International Conference on Advances in information Mining and Management, (pp. 8-14). Venice, Italy.-2012

[18] P.Pak A and Paroubek: Language Independent Approach to Sentiment Analysis. LIMSI Participation ROMIP. France-2011

[19] Xin Wang, Yanquing Zhao and Guohong Fu. (2010). A Morpheme - based Method to Chinese Sentence -Level Sentiment Classification. International Journal on Asian Language Processing , 95-105.-2010

[20] Guohong Fu and Xin Wang. (2010). Chinese Sentence–Level Sentiment Classification Based on Fuzzy Sets. The 23rd International Conference on Computational Linguistics (COLING 2010), (pp. 312-319). Beijing.-2010