ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

Proposition of a Hybrid Approach for Sentiment Analysis of Travel Domain Data

A Srinivas, M Hanumanthappa

Research Scholar, Rayalaseema University, Kurnool, Andhra Pradesh, India

Professor, Dept. of Computer Science & Applications, Bangalore University, Jnanabharathi Campus, Bangalore, India

ABSTRACT: Sentiment analysis is a process of extracting, identifying and categorizing a writer's emotion, expressed in the form of text, by implying a computational method. The dominance of machine learning and lexicon-based approaches in the field of sentiment analysis has motivated this proposition of a hybrid approach that uses supervised machine learning methods. The travel domain data is concentrated more due to various reasons mentioned in this proposal. The proposed hybrid approach is a novel approach to sentiment analysis with an inclusion of effective POS tagging is supposed to yield better results.

KEYWORDS:Sentiment analysis; POS tagging; machine learning; lexicon analysis; supervised learning.

I. INTRODUCTION

Sentiment analysis is a standout amongst the most worked subfield of Natural Language Processing (NLP) and has seen a lot of research done amid past decade. It is a procedure of extracting, identifying and distinguishing sentiments from content information. Emotion artificial intelligence or opinion mining are some alternative terms used for sentiment analysis. Sentiment analysis is imperative from business point of view for all online networking or any item based advertising. The social media mammoths like twitter are intensely relying upon sentiment analysis to continuously develop a large portion of their marketing and application development. While recent studies are concentrating more on social media data to try sentiment analysis with, our approach is to try the same on travel domain related data. Many algorithms are being used for sentiment analysis, fall under either of the two well established computing approaches: Machine Learning and Lexicon analysis.

Machine learning approach to the sentiment analysis can be done in two ways. The first one, supervised learning in its simplest form, is a type where to produce a predicted outcome training data sets are used on a set of known inputs. A well-known algorithm is used to carry out this process. The output is already known in supervised learning. It is one of the mostly implemented machine learning methods in modern industries. Decision trees, linear and rule based classifiers fall in this category of approach [1]. The second one is called as unsupervised learning, where there are no training data sets and outputs are known. That is why it is a more complex method and is being used in far smaller number of applications so far. In unsupervised learning, an Artificial Intelligence (AI) agent goes into the problem without having any prior knowledge about the problem. The other approach to sentiment analysis apart from machine learning is Lexicon-analysis. Dictionary-based approach is one of the methods under lexicon-analysis. Few words which have maximum influence on the output of the content in term of sentiment score are collected and their possible synonyms and antonym are searched in well-known dictionary repositories like thesaurus or WordNet etc. The word seed will be added with these newly found synonyms and are used every time the word is used in searching.

In Corpus-based approach of Lexicon-analysis, opinion words with context specific orientation are searched depending on syntactic patterns or patterns that occur together along with a seed list of opinion words to find other opinion words in a large corpus [2].

Apart from these mainstream approaches, Part-of-Speech tagging is an important part of sentiment analysis. In corpus linguistics, part-of-speech tagging (POST or POS tagging), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context—i.e. relationship with adjacent and related words in a phrase,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

sentence, or paragraph [3]. A simplified form of this is used in the identification of words as nouns, verbs, adjectives, adverbs, etc. Filtering a text for useful content involves tokenization and normalization. Tokenization is the name given to the process of chopping up sentences into smaller pieces (words or tokens). The segmentation into tokens can be done with decision trees or customized tokenizers like openNLP tokenizer, Stanford tokenizer etc.

The approach made for sentiment analysis will be a hybrid one including Dictionary-based lexicon analysis and Support Vector Machines (SVM). Although it may look strange involving two methods from different principles to conduct sentiment analysis, it is necessary for addressing issues related to analysis of datasets of various backgrounds. The absence of a single efficient method to fulfil the purpose has given an idea of proposing such a complex yet efficient mechanism. The figure below depicts different phases of this proposed analysis.

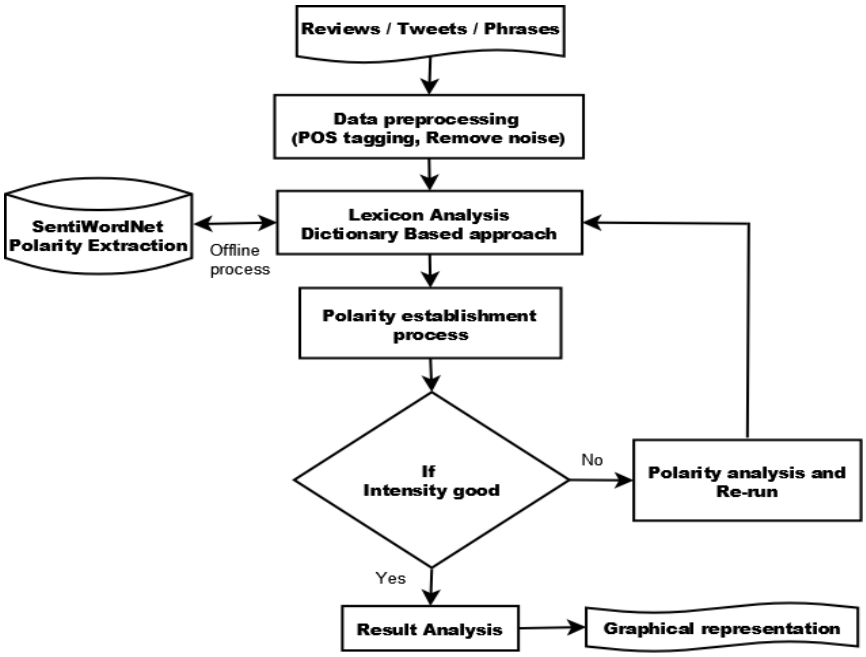

Figure 1. A hybrid approach for sentiment analysis

Following contents will briefly discuss about the different phases shown in the diagram. The whole process of sentiment analysis using hybrid approach has been divided into 5 phases namely, Data acquisition, Text preparation, Sentiment detection, Sentiment classification and result analysis.

II. DATA ACQUISITION FOR HYBRID APPROACH

Sentiment analysis carries a huge importance in various fields due to its ability to classify emotions using text as input. Consumer product marketing, social media marketing are few areas that are benefited through sentiment analysis. The proposed approach is concentrated more on user provided data related to travel domain. Travel domain having huge demand for creative and effective advertising, can easily be profited from sentiment analysis just like any other domain. The intended data collection for the analysis is from travel domain. Most of the contents of such data are user reviews. The data can be used either in text format or in JSON (JavaScript Object Notation). The reason why JSON is also considered is because of the popularity, flexibility provided by the JSON format. Most of the programming languages used today support JSON. It is one of the simplest data representation and interchange format which stores data in a simple {name : value} form. Also, it is easy to attach polarity value to the words if the data representation is in JSON.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

III. INPUT TEXT PRE-PROCESSING

Text pre-processing is an important phase in the process of sentiment analysis. Pre-processing includes text filtering, POS tagging and removing noise from the input data. A text may have many words which have no effect on sentiment analysis result. Such words are removed from the input first. Next in line are those words which are either misspelled or slang words. Stop word removal is another step in pre-processing to remove adverbs, articles, prepositions which have no contribution to the sentiment score [4]. Stemming can also be done on the text to reduce a word to its root form. For example, talking and talker have common root word talk. So, the stemmed form of both these word is talk.

Another part in the pre-processing is the Part-of-Speech tagging (POS tagging). It is an integral part of sentiment analysis because of its ability to identify and classify different words based on their nature. Sentiment analysis cannot be applied on all sorts of text contents. Some parts of the text are invaluable to the analysis process. For example, nouns, adverbs etc. don't contribute to the final sentiment score. To determine emotions it is necessary to look at few specific words. In a sentence, 'The car's seats were wonderful', the word 'wonderful' is the one useful for analysis and give rise to a positive result. Words like, 'the', 'car', 'seat' may not have much effect on the result. Sentiment analysis has several benefits from POS tagging. First, POS tagging is essential in Word Sense Disambiguation. Words with similar characters may have more than one meaning. For example, the word "bear" has more than one meaning. Its meaning depends on the 'part of the speech'. 'The bear starts running' and 'In future it may bear some fruits' are two sentences with same word 'bear' in them but with different meaning. The meaning can only be understood by an analyser only if POS tagging is done on that sentence. Second, further steps of sentiment analysis can be benefited by POS tagging of words. POS tagging not only identifies words but it can identify characters like comma, semicolon etc. These neutrally placed characters have no influence on sentiment score. Identifying them could be very useful in further processing.

POS tagging for this proposed hybrid approach for sentiment analysis is done using Stanford coreNLP [5]. The input is fed in the form of a text file and POS tagged output is provided as an xml file. Following figure 2 is a snapshot of an example of POS tagging done using Stanford coreNLP.

Input sentence: **Nothing to complain but nothing to feel too good about either.**

Result of POS tagging through Stanford coreNLP

Id	Word	Lemma	Char begin	Char end	POS	NER	Normalized NER	Speaker	Sentiment
1	Nothing	nothing	0	7	NN	O		PER0	
2	to	to	8	10	TO	O		PER0	
3	complain	complain	11	19	VB	O		PER0	
4	but	but	20	23	CC	O		PER0	
5	nothing	nothing	24	31	NN	O		PER0	
6	to	to	32	34	TO	O		PER0	
7	feel	feel	35	39	VB	O		PER0	
8	too	too	40	43	RB	O		PER0	
9	good	good	44	48	JJ	O		PER0	
10	about	about	49	54	IN	O		PER0	
11	either	either	55	61	CC	O		PER0	
12	.	.	61	62	.	O		PER0	

Figure 2. POS tagging on an input sentence and result table

To understand the result of a POS tagging process, one that is shown in above example, one must understand the meaning of abbreviations used in POS tagging. Following table number 1 shows these meanings [6].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

Table 1. Description of POS tags used in sentiment analysis

S.no	TAG	Description	S.no	TAG	Description	S.no	TAG	Description
1.	CC	Coordinating conjunction	13.	NNS	Noun, plural	25.	TO	to
2.	CD	Cardinal number	14.	NNP	Proper noun, singular	26.	UH	Interjection
3.	DT	Determiner	15.	NNPS	Proper noun, plural	27.	VB	Verb, base form
4.	EX	Existential there	16.	PDT	Predeterminer	28.	VBD	Verb, past tense
5.	FW	Foreign word	17.	POS	Possessive endings	29.	VBG	Verb, present participle
6.	IN	Preposition	18.	PRP	Personal pronoun	30.	VBN	Verb, past participle
7.	JJ	Adjective	19.	PRP$	Possessive pronoun	31.	VBP	Verb, non-3rd person singular
8.	JJR	Adjective, comparative	20.	RB	Adverb	32.	VBZ	Verb, 3rd person singular present
9.	JJS	Adjective, superlative	21.	RBR	Adverb, comparative	33.	WDT	Wh-determiner
10.	LS	List item marker	22.	RBS	Adverb, superlative	34.	WP	Wh-pronoun
11.	MD	Modal	23.	RP	Particle	35.	WP$	Possessive wh-pronoun
12.	NN	Noun, singular or mass	24.	SYM	Symbol	36.	WRB	Wh-adverb

IV. DICTIONARY BASED APPROACH

The hybrid approach comprises of two step sentiment analysis process where Lexicon-analysis being the first step. Utilization of a vocabulary is one of the two principle ways to deal with sentiment analysis and it includes ascertaining the opinion from the semantic introduction of word or expressions that happen in content [7]. With this approach, a lexicon of positive and negative words is required, with a positive or negative sentiment value allotted to each of the words. Diverse ways to deal with making lexicons have been proposed, including manual and programmed approaches. For the most part talking, in lexicon based methodologies a bit of instant message is spoken to as a pack of words. Following this representation of the message, sentiment values from the dictionary are assigned to all positive and negative words or phrases within the message. The final sentiment of the input is determined using after applying a sum or an average operation on these collected sentiment scores of individual words. Apart from a sentiment value, the aspect of the local context of a word is usually taken into consideration, such as negation or intensification.

In the lexicon-analysis corpora like SentiWordNet are used to extract polarity for each eligible word from the input sentence. Many such corpora provide a dictionary of polarized words to support Lexicon-analysis. But this process gets complex with complex input set. A complex input set is the one which has metaphors, sarcasm involved. In such cases the result of lexicon-analysis won't be efficient. Hence, it reduces the effectiveness of sentiment analysis.

V. POLARITY ESTABLISHMENT AND ANALYSIS

Machine learning is the second step of sentiment analysis process done after the lexicon-analysis. The method used here is called Support Vector Machines (SVM) of supervised learning [8]. Supervised learning is a part of machine learning methods which are easy to implement because inputs provided here are known and labelled. Another peculiar character of supervised learning is predictable outputs. SVM is a classification technique in machine learning that can act on a set of labelled data to classify it using defined parameters. SVM uses something called as a hyperplane or decision plane to classify these values. A hyperplane is the deciding factor to all the classification done using SVM. There is a possibility of defining more than one such hyperplanes for a set of labelled data. The classification is

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

represented on n-dimensional graph where each dimension is a defined feature of those set of objects. An optimal hyperplane would be the one that has highest marginal gap between itself and the nearest plots of the graph.

Keeping an option of a re-run of the algorithm in case of not meeting the intensity of the polarity assignment is the integral part of the proposed hybrid approach. This step could prove to be the best solution against the use of odd sentences in input. A simple algorithm to check the polarity intensity can be used in this step.

VI. RESULT ANALYSIS

Polarity establishment divides positive and negative words. This step is crucial for the analysis because individually polarized words contribute to the overall sentiment of the sentence. To analyze the result graphical representation will be used. Now, the methods used for sentiment analysis vary depending on their ability to decide the overall sentiment of an input. There is no single method that can be considered as ideal for sentiment analysis. Their efficiency is lying upon the way inputs are written. In other words, most of them work well when input sentences are clearly written (For ex: written without any emoticons, sarcasm, metaphor etc.). But the ability of such methods are questioned with addition of contents which are naturally not-part-of a language. In such cases it is easy to get undesired results.

Representation of the results must be done in such a way that, it must allow expressing the actual sentiment involved. This won't be a difficult task if the polarity establishment is up to the mark. Various tools and methods are available to plot graph or display the results in many form to distinguish such polarity.

Following figure 3 sums all the phases of proposed hybrid approach.

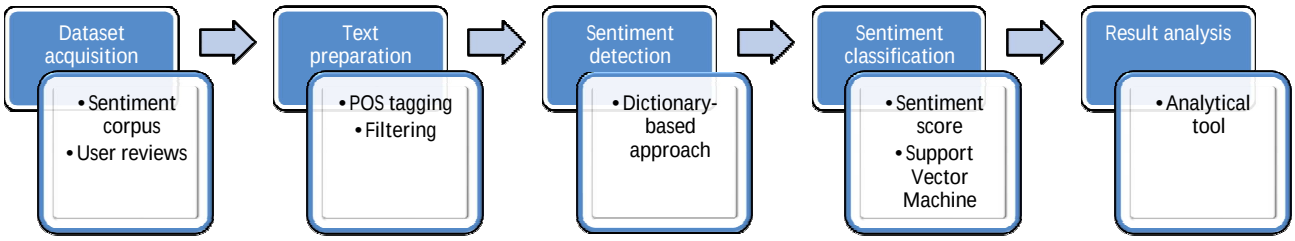

Figure 3. Phases in hybrid approach for sentiment analysis

VII. CONCLUSION AND FUTURE WORK

The proposed approach comprising of both lexicon based approach and machine learning could prove to be one of the viable approaches for sentiment analysis due to its simple design and efficiency. Even though, the hybrid approach is in the phase of proposition and it is yet to be put to practice, it is certain that the combination of predictive approach of machine learning and a well-defined corpora can yield desired results for sentiment analysis. The main intension of contents of this article is to present an outline for a hybrid approach for sentiment analysis using two predominant methods in the field. Also, the importance of POS tagging in the process of sentiment analysis is briefed in this paper.

As a future work this approach can be implemented using different methods of machine learning. Since, the lexicon-based approach doesn't offer many options in replacing the method; many supervised algorithms like Naïve Bayes, decision trees, neural networks can be used for sentiment analysis. One of these can replace Support Vector Machine in hybrid approach.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

REFERENCES

1. Walaa Medhat, Ahmed Hassan, Hoda Korashy, "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal, pp. 1093-1113, Volume 5, Issue 4, December 2014.
2. Douglas Rice, Christopher Zorn, "Corpus-based dictionaries for sentiment analysis of specialized vocabularies", New Directions in Analyzing Text as Data Workshop, London, Version 0.1, September 2013.
3. O Owoputi, B O'Connor, C Dyer, K Gimpel, N Schneider, A Smith, "Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters", Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 380-391, 2013.
4. H Saif, M Fernández, Y He, H Alani, "On stopwords, filtering and data sparsity for sentiment analysis of Twitter", pp. 810-817, LREC 2014.
5. CD Manning, M Surdeanu, J Bauer, JR Finkel, SJ Bethard, D McClosky," The Stanford CoreNLP Natural Language Processing Toolkit", Proceedings of 52nd AMACL, pp. 55-60, 2014.
6. MP Marcus, MA Marcinkiewicz, B Santorini, "Building a large annotated corpus of English: the penn Treebank", Computational linguistics, Volume19, Issue 2, pp. 313-330, June 1993.
7. E Cambria, B Schuller, Y Xia, C Havasi, "New avenues in opinion mining and sentiment analysis", IEEE Intelligent Systems, Volume 28, Issue 2, March-April 2013.
8. G Gautam, D Yadav, "Sentiment analysis of twitter data using machine learning approaches and semantic analysis", Proceedings of Seventh International Conference on Contemporary Computing,August 2014.