iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** XI **Month of publication:** November 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue XI November 2017- Available at www.ijraset.com

Sentiment Analysis Tools and Techniques: A Comprehensive Survey

Maria Christina Barretto[1], Sweta Morajkar[2]
[1, 2] *Assistant Professor, Computer Engineering Department, Don Bosco College of Engineering*

Abstract*: *Data analysis refers to a process of inspecting and transforming data in order to derive some useful information. The objective of this process is to discover interesting patterns that can be used to derive important conclusions in areas like business, science, and social science domains. Semantic analysis plays a very important role in decision making process. Due to increase in web technologies a large amount of data gets generated. In order to extract important data from this, various data mining techniques have been proposed in recent years. Many researchers have focused on finding some interesting patterns out of this data using semantic analysis techniques. These techniques basically deal with determining the contextual polarity with reference to a specific domain. The paper presents a comprehensive survey about various techniques and tools used for sentiment analysis with some newer approaches like sentiment analysis using LDA. We present a comprehensive survey of latest sentiment analysis techniques along with some comparison results.

Keywords*: *Sentiment, Opinion mining, Classification, Supervised Learning, Lexicon

I. INTRODUCTION

Data analysis addresses the methods for managing large amount of data. It is the art of processing raw data to extract some reasonable information. Data analysis is widely used in many industries and organizations to make better business decisions. Data comes from a wide variety of sources. Hence there comes a need to capture, store and analyse this data that has high volume, velocity and variety. Due to the overwhelming amount of data, it is necessary to perform accurate analysis of this data in terms of structure and content.

The exponential increase in Internet usage and exchange of user opinion has led to the collection of large amounts of data (chatting, social media communications, tweets, blogging, online transactions, e-commerce etc.) making the web a huge repository of structured and unstructured data. User behaviour about the web content is very important. Collective outcome from various users about web content can affect readers in getting an idea about some issues example social events, company strategies, product preferences, marketing campaigns and political movements. Considering all the above aspects, many researchers are focusing on extracting interesting patterns from the data to derive useful information.

Sentiment analysis is the process of recognizing and classifying different opinions/sentiments expressed online by individuals to derive the writer's attitude towards the specific entity[1]. An entity can represent any individual, item, event or topic. The attitude may be their judgement or evaluation, their affective state or the intended emotional communication. It is also referred to as opinion polarity; i.e. to analyse if someone has positive, negative or neutral opinion towards something. The expression sentiment analysis and opinion mining are used interchangeably.

Sentiment analysis can be done at three different levels; Document (coarse) level, Sentence (intermediate) level and Feature (fine) level. Document or Coarse level sentiment analysis tries to determine the sentiment from the whole manuscript or document and then classifies it as positive or negative based on the ideas expressed by the user. Each document focuses on a single entity or event and contains opinion from a single opinion holder. This binary classification of labelling an opinionated document as expressing either a positive or negative ("thumbs up" or "thumbs down") is called sentiment polarity classification.

To get a more refined view of different opinions expressed in the document, Sentence or Intermediate level sentiment analysis is used. This level of sentiment analysis filters out those sentences which contain no opinion. There are two kinds of information in a particular sentence; objective and subjective. An objective sentence states factual information about the world. A subjective statement expresses some personal feeling, belief or view. The task of determining whether a sentence is subjective or objective is called subjectivity classification. The resulting subjective sentences are further classified as expressing positive or negative opinions. This is called as sentence level sentiment classification. Hence in sentence level sentiment analysis, two sub tasks are performed; subjectivity classification and sentence level sentiment classification. This type is usually used for reviews and comments that contain one sentence and written by the user.

The above two techniques, work well when they refer to a single entity. However, in many cases opinions can be about entities and their various attributes or aspects. Most opinion mining techniques consider the opinions from a large number of opinion holders. An opinion from a single person is usually insufficient and hence a summary of opinions is desirable. This kind of a summary can be provided based on aspects, hence called aspect or feature level sentiment analysis [2]. Document and Sentence level sentiment analysis can determine the overall sentiment towards a service but might not be able to capture the full essence of the review. Therefore aspect level sentiment analysis focuses on recognizing all the sentiment expressions within a given document and the aspects or attributes to which the opinions refer.

II. LITERATURE SURVEY

Following section provides a detailed survey of various tools and techniques used for sentiment analysis.

A. Steps in Sentiment Analysis

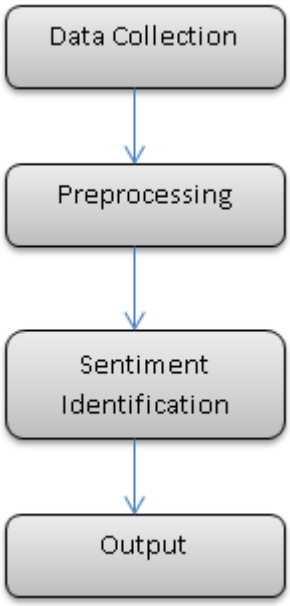


Fig. 1 Steps in Sentiment Analysis

1) *Data collection:* The first step in sentiment analysis process is to collect data. The main sources of data are from product reviews obtained from blogs, social networks and online forums. Data obtained from these sources is unorganized and expressed in different contexts and vocabularies. To use manual analysis for such data results in high time complexity and hence slow down the process.
2) *Pre-processing:* Data collected in the previous step is preprocessed. Preprocessing involves removal of non-textual content from the data obtained. Some popular preprocessing steps include removing stop-words and emoticons, removal of special characters and punctuation, removing URLs, stemming, tokenization and feature extraction.
3) *Sentiment identification:* Classification is a technique which is used to classify data into various categories. This step is used in order to classify data into three classes; positive, negative and neutral. These classes are also called opinion orientations, sentiment orientations, semantic orientations or polarities. This step is called as sentiment identification. The sentiment analysis can now be based on classifying the polarity of the text at the document, sentence, or aspect level positive, negative, or neutral.
4) *Presentation of output:* Last step in sentiment analysis is to present the output in various forms. The text results can be displayed using some graphical representation. A sentiment time line can also be generated for a certain period of time with values like frequency and averages.

B. Sentiment Analysis Approaches

The sentiment classification approaches can be classified in: a) Machine Learning b) Lexicon based

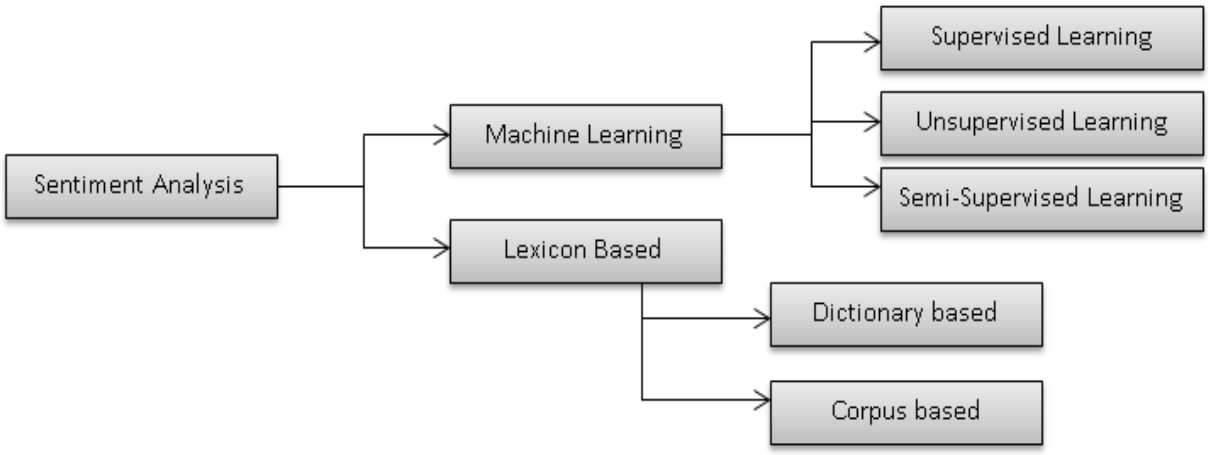


Fig.2 Sentiment Analysis Techniques

The machine learning approach to sentiment analysis deals with algorithms that allow a computer to learn. This means that an algorithm is given a set of data and is allowed to infer information about the properties of the data. This information can be used to make predictions about other data that the algorithm may come across in the future. Machine learning can be performed in a supervised, unsupervised or semi supervised manner [3]. The main advantage of this method is the ability to adapt. Machine learning (ML) techniques do have its weaknesses; since the algorithms vary in their ability to generalize over a large set of patterns and it is likely that a new pattern can be misinterpreted. Also, ML algorithms require labelled data to train. The availability of labelling data and its acquisition can be costly or even prohibitive for certain tasks.

1) *Supervised Learning:* Supervised learning is the process of inferring a function from labelled training data. The training data consist of a set of training examples. Supervised learning is a widely used solution for classification purpose and is been used in most of the sentiment classification techniques. Techniques for sentiment classification include SVM, Neural Network and Decision tree Classifiers. Other commonly used algorithms include K-Nearest Neighbour, Bayesian Network. The Naïve Bayes Classifier is one of the most commonly used classifier for text data. Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes model is easy to build and useful for very large data sets. The method calculates prior probabilities from training data and computes the posterior probability of the document based on the prior probability values. Maximum entropy classifier is a probabilistic classifier which converts labelled feature sets into vectors by using encoding. This vector is used to calculate weights for each feature that can be combined to determine the most likely label for feature set. [5], [6]
   Support vector machine is non probabilistic approach which is used to separate data linearly and nonlinearly. It determines the separators in search space which can best separate the classes. SVM is suited for text data and is used in most of the sentiment analysis techniques. [7], [8]
   A neural network has emerged as an important tool for classification. During past decade neural network classification has established as a promising alternative to various conventional classification methods. The neural network with appropriate network structure can handle the correlation/dependence between input variables.
   Decision tree is a simple and most widely used classification technique. It is a predictive model to go through the observations and to predict the value of final outcome. It defines hierarchical decomposition of training data. Decision tree has been used in many sentiment classification methodologies in recent years.
2) *Unsupervised Learning:* Unsupervised learning is a type of machine learning algorithm which is used to find inferences from datasets consisting of input data without labelled responses. In most of the classification techniques, especially text data it is very difficult to create training labelled data and it requires much of the human effort. Making use of unsupervised techniques can help overcome the disadvantage. This approach when used in Document level sentiment analysis determines the semantic orientation (SO) of specific phrases within the document. If the average SO of these phrases is above some predefined threshold, the document is classified as positive; else it is termed as negative.
   Turnkey [9] uses two arbitrary seed words (poor and excellent) to calculate the semantic orientation of phrases, where the orientation of a phrase is defined as the difference of its association with each of the seed words

3) *Semi Supervised learning:* Semi supervised learning[3] is a class of supervised learning tasks that makes use of unlabelled data for training. The category falls between supervised and unsupervised learning. Many research studies have found out some amount of labelled data along with unlabelled data can produce considerable improvement in learning. In general, there are two kinds of semi-supervised learning approaches. One is to learn with unlabelled data using bootstrap techniques like self-training, Expectation Maximization (EM) and co-training. Another category is structural learning methods which learn good functional structures using unlabelled data. For example, the graph based method which constructs a graph with labelled and unlabelled examples as nodes and their similarity as edges.

A LCCT (Lexicon-based and Corpus-based, Co-Training) model for semi-supervised sentiment classification combines the idea of lexicon-based learning and corpus-based learning in a unified co-training framework. It is capable of incorporating both domain-specific and domain independent knowledge.

The Lexicon based approach [4] to sentiment analysis does not require any prior training to mine data. It simply uses a pre-defined list of words, where each word is associated with a specific sentiment. In this approach the definition of the sentiment is based on the analysis of individual words and/or phrases. The polarity of a piece of text can be obtained from the polarity of the words that compose it. To do this, textual content is broken down into micro phrases, based on the placement of splitting cues. A sentiment is thus the sum of the polarity conveyed by each of the micro phrases. The polarity of each micro phrase depends on the sentimental score of each term in the micro phrase, which is obtained by using a lexical resource. The hybrid based approach to sentiment analysis combines both Machine Learning and Lexicon based approaches.

Lexicon based methods calculate similarity of a document from a semantic orientation of phrases in the document. The classification approach involves creating classifier from labelled instances. The approach makes use of lexicon to perform sentiment analysis by assigning weights to sentiment related words. To extract sentiment related words, two methods are used. A dictionary based approach deals with creating dictionaries with initial seed set of words which can be further expanded. First, a list of adjectives and corresponding SO values is compiled into a dictionary. Then, for any given text, all adjectives are extracted and annotated with their SO value, using the dictionary scores. The SO scores are in turn aggregated into a single score for the text. Another approach is a corpus based approach which uses seed set of sentiment words with known polarity values and then finds other sentiment words in a large corpus to find semantic words with specific context.

C. Sentiment Analysis Tools

Recent advancement in the field of sentiment analysis includes research on the latest tools for sentiment analysis. Most commonly used tools for detecting feelings polarity is based on emoticons. Emoticons are face based and symbolize happy or sad feelings. To extract the polarity of feelings, common emoticons can be used from various sources. These emoticons can be used in combination with latest techniques for data extraction in sentiment analysis which serves as training data in most of the supervised techniques. Happiness Index is a tool that uses affective norms for English words. It provides score for the text from 1 to 9 indicating the amount of happiness. Another tool is SentiWordNet. This tool is used a lexical resource for opinion mining. It is based on the dictionary WordNet. SentiWordNet assigns to each synset of WordNet three sentiment scores: positivity, negativity, objectivity.

The Positive and Negative Affect Schedule (PANAS) consists of two mood scales i.e. positive affect and the other which measures negative affect. Modified version of PANAS-t keeps track of increase or decrease in sentiments over time. This method is based on a large set of words associated with eleven moods. The method calculates the score for each sentiment for a given time period as values between -1.0 to 1.0 to indicate the change.

Septic Net is tool which is used to perform concept level sentiment analysis. Tasks include emotion recognition and polarity detection which considers semantics. It does not focus much on word co-occurrence frequencies. It can be considered as a framework or knowledge base. As a framework Sentic Net consists of a set of tools and techniques for sentiment analysis combining the knowledge from natural language processing and Machine learning approaches. It mainly focuses on semantic preserving representation of natural language concepts. As a knowledge base, sentic net provides set of semantics with many natural language concepts. Semantics are those concepts which are related to semantically to the input concept.

EWGA uses an entropy-weighted genetic algorithm to efficiently select features for sentiment classification which makes use of a wrapper-model, where the performance of a feature subset is used as its fitness function value within the genetic algorithm. Linguistic Inquiry Word Count analyses text files on a word-by-word basis using an internal dictionary of more than 2,300 of the most common words and word stems. LIWC classifies the words into dozens of linguistic and psychological categories that tap social, cognitive and affective processes.

III.ANALYSIS OF SENTIMENT ANALYSIS TECHNIQUES

TABLE I

ANALYSIS OF SENTIMENT ANALYSIS APPROACHES

Sr. No	Title	Proposed Methodology	Problem Addressed	Dataset Used	Future Scope
1	An Unsupervised approach for Sentiment Analysis in Twitter[10]	Unsupervised dependency based approach using a Sentiment Lexicon	i. Contextual Polarity Disambiguation ii.Message Polarity Classification	SemEval-2015 task Organizer texts extracted from Twitter	-
2	An Unsupervised Approach for feature based Sentiment analysis of Product Reviews[11]	Analysis of Product Reviews at feature level and Combined use of Coreference Resolution	i. Coreference resolution	Camera dataset from Amazon.com	i. Mining of Polarity of Implicit features ii.Analysis of Comparative statements
3	Sentiment Analysis of Movie Reviews[12]	Feature based heuristic for aspect level Sentiment analysis	Accuracy and Polarity Shift	Movie Review dataset	-
4	Topic Modelling based Sentiment analysis on Social Media for Stock Market Analysis[13]	Topic Model TSLDA	Accuracy	2 Datasets i. Historical Price ii.Message board dataset	Extension as non parametric topic Model estimating the number of topics inherent in data
5	Twitter Sentiment analysis using Deep Convolutional Neural Network[14]	Sentiment analysis using deep convolutional neural network	Accuracy	Test Set from SemEval 2015	Investigation using First Linear Layer
6	A Probabilistic approach to Tweets Sentiment Classification[15]	Weighted Word Pairs by using Latent Dirichlet	Accuracy	Movie Review dataset	Use of Sentiwordnet for better evaluation of words

[10]Presents an unsupervised approach for sentiment analysis in Twitter. It followed unsupervised dependency parsing approach using lexicon which is created by means of an automatic polarity expansion and natural language processing techniques .Some of the subtasks which are considered in the work include Contextual Polarity Disambiguation which determines the polarity of marked instance of a word. A feature based sentiment analysis for product reviews is provided in [11].This method focuses on combined use of coreference resolution, domain specific aspect dictionary, SentiWordNet, linguistic rules, adjectives, verbs, adverb adjective combinations and adverb verb combinations together for sentiment analysis of features.[12]proposed an aspect oriented approach that analyses the textual reviews of a movie and assigns a sentiment label. The scores on each aspect from multiple

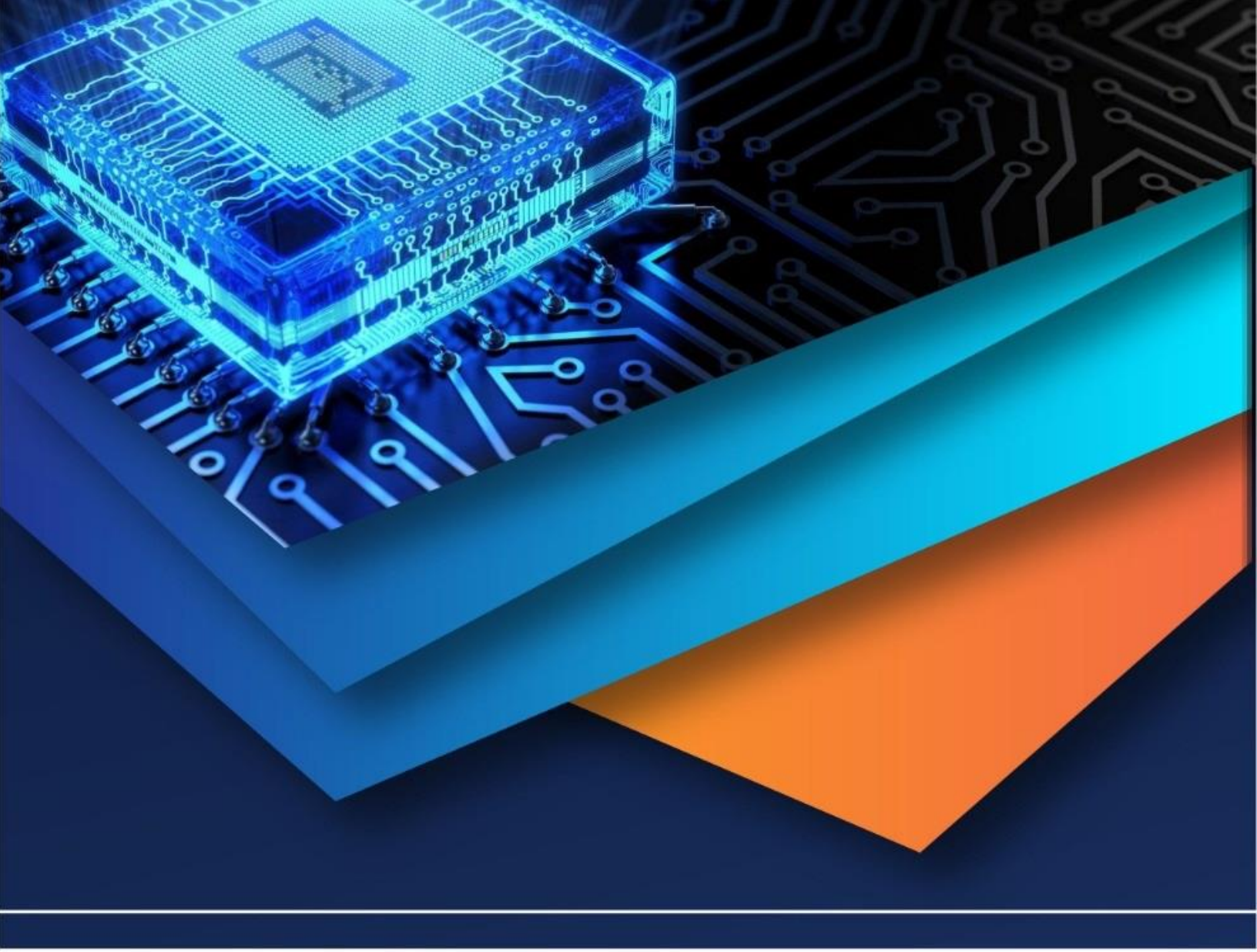
reviews are then aggregated and a net sentiment profile of the movie is generated based on certain parameters. Method makes use of Sentiwordnet with different linguistic feature selection consisting of adjectives, adverbs and verbs. A new prediction model is proposed which captures topics and sentiments simultaneously. In addition, a new model TSLDA is proposed to obtain this feature[13]. [14] describes a deep learning system for sentiment analysis of tweets. A new model for initializing the parameter weights is proposed. An unsupervised neural language model is used to train the initial word embedding's which are further tuned by the new proposed deep learning model. An approach to sentiment analysis based in weighted word pairs by the use of LDA are proposed in [15].It derives a word based graphical model for mining a positive or negative opinion towards a topic.

IV.CONCLUSION

Sentiment analysis use information from sources such as forums, microblogs, forums and news sources. The information gathered from these sources plays a very important role in expressing people opinion about a particular product. A deeper analysis of such data is required in order to extract important knowledge. The detailed study presented in this paper is based on machine learning techniques and their applications in well known domains. Major categories focus on tasks like subjectivity classification, sentiment classification and opinion spam detection. Experimental result from other researches show how different Sentiment analysis algorithms behave when applied on various sets of data. Considering the tools used for sentiments analysis, the most used tools for detecting the feelings polarity (negative and positive affect) are discussed in the paper: Emoticons, LIWC, SentiStrengh, Senti WordNet, SenticNet. Future work in sentiment analysis includes handling coreference information.

REFERENCES

[1] B Liu, L Zhang, (2012) A survey of opinion mining and sentiment analysis - Mining text data, Springer US, 415-463.
[2] K Schouten, F Fransincar, Survey on Aspect – Level Sentiment Analysis, IEEE transactions on knowledge and data engineering, vol. 28, no. 3. 2016
[3] B Yang, Semi Supervised Learning for Sentiment Classification, [online] www.semanticscholar.org
[4] C Musto, G Semeraro, M Polignano, A Comparison of Lexicon based approaches for sentiment analysis of Microblog, CEUR Workshop Proceedings, 1314, pp. 59 – 68, 2014
[5] X. Bai, "Predicting consumer sentiments from online text", Decision Support Systems, vol. 50, no. 4, pp. 732–742, 2011
[6] M.-T. Martín-Valdivia, E. Martínez-Cámara, J.-M. Perea-Ortega, and L. A. Ureña-López, "Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches", Expert Systems with Applications, vol. 40, no. 10, pp. 3934–3942, 2013.
[7] P. C. R. Lane, D. Clarke, and P. Hender, "On developing robust models for favourability analysis: Model choice, feature sets and imbalanced data," Decision Support Systems, vol. 53, no. 4, pp. 712–718, 2012.
[8] S. E. Seker, K Al-Naami, "Sentimental Analysis on Turkish Blogs via Ensemble Classifier", Proceedings the International Conference on Data Mining, 2013.
[9] Turney, P. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th ACL, pp. 417-424.
[10] Milagros Fernandez-Gavilanes, Tamara ´ Alvarez-L ´ opez, Jonathan Juncal-Mart ´ ´ınez, Enrique Costa-Montenegro, Francisco Javier Gonzalez-Casta ´ no," GTI: An Unsupervised Approach for Sentiment Analysis in Twitter" in Conference: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)
[11] Sherin Molly Babu , Shine N Das," An Unsupervised Approach for Feature Based Sentiment Analysis of Product Reviews", International Journal of Scientific Research Engineering & Technology (IJSRET), ISSN 2278 – 0882 Volume 4, Issue 5, May 2015.
[12] V. K. Singh , R. Piryani, A. Uddin, P. Waila ," Sentiment analysis of movie reviews", Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), 2013 International Multi-Conference.
[13] Thien Hai Nguyen Kiyoaki Shirai," Topic Modeling based Sentiment Analysis on Social Media for Stock Market Prediction", Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 1354–1364,Beijing, China
[14] Aliaksei Severyn, Alessandro Moschitti," Twitter Sentiment Analysis with Deep Convolutional Neural Networks", SIGIR '15 Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval Pages 959-962
[15] Francesco Colace , Massimo De Santo, Luca Greco," A probabilistic approach to Tweets' Sentiment Classification", Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY