# A Survey on Sentiment Analysis Data sets and Techniques

Vidula Dattatray Bhat.
Dept. of Information Technology
Maharashtra Institute of Technology
College of Engineering
Pune India

Vivek S. Deshpande
Dept. of Information Techology
Maharashtra Institute of Technology
College of Engineering
Pune India

Rekha Sugandhi
Dept. of Computer Science
Maharashtra Institute of Technology
College of Engineering
Pune India

***Abstract*- This survey reviews the recent progress in the field of sentiment analysis with the focus on available datasets and sentiment analysis techniques. Since many exhaustive surveys on sentiment analysis of text input are available, this survey briefly summarizes text analysis techniques and focuses on the analysis of audio, video and multimodal input. This survey also discuses different available datasets. In most of the work datasets are prepared as per specific research requirements. This survey also discusses methods used to prepare such datasets. This survey gives an overview of available datasets, methods to prepare data sets, sentiment analysis techniques, and challenges in this area.**

***Key words*- *Sentiment Analysis, Opinion Mining, Multimodal Sentiment Analysis, datasets, sentiment analysis***

## I. INTRODUCTION

Opinions always play an important role in decision making. Businesses seek consumer opinions about their products and services to improvise them, consumers seek opinions of other consumers to get the best deal. Governors and policy makers of the country also need to consider opinions of the masses before making some decisions. Emergence of social networking sites, blogs, forums, e-commerce websites have provided internet users with a platform where they can express their opinions. Thus a huge source of opinions, views, and sentiments has been created and is being updated every day. The available data is huge and unstructured. Since manual analysis of the tremendous volume of data is not possible, the process needs to be automated. The field of Sentiment Analysis (SA) or Opinion Mining (OM) has emerged from this need.

Sentiment Analysis (SA) can be applied in consumer products and services, financial services, social events, elections etc. Apart from commercial use researchers have also explored the possibility of applying SA for detecting personality traits [4] detecting leadership traits [3], detecting Child's interest in the lecture [14], etc. Because of the wide range of applications many small and large companies have started providing Opinion Mining (OM) and Sentiment Analysis (SA) tools and solutions.

The field of Sentiment Analysis is relatively new. The pioneering work in this field includes papers by Turney [16] and Pang et. al. [17] published in 2002. Since SA is still in its nascent stage there are many open problems which constitute research areas in this field.

This paper aims at reviewing the literature in the field of SA with the focus on available training data sets and classification algorithms. The paper is organized as follows- Section 2 contains a review of the data sets which are usually used for sentiment analysis.

## II. DATASETS

### *A. Movie review and other review sites*

In most of the work on sentiment analysis movie review dataset is used. Large movie review dataset is available on [23]. It contains 25000 training and 25000 testing movie reviews. In addition this data set also provides untagged data. Andrew L. Mass et. al have used this dataset in their work [5]. Movie review data available on [24] includes various versions of sentiment polarity data, sentiment scale data and sentiment subjectivity data. The data set of amazon product reviews is available on [25]. The data span a period of 18 years, including ~35 million reviews up to March 2013. Reviews include product and user information, ratings, and a plaintext review.

### *B. Blogs*

Blogs are different from reviews in structure and layout. Blogs may also contain comments which express sentiments which are contradictory to those in the blog. To analyze blog posts it is required to extract core content from the blogs. Melville et. al. [6] used the algorithm provided [7**]** to extract text only from parts of the Web-page where the ratio of HTML tags to words is above a minimal threshold.

Kale et. al. [8] described techniques to find "like minded" blogs based on blog-to-blog link sentiment for a particular domain. They used a graph of 300 blogs created from the link structure of **NielsenBuzzmetric** dataset available on [26]**.** Srinivasan Ramaswamy [9] has given a method of corpus preparation where a python module viz. Universal Feed Parser is used to parse syndicated fields.

### *C. Social Networking site posts*

Users of social networking sites like Facebook, Twitter post comments on daily events and numerous topics on the sites. Most of the work in the field has been done on Twitter datasets. Sentiment140 available on [27] is twitter datasets available for researchers. The data is automatically extracted by using positive and negative emoticons using Twitter Search API. The data consists of six fields- polarity, tweet id, tweet date, the query, user name, and text of the tweet. A facebook dataset is available on [28].

This corpus consisting of 10,000 manually annotated Facebook posts. 2,587 positive, 5,174 neutral, 1,991 negative and 248 bipolar posts are tagged. The file format is Excel.

### *D. Online Videos*

Rosas et. al. have used 105 online Spanish videos in their work [1]. The videos are collected from Youtube using search terms like 'my opinion', 'my favorite products', 'I like', 'I dislike' etc . Videos collected were converted into mp4 format. Length of videos varies between 2 to 8 minutes. These videos were preprocessed to overcome two issues- cut the introductory title and maintain the same subject of discussion in a clip. The methodology used in this paper [1] is recent and provides a guideline for beginners in the field. Videos contain text, images and audio together thus they make an ideal input for multimodal sentiment analysis system.

## III. SENTIMENT ANALYSIS TECHNIQUES

### *A. Overview*

Sentiment Analysis aims at classifying data into sentiment polarities, scales or subjects. In review analysis data is usually classified between three classes- Positive, Neutral and Negative. Data is also classified in six basic emotions- Happiness, Surprise, Sadness, Fear, Disgust and Anger. Researchers have also classified data into non-basic emotions like- boredom, anxiety, frustration etc. The cues for detecting sentiments can be either linguistic i.e. text input or non-linguistic i.e. image, audio or video input. Multimodal Sentiment Analysis is an active research field in which more than two modes of input are combined. There are three ways to combine the inputs – Data fusion, feature fusion and decision fusion. In this survey we give a brief summary of previous work on text analysis and mainly focus on the methods used for Sentiment Analysis using non-linguistic cues and the methods used for Multimodal SA.

### *B. Sentiment Analysis of Text Input*

The classification techniques used for topic based classification can also be applied to sentiment classification. The techniques are mainly classified into two classes- Machine Learning Approach and Lexicon Based Approach.

Machine Learning Approach uses data inputs to construct a model which is then used for decision making. Two major categories of machine learning algorithms are – Supervised and Unsupervised machine learning. In supervised machine learning data is labeled and aim of the algorithm is to find the mapping between the data and its label. In unsupervised learning labels are not provided and the algorithm aims at finding out the hidden pattern in the data. Major work in Sentiment Analysis has been done using supervised learning algorithms. Supervised learning algorithms can be classified into four categories - Decision tree classifiers, Rule based classifiers, Linear Classifiers and Probabilistic classifiers.

Pang et. al. [17] have compared three standard text classification algorithms – Naïve Bayes (NB), Maximum Entropy (ME) and Support Vector Machine (SVM) in their pioneering work [5] in the field of SA. Out of which Naïve Bayes and Maximum Entropy are probabilistic classifiers and Support Vector Machine is a linear classifier.

Following are some important conclusions of their work [17].

- Sentiment Classification is difficult than topic based classification.
- Considering only feature presence and not feature frequency yields better performance in sentiment classification. This is in contrast with topic-based classification where feature frequency is more important than feature presence.
- Only Bigrams are not effective to capture the context.
- Applying direct feature selection algorithm on unigrams performs better than using adjective words.
- Algorithms perform slightly well when term position information is considered.

Pang et. al. [17] compared three algorithms- Naïve Bayes, Support Vector Machine and Maximum Entropy on eight features. Support Vector Machine algorithm performed better for five out of eight features. SVM has been an algorithm of choice in most of the recent work on SA.

Opinion words or phrases are called lexicons and are used for classification of data into positive, negative or neutral classes. Depending on how the positive or negative wordlists are created lexicon based approach can be divided into following categories – manual approach, dictionary based approach, and corpus based approach

Hu Mining et. al. and Kim S. et. al. gave the workflow of dictionary based approach in [10] and [11]. First, opinion words with known orientations are manually selected. Then WordNet or thesaurus is used to expand the seed lists. The iterative process stops when no more new words can be found. The Corpus-based methods use syntactic patterns or co-occurring patterns with a seed list of opinion words in order to find other opinion words in a large corpus.

### *C. Sentiment Analysis of Non-textual Input*

Use of images, audios, videos for expressing emotions, views, and opinions is increasing. Thus opinion mining from non-textual sources has become important.

Exhaustive surveys on classification methods used for facial expression recognition are given in [2], [22]. The basic stages in facial expression detection are – face acquisition, facial expression extraction and facial expression detection. Friesen and Ekman [20] proposed the emotional facial action coding system (EFACS). EFACS provides the sets of AUs which are sets of specific facial muscle movements and consist of three parts: AU number, FACS name, and muscular basis. E.g. AU number 1 is the inner brow raiser in which frontalis; pars medialis muscle movements are used.

Most of the research has been done on six basic emotions. The study of non-basic emotion is domain and application specific e.g. study of boredom in gaming.

In many studies difference between spontaneous vs. non- spontaneous facial expressions is studied. Cohn et al. [15] showed that spontaneous smiles are smaller in amplitude, longer in duration and have long onset and offset times.

Acoustic features such as fundamental frequency (pitch), intensity of utterance, bandwidth, and duration are used for sentiment analysis using audio cues. There are two approaches for speech recognition – speaker dependent approach and speaker independent approach. Speaker dependent system is trained on a particular speaker's voice characteristics. Such systems can recognize speech in a variety of contexts but cannot recognize speech from multiple users. Vice versa speaker independent syste1ms are limited in context but can recognize speech from multiple users. Navas et al. [33] shown that the speaker-dependent approach gives better results than the speaker-independent approach where Gaussian mixture model (GMM) as a classifier 98% accuracy was achieved with prosodic, voice quality as well as Mel frequency cepstral coefficients (MFCC) used as speech features. In speaker-independent applications, the best classification accuracy achieved so far is 81% [3], obtained on the Berlin Database of Emotional Speech (BDES) with a two-step classification approach and a unique set of spectral, prosodic, and voice features, selected through the Sequential Floating Forward Selection (SFFS) algorithm.

### *D. Multimodal Sentiment Analysis*

Large number of videos is being uploaded online everyday. Videos contain text, visual and audio features which complement each other. Multimodal Sentiment Analysis refers to the combination of two or more input modes to improve the performance of the analysis. Combination of text and audio-visual inputs is an example of multimodal sentiment analysis.

The combination or fusion can be done in three ways – Data fusion, Feature fusion and Decision fusion. Decision fusion is used in most of the work. In feature level fusion a joint feature vector is created by combination of separate input features. A typical example of future level fusion is [13**]**, which combines prosodic and facial expression features. Multimodal sentiment analysis is an emerging research field. Rosas et. al. have shown in their recent work [1] that combination of input modes improves the accuracy of the analysis.

## IV. DISCUSSION

In this survey we have studied sentiment analysis of textual and non-textual input separately with the focus on available data sets and analysis techniques.

Our scope in this survey is limited to English language. Many annotated data sets of twitter posts, movie reviews etc. are available for text input. Many domain specific corpora are available for text input analysis. The annotated data sets for images, audio and video inputs across multiple domains are comparatively scarce. In most of the work research specific data sets are created and used.

In text analysis methods, supervised algorithms are used in most of the work. Three highly used algorithms are Support Vector Machine (SVM), Naïve Bayes (NB) and Maximum Entropy (ME).

SVM can be used for diverse information sources; it is sensitive to sparse data. Naïve Bayes (NB) usually performs well with small data sets.

In case of sentiment analysis from speech 98% classification accuracy is achieved using speaker dependent approaches where as 81% classification accuracy is achieved so far using speaker independent approaches.

The problem of detecting hidden emotions like sarcasm or irony has always challenged the researchers in the field. These emotions are called hidden since they are not directly expressed in the text. Human beings percept these emotions by using two cues – non verbal communication and context. The same cues can be used to detect hidden emotions using multimodal sentiment analysis for detecting the non verbal cues and using context. Multimodal sentiment analysis is an emerging research area. Most of the research up till now has been successful in proving that a combination of two or three input modes improves the accuracy of the analysis. The role of context in sentiment analysis has been explored and it has been proved that, context improves the classification accuracy. Combining multimodal cues and context with the focus on detecting hidden emotions is a probable future direction of research.

## REFERENCES

[1] Ver´onica P´erez Rosas, Rada Mihalcea, Louis-Philippe Morency, "Multimodal Sentiment Analysis of Spanish Online Videos", IEEE Intelligent Systems, 2013

[2] Shangfei Wang et. al., Analyses of a Multimodal Spontaneous Facial Expression Database, IEEE transactions on affective computing 2013, vol. 4, no. 1

[3] Felix Weninger, Jarek Krajewski, Anton Batliner, and Bjorn Schuller, "The Voice of Leadership: Models and Performances of Automatic Analysis in Online Speeches", IEEE Transactions on Affective Computing, Vol. 3, No. 4, October-December 2012

[4] Gelareh Mohammadi and Alessandro Vinciarelli, "Automatic Personality Perception: Prediction of Trait Attribution Based on Prosodic Features", IEEE Transactions on Affective Computing, Vol. 3, No. 3, July-September 2012

[5] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts, "Learning Word Vectors for Sentiment Analysis" The 49th Annual Meeting of the Association for Computational Linguistics, 2011

[6] Prem Melville, Wojciech Gryc, Richard D. Lawrence, "Sentiment Analysis of Blogs by Combining Lexical

Knowledge with Text Classification", ACM International Conference on Knowledge Discovery and Data Mining, 2009

[7] Extracting the main content from a webpage. http://wshadow.com/blog/2008/01/25/extractingthe-main-content-from-a-webpage/, 2008

[8] Anubhav Kale, Amit Karandikar, Pranam Kolari, Akshay Java, Tim Finin, Anupam Joshi, "Modeling Trust and Influence in the Blogosphere Using Link Polarity", International Conference on Weblogs and Social Media 2007

[9] Shrinivas Ramaswamy, "Blog Analysis - Trends and Predictions", Applied Natural Language Processing Project Report,http://courses.ischool.berkeley.edu/i256/f06/projects/ramaswamy.pdf, 2006

[10]Hu Minging, Liu Bing, "Mining and summarizing customer reviews" Proceedings of ACM SIGKDD international conference on Knowledge Discovery and Data Mining, 2004.

[11]Kim S, Hovy E. "Determining the sentiment of opinions", In: Proceedings of interntional conference on Computational Linguistics, 2004.

[12]Vinay Kumar Bettadapura, "Face Expression Recognition and Analysis:The State of the Art", Computer Vision and Pattern Recognition, arXiv:1203.6722 [cs.CV], 2012

[13]Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, "Ana[7] B. Fasel, Juergen Luettin, "Automatic facial expression analysis: a survey", "Pattern Recognition", Volume 36, Issue 1, 2006

[14]A. Kapoor, R.W. Picard, and Y. Ivanov, "Probabilistic Combination of Multiple Modalities to Detect Interest," Proc. 17th Int'l Conf. Pattern Recognition, vol. 3, pp. 969-972, 2004.

[15]Cohn, J.F. and Schmidt, K.L.(2004), "The Timings of Facial Motions in Posed and Spontaneous Smiles", International Journal of Wavelets, Multiresolution and Information Processing, 2004

[16]Peter Turney, "Thumbs Up or Thumbs Down: Semantic Orientation Applied to Unsupervised Classification of Review", In Proceedings of ACL, 2002

[17]Bo Pang, Lillian Lee, Shivkumar Vaithyaanathan, "Thumbs up?: sentiment classification using machine learning techniques.", EMNLP '02 Proceedings of the ACL-2002

[18]Namrata Godbole, Manjunath Srinivasaiah, Steven Skeina, "Large-Scale Sentiment Analysis Mikhail Bautin, Lohit Vijayrenu, Steven Skeina,

[19]Kinam Park, "Automatic extraction of user's search intention from web search logs", Multimedia Tools App DOI 10.1007/s11042-010-0723-8

[20]Paul Ekman, Wallece Friesen, "Journal of Personality and Social Psycology, Vol 39 No. 61980", 2000

[21]Martin .J, "Blogging for dollars", Fortune Small Business, 15(10), 88–92, 2005

[22]Maja Pantic and Leon J.M. Rothkrantz, "Automatic Analysis of Facial Expressions: The State of the Art", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 12, December 2000 lysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information", International Conference on Multimodal Interfaces, 2004

[23]Large Movie Review Data set by Stanford http://ai.stanford.edu/~amaas/data/sentiment/

[24] Movie Review Data Set by Cornell http://www.cs.cornell.edu/people/pabo/movie-review-data/

[25] Amazon product review data set http://snap.stanford.edu/data/web-Amazon.html.

[26]NielsenBuzzmetric Weblog dataset http://www.icwsm.orgss/data.html