AN OVERVIEW OF SENTIMENT ANALYSIS: APPROACHES AND APPLICATIONS

Sunayana Bhandari | Dr. Subhajit Ghosh

[1](*School of Computing Science, Galgotias University, Greater Noida, U.P., India, sunayana.bhandari7@gmail.com*)

[2](*School of ComputingScience,Galgotias University, Greater Noida, U.P.India,subhajit.ghosh@galgotiasuniversity.edu.in)*

Abstract—*Sentiment analysis is a recent area of research that deals with interpreting user sentiments in web articles, tweets, blog post, product review and news reports. It divides the data based on its polarity i.e. positive, negative or neutral. These sentiments are used by organizations to understand user point of views and improve business performance. This survey paper highlights the fundamentals of sentiment analysis, various sentiment analysis approaches and methodologies developed and used so far; and its various areas of applications. It compares sentiment analysis with certain other data analysis techniques.*

Keywords— Sentiment Analysis; Supervised sentiment analysis; Semi supervised sentiment analysis; Unsupervised sentiment analysis; Coarse grained Sentiment Analysis; Fine grained sentiment analysis.

1. INTRODUCTION

The internet contains a large collection of user generated data in form of social media such as blog post, news articles, twitter data content, Face book data, product reviews on online shopping sites. Due to open-ended format, easy availability and open access, these data sources play a vital role in participant interaction and exchange of user opinions, experience as well as factual information regarding any product or service that in turn highly influences the customer decision making. Therefore, these user point-of-views are used by manufacturers, companies, small and large business owners and organizations to assess customer feedback and review, evaluate the market response and study user complaints, preferences, liking, favors to provide improved product and better customer assistance and support and directly increase its business profits. Service providers use this information to understand customer needs and provide better services and enhance its market presence. The key issue with web based data is its unstructured and semi-structured form along with ambiguity and ubiquity in content nature. The data being wide spread across the internet and present in different formats makes it difficult and time consuming to be manually analyzed.

Sentiment analysis is an analytical tool based on text analysis used to classify the user data into categories based on content polarity, i.e. positive, negative or neutral through data preprocessing, natural language processing and various classification techniques aggregated over a large data set. Semi-automated tools are designed whether using supervised learning through training set and test sets or unsupervised using lexicon comparison approach. New research is being focused on a more fine-grained aspect based sentiment analysis approach.

In addition to product/service based sentiment evaluation, sentiment analysis approach is also used in various other areas such as in identification of cyber threats and radical conflicts; by political parties to use online voter responds to structure their campaign plan and predict voting results; by administrative organizations for compiling user complains and suggestions provided online; by media using social media trends to evaluate and predict public sentiments and opinions; and in financial analysis to predict stock pricing.

Measure of effectiveness of any sentiment analysis approach is done through four basic measures namely Precision (P), Recall (R), Accuracy (A) and Averaged function F_1[1]. Two averaging techniques used over data set are micro-averaging and macro-averaging [1].

2. COARSE-GRAINED SENTIMENT ANALYSIS

Coarse grained analysis is aimed towards finding document level or sentence level polarity.

A. Supervised learning sentiment analysis

Supervised learning deals with machine learning based classification. It consists of two types of data sets [1]. 1) A training set that used automated or semi-automated classifiers to distinguish between the various characteristic features present in the given documents set. 2) A test set used to validate the correctness of the classifier. Various supervised sentiment learning approaches developed as Bayesian models based on the Poisson and Negative-Binomial distributions for high-frequency sentiment words [2], a two-stage Markov Blanket Classifier (MBC) to capture conditional dependencies amongst words rather than using only keywords or high-frequency words [3]. A combined supervised sentiment analysis approach has been proposed combining rule based classification (RBC) with statistics based classifier (SBC) to form inductive rule based algorithm using induction algorithms ID3 [4] and RIPPER [5]. Used with SVM^{light} [6] and four different datasets, the best results were indicated by micro-averaged F_1 as 90.00% and macro-averaged F_2 as 89.98% while using RBC, induced SBC using ID3, GIBC and SVM in sequence [1].

A multi-swarm particle swarm optimization (MSPSO) method [7] has been created to select emotional features in online course reviews. Multi diverse particle swarms were generated on several cross training subsets used to extract the best discriminative features by the F-Measure fitness function with removed feature redundancy. The dataset collected from Massive Open Online Course (MOOC) platform had been used with variable length feature extraction done using n-gram (n=1, 2). The proposed method when compared against IG, MI, CHI square statistic, genetic algorithm (GA) and single swarm PSO method (SSPSO) showed MSPSO obtaining over 88% micro-F-measure for a reduced 3000 feature set [7].

B. Semi-supervised sentiment analysis

Semi-supervised approaches uses features of both machine learning based classification as well of part-of-speech analysis for determining sentiment polarity. Various approaches for semi-supervised sentiment analysis have been proposed including quantitative analysis of subjective term definitions [8] and also on some on-line dictionaries [3]. A sentiment analysis technique has been proposed that using 3-gram feature to capture local context and improvement over PV-DM [9] for integrating distributed semantic feature of word sequence along with part-of-speech (POS) sequence for capturing global context in a text data. The comparative result over a large monolingual corpus as input data set provides 93.18% accuracy over 92.10% accuracy from using PV-DM with POS sequence feature and 91.92% accuracy from using PV-DM without POS sequence feature[10].

C. Unsupervised sentiment analysis

Unsupervised learning methodology doesn’t require prior training to be given to the system. It studies semantic orientation that is the inclination of the words towards positive or negative sentiment.

Unsupervised approaches are usually lexicon based classification determining word polarity based on existing WordNet. or new created lexicons.

3. FINE-GRAINED SENTIMENT ANALYSIS

Fine-grained aspect (or feature-based) sentiment analysis as opposed to coarse-grained sentiment analysis includes identifying the target entity of any opinion or sentiment expressed in addition to its polarity and intensity. Fine-grained analysis is important as context dependent opinions as well as context indistinct-dependent opinions [11] need special methods for context determination. There have been fine-grained analysis models based on hidden conditional random fields (HCRFs) [12] that provide sentence-level fine-grained analysis for document-level coarse-grained analysis. An HCRF based model with sentence-levels analysis derived solely from document-level supervised approach using labels as form of latent variables have shown to reduce sentence classification error by 13% and 22% when compared with machine-based and lexicon-based models respectively [13]. Another proposed set of models infusing fully supervised and partially structured conditional models [14] using cascading and interpolation approach are tested against the baseline HCRF based models.

Fine-grained analysis had been done on news articles using novel sentiment annotation scheme taking into account both explicit and implicit sentence expressions [15]. The results were compared against two coarse-grained analysis approaches: lexicon-based approach and supervised machine learning based on bag-of-words and sentiment lexicon feature approach also manually annotated gold standard labels. It concluded that fine-grain analysis performs reliably when taking into account only the implicit expressions but outperforms coarse-grain approaches in case of both explicit and implicit expressions by being in perfect agreement to gold standard labels.

4. TOPIC SENTIMENT ANALYSIS

Topic-sentiment’ analysis has been conducted on twitter data for hashtag analysis based on novel graph model [16] to recognize sentiment polarity for hashtag containing tweets and co-occurrence relationship between hash tags and incorporating the literal hashtag meanings as semi-supervised learning set.

Similarly, it has also been conducted over social media feedback for companies in order to develop a stock prediction model using Latent Dirichlet Allocation (LSA) model, joint sentiments/topic model (JST) and Aspect-based sentiment model to extract topic sentiment features[17]. The resultant model provided 2.07% better accuracy over predictions based on pricing history only and 3.03% better accuracy over predictions based on human sentiment methods.

5. TERM WEIGHING SCHEMES

Delta TFIDF was developed as a general purpose technique to weigh word score [18]. It was used to study on Pang and Lee’s movie review, subjectivity detection and congressional debate transcripts as data-sets. The results were compared against standard bag of unigram and bigram words represented using 10 fold cross validation and t-tailed tests. Used with support vector machine, it outperformed term counts and TFIDF baseline, with accuracy of 88.1%, 91.26% and 72.47% for sentiment polarity classification, subjectivity detection on movie review and text segment classification for congressional debate transcripts, respectively

Term weighing schemes had been proposed based on class distribution of certain terms over the whole document and class document set to provide positive discrimination based on term frequency. Proposed schemes outperformed in terms of accuracy against traditional term weighing schemes such as tfidf, tficf, MI, OR, WLLR and CHI with single classifiers such as with SVM, PNN, GMM and combined multiple classifiers based on simple voting approach and Borda count approach. The term weighing scheme based class density relative to all class documents had slightly better performance as compared to others [19].

6. DATA PREPROCESSING

Many data preprocessing techniques have been used to make sentiment analysis easier such as sentence compression using Sent_Comp to remove redundancy and non-sentiment bearing words keeping the sentence polarity intact and phrases by applying discriminative conditional random field model (CRF-model) and generating shorter sentences that are easy to parse on product review[20]. When used as an automated step for data compression followed by classic sentiment classification methods from Pang [12] and Mohammad [21] it generated accuracy of 87.95% and 90.67% respectively which was lesser but compared to accuracy measure 88.78% and 91.43% for system without using Sent_Comp for compression. But Sent_Comp had been found relatively more effective for aspect-based sentiment analysis task that rely heavily on syntactic features.

7. CROSS - LINGUAL SENTIMENT ANALYSIS

Cross-lingual sentiment classification is sentiment analysis on corpus in languages other than English.

Automatic machine translation services are usually used for projection of corpus from one language to English since most work in the area is based on English language. Cross-lingual classification has limited accuracy due to difference in term distribution across the translated document when compared to targeted document.

A proposed cross-lingual sentiment model named density based active self learning (DBASL) used the combination of uncertainty-based active learning by selecting an unlabelled example with maximum entropy and density for labeling, and semi supervised self-training approaches making use of unlabelled sentiment data from the target language. [22] Using English-French, English-Chinese and English-Japanese book review data set, the method depicted highest accuracy 78.63%, 71.36% and 70.04% respectively when compared against baseline models such as Active self-training (AST) model, Active learning with uncertainty sampling (AL) model, Self-training (ST) model, Structural correspondence learning model (SCL), Random sampling (RS) model, and Support vector machine with machine translation (SVM-MT).

8. SENTIMENT ANALYSIS VERSUS STAR RATINGS

Another approach for classifying web content based on polarity is star rating approach. Star rating is used to rate any product/service based on its favorableness usually on a scale of 1 to 5, where 1 indicates negative sentiments whereas 5 indicates positive sentiments of user towards the rated content. The ratings provide a basis for user to select the more favorable product/ service. When compared to sentiment analysis, these rating could only be provided in case of well- structured data such as online product feedback, movie reviews, reviewing blogs, etc., other forms of social media such as descriptive blogs, tweets, news reports, etc lack the presence of star ratings. Sentiment analysis on the other hand includes all forms of unstructured or semi-structured data and categorizes the sentiments present in them. Comparative studies have been done on to establish sentiment analysis scores as an alternative to star ratings as well as a surrogate to star ratings where such ratings are not present. Based on semantic proximity studied between star ratings and sentiments evaluated for product comments, the comparison had been done by chi square analysis and two tail bivariate correlation analysis using SPSS for product review, hotel review and doctor reviews taken from amazon, tripadvisor and rateMDs, respectively. The analysis concluded that sentiment analysis score could be used as an alternative where star rating is not present as the normalized sentiment analysis mean is close to star rating mean except for explicit ratings where due to presence of more neutral language in review comments, the sentiment analysis score had limited ability [23].

9. APPLICATION FOR SENTIMENT ANALYSIS

Apart from sentiment analysis being used in product/service reviews and helping businesses study customer feedback, many other domains also use sentiment analysis as a major data evaluation scheme.

A. Cyber security domain:

Bilingual sentiment Analysis Lexicon (BiSAL) was developed to identify words bearing radical sentiments and cyber threats with the help of 2 sentiment lexicon for English language (SentiLEN) and Arabic language (SentiLAR) [24]. The method proposed included seed identification, morphological variants identification, and sentiment score determination through available sentiment corpora AFFIN, SentiWordNet, General Inquirer and SentiStreng for English and a semi-automated analysis by Arabic language experts for Arabic words. The resultant is the complete SentiLEN data set including 279 root words and a complete SentiLAR data set including 1019 root words.

B. Stock market predictions

Stock prediction models could be made combining stock pricing history and social media sentiment analysis, as stock pricing of any company highly depends on customer feedback heavily present on social media. The fine-grain sentiment analysis on news articles contained input data as news articles from Belgian financial newspaper De Tijd (May 2012) concerning four companies: KBC, Delhaize, AB InBev and Belgacom and evaluating their performance [15].

The ‘topic-sentiment’ approach was evaluated on a large scale with taking into account 18 stocks over one year transaction to predict stock pricing taking into account sentiments depicted on social media about the companies under consideration [8].

C. Measurement of people subjective well-being (SWB)

SWB is defined as the way a person evaluates his/her own life, including emotional experiences of pleasure versus pain in response to specific events and cognitive evaluations of what a person considers a good life [25,26]. A text based sentiment analysis had been done to measure the subjective well-being of Chinese people following the classic PANAS framework in psychology. A new lexicon namely Ren-CECps-SWB 2.0 was developed that would provide weight to eight basic emotions through online questionnaires and Delphi method. It was tested on 7 years data of grassroots blogs on Sina.com and compared against the SWB measurement through FGNH (Facebook's Gross National Happiness) indexes and Dodds and Danforth’s SWB measurement based on economic utility theory [27]

10. CONCLUSION

The wide ranging application of sentiment analysis necessitates a comprehensive review of the approaches that have developed in the area. The paper is an attempt in that direction.

REFERENCES

[1] Rudy Prabowo, and Mike Thelwall, “Sentiment analysis: A combined approach.”, Journal of Informetrics 3 (2009) 143–157

[2] Edoardo Airoldi, E., Cohen, and W., Fienberg, “S.: Bayesian models for frequent terms in text” (manuscript, 2005)

[3] Edoardo Airoldi, Bai, and R. Padman, “Markov blankets and meta-heuristic search: Sentiment extraction from unstructured text,” Lecture Notes in Computer Science, vol. 3932, pp. 167–187, 2006.

[4] J. R. Quinlan, “Induction of decision trees.”, Machine Learning, 1, (1986). 81

[5] W. W. Cohen, “Fast effective rule induction.”, A. Prieditis & S. Russell (Eds.), Proceedings of the 12th international conference on machine learning (ICML 1995), July 9–12, 1995 (pp. 115 123). Tahoe City, California, USA.

[6] T. Joachims, “Making large-scale SVM learning practical.”, B. Sch¨olkopf, C. J. C. Burges, & A. J. Smola (Eds.), Advances in kernel methods: support vector learning. (1998). The MIT Press.

[7] Zhi Liu, Sanya Liu, Lin Liu, Jianwen Sun, Xian Peng and Tai Wang, “Sentiment recognition of online course reviews using multi-swarm optimization based selected features”, Neurocomputing, 2015. 12.036

[8] Andrea Esuli, and Fabrizio Sebastiani, "Determining the semantic orientation of terms through gloss classification." ,Proceedings of the 14th ACM international conference on Information and knowledge management. ACM, 2005.

[9] Q.V. Li, T. Mikolov, “Distributed representations of sentences and documents”, arXiv preprint arXiv: 1405.4053, 2014

[10] Zhijian Cui, Xiaodong Shi and Yidong Chen, “Sentiment Analysis via Integrating Distributed Representations of Variable-length Word Sequence”, Neurocomputing, 2015.07.129

[11] Chunxu Wu, Lingfeng Shen, and Xuan Wang, “A New Method of Using Contextual Information to Infer the Semantic Orientations of Context Dependent Opinions”, International Conference on Artificial Intelligence and Computational Intelligence, 2009.

[12] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques,” *Proc. EMNLP’02*, 2002, pp. 79–86.

[13] Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. 2007. Structured models for fine-to-coarse sentiment analysis. In Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL).

[14] Oscar Tackstrom, and Ryan McDonald, “Semi-supervised latent variable models for sentence-level sentiment analysis”, The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2, 2011, Pages 569-574

[15] Marjan Van de Kauter, Diane Breesch, and Véronique Hoste, “Fine-grained analysis of explicit and implicit sentiment in financial news Articles”, Expert Systems with Applications 42 (2015) 4999–5010

[16] Xiaolong Wang, et al. "Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach." Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011.

[17] Thien Hai Nguyen, Kiyoaki Shirai, and Julien Velcin, “Sentiment analysis on social media for stock movement prediction”, Expert Systems With Applications 42 (2015) 9603–9611

[18] Justin Martineau, and Tim Finin “Delta TFIDF: An Improved Feature Space for Sentiment Analysis”, Proceedings of the Third International ICWSM Conference (2009)

[19] Mohamed Abdel Fattah, “New term weighting schemes with combination of multiple classifiers for sentiment analysis”, Neurocomputing 167 (2015) 434–442

[20] Wanxiang Che, Yanyan Zhao, Honglei Guo, Zhong Su, and Ting Liu, “Sentence Compression for Aspect-Based Sentiment Analysis”, ieee/acm transactions on audio, speech, and language processing, vol. 23, no. 12, december 2015

[21] S. M. Mohammad, S. Kiritchenko, and X. Zhu, “NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets,” *CoRR*, vol. abs/1308.6242, 2013.

[22] Mohammad Sadegh Hajmohammadi, Roliana Ibrahim, Ali Selamat, and Hamido Fujita, “Combination of active learning and self-training for cross-lingual sentiment classification with density analysis of unlabelled samples”, Information Sciences 317 (2015) 67–77

[23] Parisa Lak, and Ozgur Turetken, “Star Ratings versus Sentiment Analysis - A Comparison of Explicit and Implicit Measures of Opinions”, 47th Hawaii International Conference on systescience, 2014

[24] Khalid Al-Rowaily, Muhammad Abulaish, Nur Al-Hasan Haldar, and Majed Al-Rubaian, “BiSAL e A bilingual sentiment analysis lexicon to analyze. Dark Web forums for cyber security”, Digital Investigation 14 (2015) 53e62

[25] E. Diener, “Subjective well-being”, Psychol. Bull. 95, 1984, pp. 542–575.

[26] E. Diener, “Subjective well-being: the science of happiness and a proposal for a national index”, Am. Psychol. 55 (1), 2000, p. 34.

[27] Jiayin Qi, Xianglin Fu, and Ge Zhu “Subjective well-being measurement based on Chinese grassroots blog text sentiment analysis”, Information & Management 52 (2015) 859–869