A Survey of Sentiment Analysis Process and Technologies

Priyanka Namdev [1], Prof. Lakhan Singh [2],
[1,2]*Department of Computer Science & Engineering,*
SAM College of Engineering, Bhopal, India

***Abstract*—Sentiment analysis of social networking is a newly rising research area of computer science which has recently attracted many researchers. Social networking like Twitters and Facebook present platforms for users where they can bring out, issue and publish their opinions and thoughts. In terms of thoughts and opinions expressed by the users, sentiment analysis undertake the problem by analyzing the text mining process. In this paper, an analytical survey is presented for sentiments analysis of social networking in context with methods and technologies. Finally, a concluding scope of sentiment analysis is presented for future research trends and its relative subject areas.**

***Keywords*—Sentiment Analysis, Social Networking, Data Mining.**

I. Introduction

Sentiment analysis also known as opinion minings generally use text analysis, natural language processing and computational linguistics to identify and study subjective information [1][2]. Sentiment analysis grasps the problem of analyzing the views posted on social networking in terms of the expressed sentiments. Twitter as on social networking service is a novel domain for sentiment analysis and very challenging [3][4]. Length limitation of sentences are one of the major challenges, according to which tweets can be up to 140 characters.

However, the small length messages and the normal type have caused the emergence of textual informalities that are widely encountered in social networking [5][6]. Thus, the techniques proposed for sentiment analysis must take into account these unique properties. Many sentiment analysis use a algorithm from the field of soft computing or machine learning, known as classifier [7]. Figure 1 represents the most

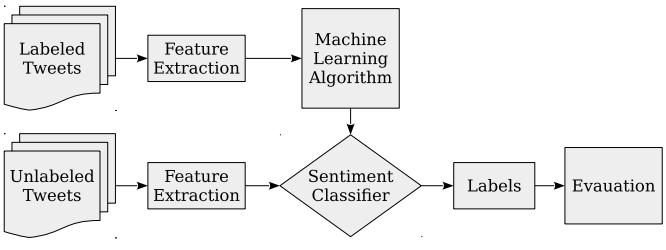

Fig. 1. Typical Process for Sentiment Classification

common sentiment analysis process. The initial step includes tweets collections from Twitter and labeling it by sentiment [8]. The labeled tweets represent the training data. Although API (Application Programming Interface) of social networking facilitates the process of collecting tweets, assigning labels is challenging and should be addressed carefully [9].

II. Survey of Sentiment Analysis Methods and Algorithms

In this section, an comprehensive survey is presented on various types of modern sentiment analysis algorithms and methods based on their classifications. In recent literature, so many works have been published on sentiment analysis of user-generated content [10], with reviews and social discussions [11], but analysis can be performed at different levels [12].

Asghar *et al.* [13] presented RIFT: a rule induction framework for Twitter sentiment analysis, this framework exposed and categorizes sentiments expressed by users in tweets relating to a product. This information is of tremendous value to assist business and government alike, to collect and analyse user feedback about products and services.

The proposed framework consisted of four modules:

- Noise reduction steps applied to the acquired text.
- Feature selection techniques, to select the most suitable text patterns collected from tweets.
- Rule induction framework, for the construction of the decision table and rule induction.
- Sentiment classification by proposing an enhanced version of Learning from Examples Module version:2 (LEM2), with Corpus-based Rule (CBR), (i.e. LEM2 + CBR).

The proposed technique greatly assists in classifying tweets by incorporating slang, emoticons and opinion words, and improves the performance of sentiment classification by focusing the rule induction framework on different data sets. The improved results about the reduced number of rules, improved accuracy, and maximum coverage show that the obtained classification results are much better than using baseline methods. Therefore, the framework can be used to classify the review process in any domain. A possible limitation of the approach, however, is the need to increase the number of conditional attributes in the decision table, which, if increased from eight could increase classification accuracy. The general-purpose nature of SentiWordNet may also result in inaccurate scoring (i.e. neutral scores) obtained of certain opinion words, such as “enjoy”, “like”, and “best”. To address this anomaly, domain-specific techniques should be further

investigated. A possible enhancement would be to exploit the use of blogs and review techniques for greater efficiency for sentiment classification. Other options for the RSES software, may be to shorten/eliminate descriptors from the antecedent part of the rules), generalisations, (i.e. eliminate descriptors from antecedent parts of rules) thereby making the rules more general and improving the filtering (i.e. removing rules having insufficient support on the training sample).

N. Öztürk *et al.* [14] studied a series of sentiment analyses using Twitter data performed in the subject of the Syrian civil war and following refugee crisis, which are currently among the most tragic and pressing issues in the world. We collected relevant tweets in two languages: Turkish and English. Upon a comprehensive twitter search, a total of 2,381,297 tweets were collected for analysis. Out of all, 1,353,367 of them were in English and 1,027,930 them were in Turkish. After removing duplicates, re-tweets, and the tweets with missing information as a part of data cleaning, 250,857 English tweets and 97,850 Turkish tweets, and the total of 348,707 tweets were used in the analyses.

Upon sentiment analysis of retrieved tweets for each language, we observed that Turkish tweets were carrying more positive sentiments about Syrians and refugees when compared to English tweets with the ratio of 35% of all tweets versus only 12%, respectively. While the sentiments of Turkish tweets were almost evenly distributed among the positive, neutral and negative categories, the English tweets were largely composed of neutral and negative sentiments. Furthermore, we found out that the details of the war happening near the borders of Turkey attracted more attention of Turkish speaking community, whereas English speaking community argued the legality of immigrants, policies and politics more frequently. Nonetheless, both communities talked about refugee children and shared their concerns about the children.

C. Diamantini *et al.* [15] worked in the development of an integrated system for information discovery from multiple social networks, which allows for the analysis of users' opinions and characteristics and is based on exploratory data analysis techniques. Furthermore, the paper introduces a novel set of features that are demonstrated to improve the classification accuracy of state-of-the-art noise detection algorithms for Twitter. In the system, the traditional lexicon-based sentiment

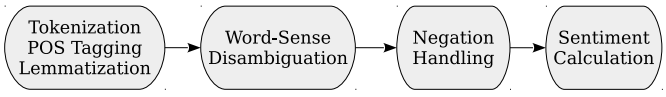

Fig. 2. The Sentiment Analysis Process

analysis is enhanced by two algorithms, which are for, respectively, the disambiguation of polysemous words and the correct handling of negated sentences as represented in Figure 2. The former algorithm detects the most suitable semantic variant of a polysemous word with respect of its context, by searching for the shortest path in a lexical resource from the polysemous word to its nearby words. The latter, on the other hand, detects the right scope of negation through the analysis of their parse trees. Experiments performed on four datasets show that coupling these algorithms results in a +6.7% improvement of classification accuracy in 3-class sentiment analysis.

Goldar *et al.* [16] presented a review of the real-time analytical and comprehensive analysis of big data parallelization techniques and methods. Basic tools to process and execute of big data are summarized and its unique characterizations are investigated based on the analytical methods and experiments. The advanced and future approaches of big data are reviewed as well.

S. Rani *et al.* [17] developed a sentiment analysis system for improvement of teaching and learning. During analysis phase, the sentiment analysis system analyzes the preprocessed data to identify instances of sentiment and emotion. It uses the Emotion Lexicon, also known as EmoLex, to associate words with positive or negative sentiment and the eight basic emotions. The lexicon supports 40 languages including several Indian ones like Hindi, Tamil, Gujarati, Marathi and Urdu. It includes annotations for 14,182 unigram words for English and 8,116 for Hindi.

Each word in the lexicon has an emotion vector $(\vec{E})$ containing a Boolean value (b) for each sentiment (s) and emotion (e):

$$\vec{E} = \vec{E}_e + \vec{E}_s, \tag{1}$$

where, $\vec{E}_e \in \{b_o, \ldots, b_7\}$,
and $\vec{E}_s \in \{b_8, b_9\}, \forall b_i \in \{0, 1\}$.
If a word in a student comment matches a word in the lexicon, the corresponding emotion vector is returned; if the word matches more than one word in the lexicon, the sum of the corresponding emotion vectors is returned. In this way, an emotion vector is created for each comment representing the different emotions and sentiments contained within.

To enable temporal analysis of sentiments and emotions, the system generates a mean emotion vector $(\vec{E}_j)$ for each month and year:

$$\vec{E}_j = \frac{1}{n} \sum_{i=0}^{n-1} \sum_{j=0}^{p-1} \vec{E}_{ji}, \ \forall \vec{E}_{ji} \in N \text{where } N \geq 0. \tag{2}$$

Here, n represents the number of comments in each month and year and p represents the emotion and sentiment parameters.

Hogenboom *et al.* [18] demonstrated that sentiment analysis can benefit from deep analysis of a text's rhetorical structure, enabling the distinction be made between important text segments and less-important ones in terms of their contribution to a text's overall sentiment. This is a significant step forward with respect to existing work, which is limited to guiding sentiment analysis by shallow analyses of rhetorical relations in (mostly sentence-level) rhetorical structure trees.

III. Problem Formulation

There is basically no problems in current sentiment analysis methods and algorithms however some limitations in sentiment analysis can be point out as

- The problem in sentiment analysis is classifying the polarity of a given text at the document, sentence or feature/aspects level.
- Whether the expressed opinion in a document, a sentence or an entire feature/aspect is positive, negative or neutral.
- Given a set of tweets containing multiple features and varied opinions, the objective is to extract expressions of opinion describing a target feature and classify it as positive or negative.
- There is currently no automated domain–independent sentiment classification tool with high accuracy that does not need a manually-annotated corpus.
- Such a tool is needed for opinion search, recommendation, summarization and mining of the increasingly web opinionated content.

IV. Proposed Approach

Sentiment analysis can be considered as a classification problem, since the goal in the typical scenario is to classify the opinion expressed in a tweet as positive or negative. The most frequently used evaluation metrics are accuracy, precision, recall, and $\mathcal{F}-$Score, adopted from traditional classification problems. Table I describes the performance of this method

TABLE I
Confusion Matrix for Performance of Sentiment Analysis

	Predicted as *Positive*	Predicted as *Negative*
Are *Positive*	TP	FN
Are *Negative*	FP	TN

on a set of test data for which the sentiment is known. This table, also called a confusion matrix, shows the number of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) instances that are used to compare the predictions of the method with the ground truth. TP represents the number of instances that were predicted as positive and were indeed positive, whereas FP is the number of instances incorrectly predicted as positive. TN and FN have a corresponding meaning for the negative class.

Precision: Precision represents the exactness of the method and is calculated as the ratio of instances that were predicted as positive and were indeed positive divided by the total number of instances that were predicted as positive. That is:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

Recall: Recall, which is also known as sensitivity, denotes the fraction of positive instances that were predicted to be positive and is calculated as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

Accuracy: Accuracy is the most frequently used evaluation metric and measures how often the method being evaluated made the correct prediction. It is calculated as the sum of the true predictions divided by the total number of predictions. That is:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{5}$$

$\mathcal{F}-$**Score:** Usually, calculating recall and precision is not enough. A combination of the two is more appropriate to evaluate the performance of the methods. The $\mathcal{F}-$score is the metric that combines recall and precision. This metric is also known as *harmonic* $\mathcal{F}-$score, $\mathcal{F}_1-$Score, or $\mathcal{F}-$Measure accuracy and is calculated as:

$$\mathcal{F} - \text{Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

Finally, when the sentiment classification is formulated as a multi-class problem (i.e., it aims to classify a tweet as positive, negative, or neutral), it is common practice to calculate the positive, negative, and neutral $\mathcal{F}-$Score. However, there are approaches that do not predict the neutral class. This does not mean that the task is reduced to predicting only positive and negative tweets. These approaches should still be evaluated on the whole ground truth that includes neutral tweets.

V. Conclusion

In this survey, we presented the comprehensive survey and theoretical study of different strategies by which sentiment analysis can be estimated of social networking. There are still major challenging tasks that are required to be improved in the sentiment analysis. Those challenging tasks can be highlighted in future research directions in this field.

References

[1] B. Liu, *Sentiment Analysis and Opinion Mining*. Morgan & Claypool, 2012.

[2] R. Biagioni, *Sentiment Analysis*. Cham: Springer International Publishing, 2016, pp. 7–16.

[3] M. Bouazizi and T. Ohtsuki, "A pattern-based approach for multi-class sentiment analysis in twitter," *IEEE Access*, vol. 5, pp. 20 617–20 639, 2017.

[4] D. Gonzalez-Marron, D. Mejia-Guzman, and A. Enciso-Gonzalez, "Exploiting data of the twitter social network using sentiment analysis," in *Applications for Future Internet*, E. Sucar, O. Mayora, and E. Munoz de Cote, Eds. Cham: Springer International Publishing, 2017, pp. 35–38.

[5] C. Wang, Z. Xiao, Y. Liu, Y. Xu, A. Zhou, and K. Zhang, "Sentiview: Sentiment analysis and visualization for internet popular topics," *IEEE Transactions on Human-Machine Systems*, vol. 43, no. 6, pp. 620–630, Nov 2013.

[6] A. Trilla and F. Alias, "Sentence-based sentiment analysis for expressive text-to-speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 223–233, Feb 2013.

[7] M. Brooks, J. J. Robinson, M. K. Torkildson, S. R. Hong, and C. R. Aragon, "Collaborative visual analysis of sentiment in twitter events," in *Cooperative Design, Visualization, and Engineering*, Y. Luo, Ed. Cham: Springer International Publishing, 2014, pp. 1–8.

[8] L. C. Yu, J. Wang, K. R. Lai, and X. Zhang, "Refining word embeddings using intensity scores for sentiment analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 671–681, March 2018.

[9] M. Trupthi, S. Pabboju, and G. Narasimha, "Sentiment analysis on twitter using streaming api," in *2017 IEEE* 7^{th} *International*

Advance Computing Conference (IACC), Jan 2017, pp. 915–919.

[10] F. Colace, L. Casaburi, M. D. Santo, and L. Greco, "Sentiment detection in social networks and in collaborative learning environments," *Computers in Human Behavior*, vol. 51, pp. 1061 – 1067, 2015, computing for Human Learning, Behaviour and Collaboration in the Social and Mobile Networks Era.

[11] Z. Jianqiang and G. Xiaolin, "Comparison research on text preprocessing methods on twitter sentiment analysis," *IEEE Access*, vol. 5, pp. 2870–2879, 2017.

[12] K. Schouten and F. Frasincar, "Survey on aspect-level sentiment analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 3, pp. 813–830, March 2016.

[13] M. Z. Asghar, A. Khan, F. Khan, and F. M. Kundi, "Rift: A rule induction framework for twitter sentiment analysis," *Arabian Journal for Science and Engineering*, vol. 43, no. 2, pp. 857–877, Feb 2018.

[14] N. Öztürk and S. Ayvaz, "Sentiment analysis on twitter: A text mining approach to the syrian refugee crisis," *Telematics and Informatics*, vol. 35, no. 1, pp. 136 – 147, 2018.

[15] C. Diamantini, A. Mircoli, D. Potena, and E. Storti, "Social information discovery enhanced by sentiment analysis techniques," *Future Generation Computer Systems*, 2018.

[16] P. Goldar, Y. Rai, and S. Kushwaha, "A review on parallelization of big data analysis and processing," *International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)*, vol. 23, no. 4, pp. 60–65, August 2016. [Online]. Available: http://www.ijetcse.com/wp-content/plugins/ijetcse/file/upload/docx/151A-Review-on-Parallelization-of-Big-Data-Analysis-and-Processing-pdf.pdf

[17] S. Rani and P. Kumar, "A sentiment analysis system to improve teaching and learning," *Computer*, vol. 50, no. 5, pp. 36–43, May 2017.

[18] A. Hogenboom, F. Frasincar, F. de Jong, and U. Kaymak, "Using rhetorical structure in sentiment analysis," *Commun. ACM*, vol. 58, no. 7, pp. 69–77, Jun. 2015.