A Review of Predictive Sentiment Analysis of Social Media Data

Rashmi Nande [1], Paritosh Goldar [2], Manish Kumar Sahu [3],
[1,2,3]Department of Computer Science & Engineering,
[1]RKDF University, Bhopal, India

Abstract: With recent expeditious development in Social Media users, the researcher have attracted towards the sentiments analysis of social media users data for particular event, political issues, product or movie reviews. Facebook, Twitter, Instagram, Youtube etc. are is one of the popular and most widely used social networking platform to reveal express users thoughts. With the help of sentiment analysis of social data, the valuable thoughts of society can be predicted, and analytical models will be beneficial for many predictive analysis. In this paper, an exploratory review and study toward sentiment analysis of social media data is presented with their methods and approaches. This review provides an overview of several available methods used in mathematical, statistical as well as sentiment analysis.

Keywords: Sentiment Analysis, Predictive Analysis, Social Networking, Media, Emotions.

I Introduction

The exponential growth in the use of digital devices, together with ubiquitous online access, provides unprecedented ground for the constant connectivity of people and offers tremendous capabilities for publicly expressing opinions, attitudes, or reactions regarding many aspects of everyday human activities [1]. Social media, such as blogs, forums, and social network platforms (eg, Facebook, Twitter, LinkedIn, Youtube, Instagram) are quickly becoming an integral part of people's lives, the virtual spaces where daily individuals share opinions and information and maintain and/or expand their relational network [2]. The massive use of online social networks and the abundance of data collected through them has raised exponentially the attention of the scientific and business community toward them [3–5].

Nowadays, the constant refinement of analytical tools is offering a richer array of opportunities to analyze these data for many different purposes [6]. Differences in features and characteristics of online social networks are reflected in the huge amount of different statistics and metrics that it is possible to track and analyze. The most adopted metrics are numeric, relatively easy to obtain, and freely available, such as engagement and influence metrics [7]. However, metrics of this types are often defined as "vanity metrics", since they do not interpret or contextualize the data collected. For this reason, other types of methods of analysis has been introduced. Among them, one of the most used is sentiment analysis [8], which is the analysis of the feelings (ie, opinions, emotions and attitudes) behind the words using natural language processing tools. SA is considered a quality metric, which looks behind numbers to understand how information about emotion and attitudes is conveyed in language [9, 10].

II Predictive Analysis

Predictive analytics is the use of data, mathematical algorithms and machine learning to identify the likelihood of future events based on historical data [11]. The main goal of predictive analytics is to use the knowledge of what has happened to provide the best valuation of what will happen. In other words, predictive analytics can offer a complete view of what is going on and the information we need to succeed [12].

2.1 Predictive Analytics Process

Figure 1 represents the press involved in predictive analytics.

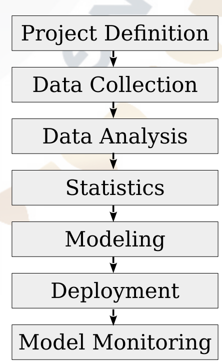

Figure 1: Predictive Analytics Process

1. **Define Project:** Define the project outcomes, deliverable, scope of the effort, business objectives, identify the data sets that are going to be used.

2. **Data Collection:** Data mining for predictive analytics prepares data from multiple sources for analysis. This provides a complete view of customer interactions.

3. **Data Analysis:** Data Analysis is the process of inspecting, cleaning and modelling data with the objective of discovering useful information, arriving at conclusion

4. **Statistics:** Statistical Analysis enables to validate the assumptions, hypothesis and test them using standard statistical models.

5. **Modeling:** Predictive modelling provides the ability to automatically create accurate predictive models about future. There are also options to choose the best solution with multi-modal evaluation.

6. **Deployment:** Predictive model deployment provides the option to deploy the analytical results into everyday decision making process to get results, reports and output by automating the decisions based on the modelling.

7. **Model Monitoring:** Models are managed and monitored to review the model performance to ensure that it is providing the results expected.

III Predictive Sentiment Analysis

Sentiment Analysis refers to the use of text analysis and natural language processing to identify and extract subjective information in textual contents [13]. There are two type of user-generated content available on the web - facts and opinions. Facts are statements about topics and in the current scenario, easily collectible from the Internet using search engines that index documents based on topic keywords. Opinions are user specific statement exhibiting positive or negative sentiments about a certain topic. Generally opinions are hard to categorize using keywords. Various text analysis and machine learning techniques are used to mine opinions from a document [14][15]. Sentiment Analysis finds its application in a variety of domains.

- **Business:** Businesses may use sentiment analysis on blogs, review websites etc. to judge the market response of a product. This information may also be used for intelligent placement of advertisements. For example, if product “A” and “B” are competitors and an online merchant business “M” sells both, then “M” may advertise for “A” if the user displays positive sentiments towards “A”, its brand or related products, or “B” if they display negative sentiments towards “A”.

- **Government:** Governments and politicians can actively monitor public sentiments as a response to their current policies, speeches made during campaigns etc. This will help them make create better public awareness regarding policies and even drive campaigns intelligently.

- **Financial Markets:** Public opinion regarding companies can be used to predict performance of their stocks in the financial markets. If people have a positive opinion about a product that a company A has launched, then the share prices of A are likely to go higher and vice versa. Public opinion can be used as an additional feature in existing models that try to predict market performances based on historical data.

IV Related Work

Many researcher carried out their research work in sentiments analysis using social media. Several reseacher have emphsize their attention on stastical results from social media using various sentiments analysis methods. Malhar Anjaria *et al.* [16] introduce the novel approach of exploiting the user influence factor in order to predict the outcome of an election result. Athours also propose a hybrid approach of extracting opinion using direct and indirect features of Twitter data based on Support Vector Machines (SVM), Naive Bayes, Maximum Entropy and Artificial Neural Networks based supervised classifiers.

Min Song *et al.* [17] employ temporal Latent Dirichlet Allocation (LDA) to analyze and validate the relationship between topics extracted from tweets and related events. They developed the term cooccurrence retrieval technique to trace chronologically cooccurring terms and thereby compensate for LDAs limitations. Finally, authors identify thematic coherence among users identified in sending receiving mentions.

Li Bing *et al.* [18] proposed a method to mine Twitter data for prediction of the movements of the stock price of a particular company through public sentiments. Authors also explain how stock price of one company to be more predictable than that of another company and they proposed to used a data mining algorithm to determine the stock price movements of 30 companies listed in NASDAQ and the New York Stock Exchange can actually be predicted by the given 15 million records of tweets (i.e., Twitter messages). They did so by extracting ambiguous textual tweet data through NLP techniques to define public sentiment, then make use of a data mining technique to discover patterns between public sentiment and real stock price movements.

Go, Bhayani and Huang (2009) [19] were among the first to explore sentiment analysis on Twitter. They classify Tweets for a query term into negative or positive sentiment. They collect training dataset automatically from Twitter. To collect positive and negative tweets, they query twitter for happy and sad emoticons. Happy emoticons are different versions of smiling face, like “`:)`”, “`:-)`”, “`: )`”, “`:D`”, “`=)`” etc. Sad emoticons include frowns, like “`:(`”, “`:-(`”, “`:(`” etc. They try various features – unigrams, bigrams and Part-of-Speech and train their classifier on various machine learning algorithms – Naive Bayes, Maximum Entropy and Scalable Vector Machines and compare it against a baseline classifier by counting the number of positive and negative words from a publicly available corpus. They report that Bigrams alone and Part-of-Speech Tagging are not helpful and that Naive Bayes Classifier gives the best results.

V Proposed Approach

Several feature sets and machine learning classifiers can be used to determine the best combination for sentiment analysis of social media data. Many pre-processing steps like – emoticons, punctuations, specific tweets, terms, slogans and stemming can be analyzed also. Some data analytical features such as – unigrams, bigrams, trigrams, positive and negation detection, joy and sadness expressions can be investigated. The identical train our classifier using various machine-learning algorithms such as – Naive Bayes, decision trees and maximum entropy approaches can be implemented for improved predictive sentiment analysis.

VI Conclusion

In this paper, a sentiment classifier for social networking data using various mathematical and statistical methods with labeled data sets are reviewed. We also investigate the relevance of using a double step classifier and negation detection for the purpose of sentiment analysis. The presented exploratory review and study toward sentiment analysis of social media data represented with their proposed methods and approaches. This review provides an basic overview of popular methods used in mathematical, statistical as well as predictive sentiment analysis of social media data.

REFERENCES

[1] D. Meshi, D. I. Tamir, and H. R. Heekeren, "The emerging neuroscience of social media," *Trends in Cognitive Sciences*, vol. 19, no. 12, pp. 771–782, 2015. [Online]. Available: https://doi.org/10.1016/j.tics.2015.09.004

[2] S. Ghani, B. C. Kwon, S. Lee, J. S. Yi, and N. Elmqvist, "Visual analytics for multimodal social network analysis: A design study with social scientists," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2032–2041, Dec 2013. [Online]. Available: https://doi.org/10.1109/TVCG.2013.223

[3] A. M. Kaplan and M. Haenlein, "Users of the world, unite! the challenges and opportunities of social media," *Business Horizons*, vol. 53, no. 1, pp. 59–68, 2010. [Online]. Available: https://doi.org/10.1016/j.bushor.2009.09.003

[4] G. Merchant, "Unravelling the social network: theory and research," *Learning, Media and Technology*, vol. 37, no. 1, pp. 4–19, 2012. [Online]. Available: https://doi.org/10.1080/17439884.2011.567992

[5] M. Salampasis, G. Paltoglou, and A. Giachanou, "Using social media for continuous monitoring and mining of consumer behaviour," *International Journal of Electronic Business*, vol. 11, no. 1, pp. 85–96, 2013.

[6] M. L. P. Felciah and R. Anbuselvi, "A study on sentiment analysis of social media reviews," in *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, March 2015, pp. 1–3. [Online]. Available: https://doi.org/10.1109/ICIIECS.2015.7192949

[7] S. Ye and S. F. Wu, "Measuring message propagation and social influence on twitter.com," in *Social Informatics*, L. Bolc, M. Makowski, and A. Wierzbicki, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 216–231.

[8] F. Pozzi, E. Fersini, E. Messina, and B. Liu, "Challenges of sentiment analysis in social networks: An overview," in *Sentiment Analysis in Social Networks*, F. A. Pozzi, E. Fersini, E. Messina, and B. Liu, Eds. Boston: Morgan Kaufmann, 2017, pp. 1–11. [Online]. Available: https://doi.org/10.1016/B978-0-12-804412-4.00001-2

[9] A. Abbasi, D. Adjeroh, M. Dredze, M. J. Paul, F. M. Zahedi, H. Zhao, N. Walia, H. Jain, P. Sanvanson, R. Shaker, M. D. Huesch, R. Beal, W. Zheng, M. Abate, and A. Ross, "Social media analytics for smart health," *IEEE Intelligent Systems*, vol. 29, no. 2, pp. 60–80, Mar 2014. [Online]. Available: https://doi.org/10.1109/MIS.2014.29

[10] Y. E. Sagduyu, A. Grushin, and Y. Shi, "Synthetic social media data generation," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 3, pp. 605–620, Sept 2018. [Online]. Available: https://doi.org/10.1109/TCSS.2018.2854668

[11] K.-R. Koch, *Predictive analysis.* Berlin, Heidelberg: Springer Berlin Heidelberg, 1990, pp. 49–51. [Online]. Available: https://doi.org/10.1007/BFb0048707

[12] M. Muchová, J. Paralič, and M. Nemčík, *Using Predictive Data Mining Models for Data Analysis in a Logistics Company.* Cham: Springer International Publishing, 2018, pp. 161–170. [Online]. Available: https://doi.org/10.1007/978-3-319-67220-5_15

[13] A. R. Samanpour, A. Ruegenberg, and R. Ahlers, *The Future of Machine Learning and Predictive Analytics.* Berlin, Heidelberg: Springer Berlin Heidelberg, 2018, pp. 297–309. [Online]. Available: https://doi.org/10.1007/978-3-662-49275-8_30

[14] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. pages 1–135, 2008.

[15] H. Dominik, *Sentiment Analysis.* Wiesbaden: Springer Fachmedien Wiesbaden, 2017, pp. 111–112. [Online]. Available: https://doi.org/10.1007/978-3-658-19287-7_17

[16] M. Anjaria and R. M. R. Guddeti, "Influence factor based opinion mining of twitter data using supervised learning," in *2014 Sixth International Conference on Communication Systems and Networks (COMSNETS)*, Jan 2014, pp. 1–8.

[17] M. Song, M. C. Kim, and Y. K. Jeong, "Analyzing the political landscape of 2012 korean presidential election in twitter," *IEEE Intelligent Systems*, vol. 29, no. 2, pp. 18–26, Mar 2014.

[18] L. Bing, K. C. C. Chan, and C. Ou, "Public sentiment analysis in twitter data for prediction of a company's stock price movements," in *2014 IEEE 11th International Conference on e-Business Engineering*, Nov 2014, pp. 232–239.

[19] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *Processing*, pp. 1–6, 2009. [Online]. Available: http://www.stanford.edu/~alecmgo/papers/TwitterDistantSupervision09.pdf