A Novel Proof of Concept for Twitter Analytics Using Popular Hashtags: Experimentation and Evaluation

Kiran Ahire, Manali Bagul, Swapnil Dhanawate, and Suja Sreejith Panicker

Abstract Twitter analytics is a classic research area especially with the widespread presence of Big Data in various online media such as—social network sites, online portals for shopping, e-commerce, forums, chats, recommendation systems, and online services. Ascertaining the sentiment behind, the various types of tweets by different persons can provide great insights on various aspects including behavioral patterns. Besides highlighting the newest trends in the field, we retrieved real-time twitter data pertaining to three currently popular hashtags in the Indian context and carried out extensive experimentation analysis about the prevailing sentiment of a strata of population. Inclusion of current challenges, future trends and applications of sentiment analysis from Twitter data makes this novel work very useful for fellow researchers.

Keywords Machine learning · Natural language processing · Sentiment analysis · Twitter data analytics · Opinion mining · Data visualization · Recommendation systems · Attitude analysis · Polarity determination · Sentiment classification

K. Ahire · S. Dhanawate
Department of Computer Engineering, Maharashtra Institute of Technology, Pune, Maharashtra, India
e-mail: Ahirekiran930@gmail.com

S. Dhanawate
e-mail: swapnildhanawate@gmail.com

M. Bagul · S. S. Panicker (✉)
School of Computer Engineering and Technology, MIT World Peace University, Pune, Maharashtra, India
e-mail: suja.panickar@mitwpu.edu.in

M. Bagul
e-mail: manalibagul123@gmail.com

V. Goyal et al. (eds.), *Proceedings of International Conference on Communication and Artificial Intelligence*, Lecture Notes in Networks and Systems 192,
https://doi.org/10.1007/978-981-33-6546-9_31

1 Introduction

Analytics of twitter data is predominantly conducted to ascertain the underlying sentiment behind the tweets/images. Hence, this effectively narrows to sentiment analysis. Sentiment analysis is computational study of the various opinions, emotions, sentiments, and attitude which is expressed by different users in the form of texts pertaining to an entity of interest. Sentiment analysis is also called as review mining, opinion mining, or attitude analysis [1–5].

Motivation for the surge in voluminous user content globally is attributed to technological advancements as also increased Internet activities like—discussion forums, conferencing, online transactions, e-commerce, chatting, surveillances, ticket booking, websites of merchants, widespread and continual communications on various social media, and the variety of other online activities [1, 3, 6, 7].

Current work is organized as follows: Sect. 2 covers motivation, Sect. 3 covers literature survey, Sect. 4 covers experimentation and results, Sect. 5 presents observations, Sect. 6 highlights the novelty, Sect. 7 presents the various applications, Sect. 8 presents challenges, Sect. 9 presents research contribution, and Sect. 10 covers conclusion.

2 Motivation

This novel technique will help people to analyze various data from Twitter and help understand the public opinion or sentiment of people behind the specific keywords, and this will be useful in various sectors like business, marketing, forecasting, politics, and tourism.

3 Literature Survey

There are mainly two approaches found in existing literature [1, 7–18] for performing sentiment analysis–lexicon based and machine learning based. Concept of polarity is used in the former while suitable classification models are developed in the latter.

Detailed survey of recent work is presented in Table 1, and research gaps are highlighted.

4 Experimentation and Results

We used Tweepy to fetch the tweets in real time for three currently popular hashtags in India: #MakeInIndia, #AtmNirbharIndia, #VocalforLocal. The tweepy.Cursor()

Table 1 Survey of sentiment analysis in recent works

Ref. No.	Year	Machine learning algorithm	Datasets used	Result	Research gap
[18]	2020	NLP, Decision Tree, SVM, Random Forest	Collecting datasets (tweets—40,000)	Highest accuracy 99.4 obtained with Random Forest	SVM technique having less accuracy for analysis
[19]	2020	Deep recurrent neural network classifier	Real data	Highest accuracy of 93% obtained with Hadoop-based deep RNN	Adding more features in the feature extraction process
[20]	2020	Valence Aware Dictionary and Sentiment Reasoner	Geo tweet Harvard Center for Geographic Analysis	Overall posted positive tweets are increasing from Monday to Sunday and peaks on Sunday	Data noise, Advanced bots detection methods should be employed to refine dataset, Negative sentiment analysis
[21]	2020	NLP, Map visualization	Real data Kumamoto Earthquake	Map visualization with statistical relevancy index	To gather more exhaustive requirements for each area and the information become more reliable
[22]	2020	SVM	Real data	Negative and positive comments	Improve the precision in feeling mining in python structure
[23]	2020	Feature ensemble model, GloVe Model	Real data	Proposed method had F1 score of 0.81	Apply Parsimonious Extreme Machine Learning
[24]	2020	Logistic Regression Classifier algorithm, SVM	Stanford dataset, Sanders Twitter sentiment corpus dataset	Logistic Regression with highest accuracy of 91%	Combine emergency detection and polar sentiment analysis. Detect public health emergencies such as 2019 nCov
[25]	2020	DNN, Naïve Bayes	Kaggle dataset, Twitter streaming API	11.2% positive for females, 12.65% positive for males	Producing new database that will contribute to further studies in this subject

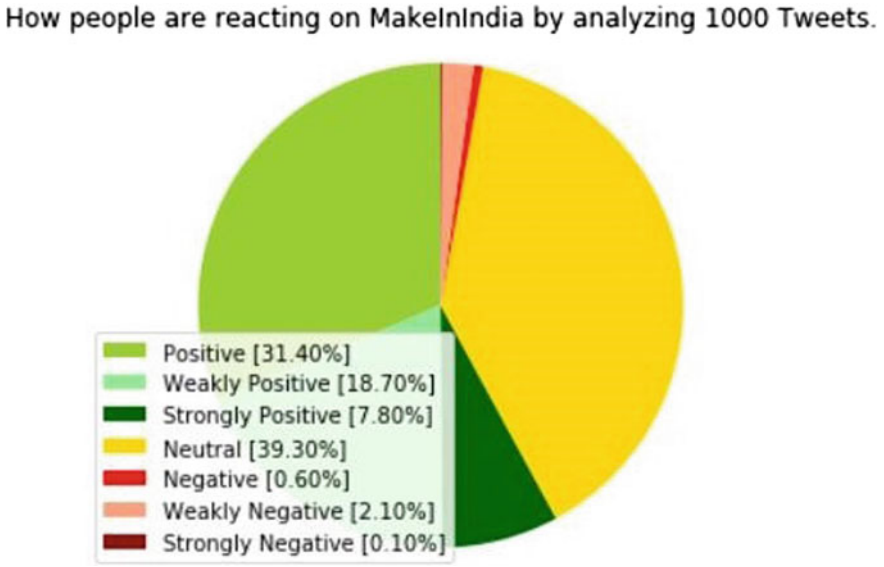

Fig. 1 Analysis of 1000 tweets for #MakeInIndia

function was used to fetch all latest tweets. Preprocessing was performed using the 're' library of python. TextBlob was used for polarity determination. We wrote a python program to encode the seven class labels as follows: -1 negative, -0.6 to -1 strongly negative, 0 to -0.3 weakly negative, 0 neutral, 0 to 0.3 weakly positive, 0.6 to 1 strongly positive, and 1 for positive and performed three experiments as under.

4.1 Experiment 1: #MakeInIndia

We fetched 1000 tweets in real time and have analyzed the same for ascertaining the sentiment. Visualization results for seven sentiment classes are as illustrated in Fig. 1

4.2 Experiment 2: #AtmNirbhar

We fetched 1000 tweets in real time and have analyzed the same for ascertaining the sentiment. Visualization results are as illustrated in Fig. 2

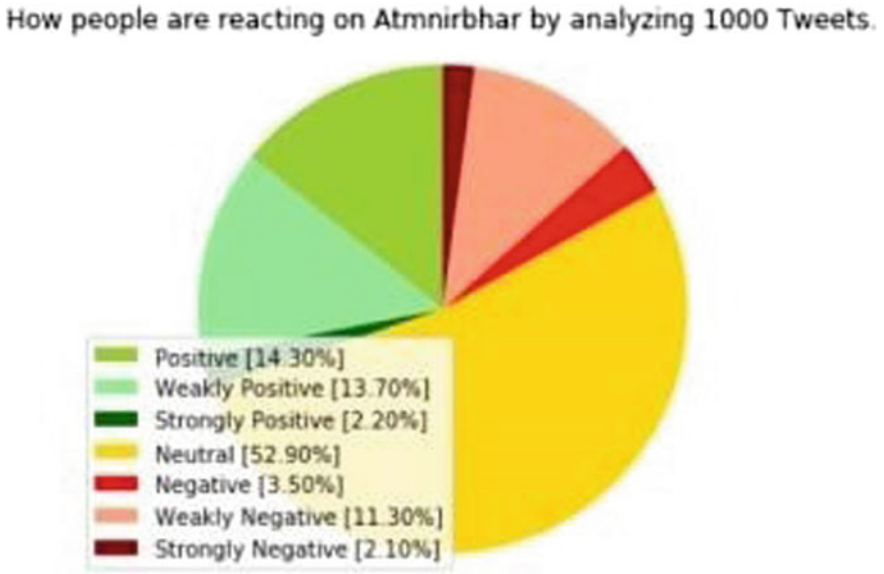

Fig. 2 Analysis of 1000 tweets for #AtmNirbhar

4.3 Experiment 3: #VocalforLocal

Figure 3 illustrates the outcome of analyzing 200 tweets.

We performed comparative analysis of the two hashtags with respect to seven sentiment classes as illustrated in the stacked bar chart in Fig. 4

To validate the obtained results, we assigned the task of annotation to two human experts and noted the findings. Figures 5 and 6 illustrate the differences in annotation between the two experts using RMSE and standard deviation, respectively.

5 Observations

- From Figs. 1, 2, 3 and 4, we infer that the highest positive percentage of tweets was for #MakeInIndia while the highest negative tweets were for #Atmnirbhar
- From Table 1, it is observed that although some standard datasets do exist, most researchers prefer to gather tweets in real time. Tweepy was observed to be the predominant choice. Also, SVM and Random Forests have frequently yielded high accuracy of over 95%

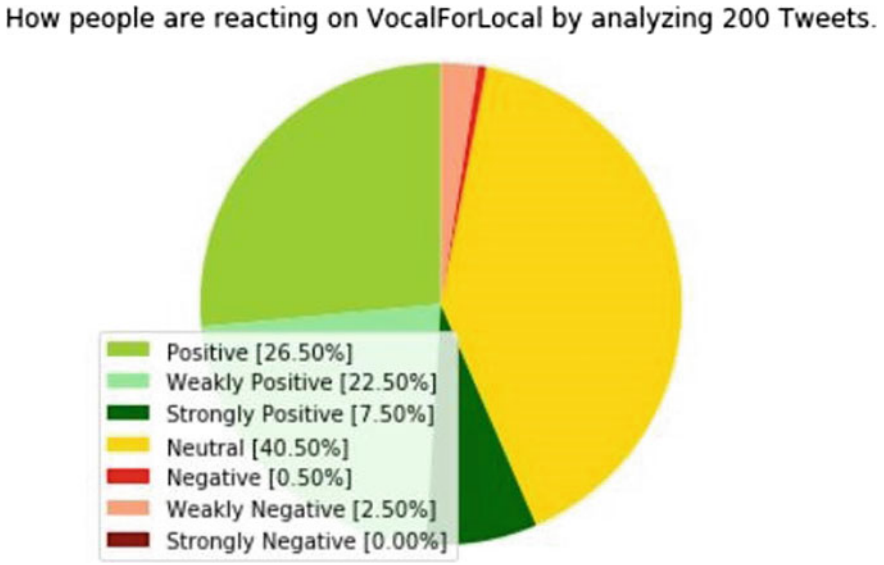

Fig. 3 Analysis of 200 tweets for #MakeInIndia

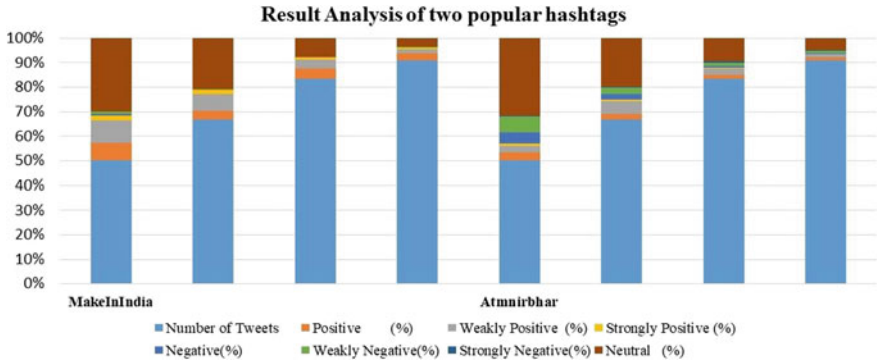

Fig. 4 Comparative analysis of two hashtags with respect to seven sentiment classes

6 Novelty

This technique gives the result visualization in the form of pie-chart along with seven classes which gives the clearer idea about the sentiment behind keyword, and this novel approach of result visualization helps people to understand result in detail.

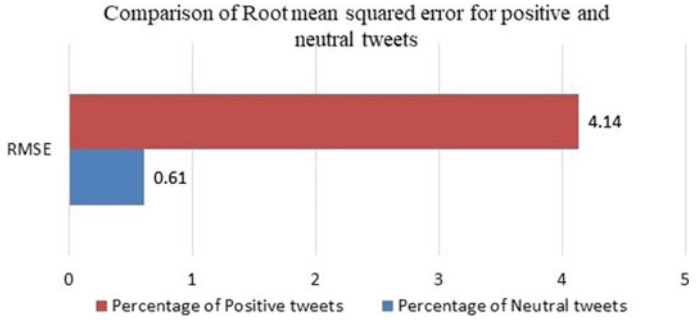

Fig. 5 Differences in RMSE values for the two human experts

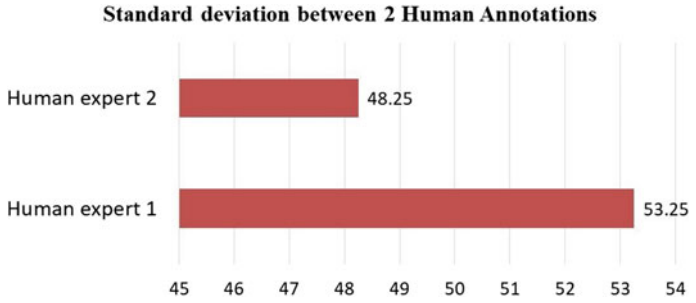

Fig. 6 Inter-individual differences in annotation by two human experts

7 Applications

Twitter data analytics has variety of applications such as

- For generating reputation for brands or products [26–28],
- For increasing the customer engagement, having better informed decisions toward risk analysis, efficient credit ratings for various customers, and performing competitive analysis [29],
- Increasing productivity and efficiency of restaurants [30],
- For better market intelligence and improve customer satisfaction [3, 36],
- Increased tourism [37],
- Monitoring and analyzing public opinions concerning political issues [3],
- To forecast the price changes as per news sentiments [1],
- To develop new products, services and promote products as per the customers reviews [1] and social advertising [38, 39].

8 Challenges in Twitter Data Analytics

i. Determining the contextual information for sentiments and forming a generalized foundation globally is difficult [30].
ii. There is increased difficulty due to the widespread use of onomatopoeias, idioms, homophones, alliterations, and acronyms [30]. Hence, complex NLP techniques are required to decipher the correct context and meaning of various words.
iii. Aspect-based sentiment analysis is an important challenge [36].
iv. Opinion summarization, subjectivity classification, and opinion retrieval [36]
v. Lack of large annotated data to train models across various domains [40]

9 Research Contribution

- Current work is a novel approach of visualizing and analyzing the three currently popular hashtags in India. Our extensive experimentation and analysis about the prevailing sentiment shall be greatly beneficial for fellow researchers.
- We have also covered important aspects such as—current challenges, future trends, and applications of sentiment analysis.

10 Conclusion

We have successfully implemented the proof of concept toward gathering tweets in real time and attempting to analyze the sentiment of a part of population using the lexicon-based technique. We have performed extensive experimentation and analyzed the sentiments for 100, 200, 500, and 1000 different sets of tweets for three most currently most popular hashtags. Ample data visualization performed in this work would be great asset to fellow researchers thereby carving the path for future research.

References

1. Kumar R, Vadlamani R (2015) A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowledge Based Systems
2. Liu B (2010) Sentiment analysis: a multi-faceted problem. IEEE Intell Syst 25(3):76–80
3. Tubishat M, Idris N, Abushariah MAM (2018) Implicit aspect extraction in sentiment analysis: review, taxonomy, opportunities, and open challenges. Inf Process Manage 54:545–563
4. Montoyo A, Martínez-Barco P, Balahur A (2012) Subjectivity and sentiment analysis: an overview of the current state of the area and envisaged developments. Decis Support Syst 53:675–679
5. Kumar S, Morstatter F, Liu H (2013) Twitter data analytics. Springer

6. Liu J (2008) Opinion spam and analysis. In: Proceedings of the international conference on web search and web data mining, ACM
7. Tan LKW, Na JC, Theng YL et al (2012) Phrase-level sentiment polarity classification using rule-based typed dependencies and additional complex phrases consideration. J Comput Sci Technol 27(3):650–666
8. Wang T et al (2014) Product aspect extraction supervised with online domain knowledge. Knowledge-Based Syst 71:86–100
9. Kunte AV, Panicker S (2020) Analysis of machine learning algorithms for predicting personality: brief survey and experimentation. In: 2019 global conference for advancement in technology (GCAT)
10. Kunte A, Panicker S (2020) Personality prediction of social network users using ensemble and XGBoost. In: Das H, Pattnaik P, Rautaray S, Li KC (eds) Progress in computing, analytics and networking. Advances in intelligent systems and computing, vol 1119. Springer, Singapore
11. Kunte AV, Panicker SS (2019) Using textual data for personality prediction: a machine learning approach. In: 2019 4th international conference on information systems and computer networks (ISCON)
12. Panicker S, Kunte A (2019) Personality prediction using social media. In: 2019 5th international conference for convergence in technology (I2CT), Pune (in Press)
13. Mane VL, Panicker SS (2015) Knowledge discovery from user health posts. In: IEEE 9th international conference on intelligent systems and control (ISCO)
14. Mane VL, Panicker SS (2015) Summarization and sentiment analysis from user health posts. In: 2015 international conference on pervasive computing (ICPC). IEEE
15. Salunke V, Panicker SS (2021) Image sentiment analysis using deep learning. In: Ranganathan G, Chen J, Rocha A (eds) Inventive communication and computational technologies. Lecture notes in networks and systems, vol 145. Springer, Singapore
16. Dangra BS, Rajput D, Bedekar MV, Panicker SS (2015) Profiling of automobile drivers using car games. In: International conference on pervasive computing (ICPC). IEEE
17. Bedekar MV, Atote B, Zahoor S, Panicker S (2016) Proposed used of information DisPersal Algorithm in user profiling. In: International conference on ICT for sustainable development, Goa, India
18. Khan M, Malviya A (2020) Big data approach for sentiment analysis of twitter data using Hadoop framework and deep learning. In: 2020 international conference on emerging trends in information technology and engineering (ic-ETITE), Vellore, India
19. Hu T, She B, Duan L, Yue H, Clunis J (2020) A systematic spatial and temporal sentiment analysis on geo-tweets. IEEE Access 8, 8658–8667
20. Murakami A, Nasukawa T, Watanabe K, Hatayama M (2020) Understanding requirements and issues in disaster area using geotemporal visualization of Twitter analysis. IBM J Res Develop
21. Kumar TS, Nabeem PM, Manoj CK, Jeyachandran K (2020) Sentimental analysis (opinion mining) in social network by using SVM algorithm. In: 2020 fourth international conference on computing methodologies and communication (ICCMC), Erode, India
22. Phan HT, Tran VC, Nguyen NT, Hwang D (2020) Improving the performance of sentiment analysis of Tweets containing fuzzy sentiment using the feature ensemble model. In: IEEE Access, vol 8, pp 14630–14641
23. Bhatnagar D, SubaLakshmi RJ, Vanmathi C (2020) Twitter Sentiment Analysis Using Elasticsearch, LOGSTASH And KIBANA. In: 2020 international conference on emerging trends in information technology and engineering (ic-ETITE)
24. Oyasor J, Raborife M, Ranchod P (2020) Sentiment analysis as an indicator to evaluate gender disparity on sexual violence tweets in South Africa. In: 2020 international SAUPEC/RobMech/PRASA conference, Cape Town, South Africa
25. Joshi PA, Simon G, Murumkar YP (2018) Generation of brand/product reputation using Twitter data. In: 2018 international conference on information, communication, engineering and technology (ICICET), Pune
26. Wang W, Li B, Feng D, Zhang A, Wan S (2020) The OL-DAWE model: tweet polarity sentiment analysis with data augmentation. In: IEEE Access, vol 8, pp 40118–40128

27. Li YM, Shiu YL (2012) A diffusion mechanism for social advertising over microblogs. Decis Support Syst 54:9–22
28. Du J et al (2013) Box office prediction based on microblog. In: Expert systems with applications
29. Al-Moslmi T, Omar N, Abdullah S, Albared M (2017) Approaches to cross-domain sentiment analysis: a systematic literature review. IEEE Access 5:16173–16192
30. Li SK, Guan Z, Tang LY et al (2012) Exploiting consumer reviews for product feature ranking. J Comput Sci Technol 27(3):635–649