International Journal of Computing and Business Research (IJCBR)
ISSN (Online) : 2229-6166
International Manuscript ID : 22296166V8I1201802
Volume 8 Issue 1 January - February 2018

Effectual Implementation of Emotions Mining and Predictive Analytics from Twitter Social Media

Amit Singla [1], Prof. S. S. Sarangdevot [2], Dr. Vishal Goar [3]

[1] Research Scholar
Janardan Rai Nagar Rajasthan Vidyapeeth University
Udaipur, Rajasthan, India

[2] Professor
Janardan Rai Nagar Rajasthan Vidyapeeth University
Udaipur, Rajasthan, India

[3] Assistant Professor
Department of Computer Applications
Government Engineering College, Bikaner, Rajasthan, India

Abstract

The domain of sentiment extraction or user belief mining is in research now days with the frequency usage of social media for assorted segments. Using sentiment extraction and text mining, the overall score of prominence can be evaluated with effectual predictions based results. Twitter is one of the leading and prominent social media platforms that is used for the distribution, dissemination and broadcasting of views in multiple formats. A number of celebrities, political speakers, leaders and key personalities are using Twitter so that their views and sentiments can be transferred to the whole world. Even the media groups and news channels are using Twitter for the distribution of news in form of tweets to all the devices and handhelds. In this research work, an effectual approach for the mining of social media tweets is presented so that the understandable as well as prediction based popularity extraction can be

implemented. The present work is based on the matching of positive and negative words from social media tweets which are proposed to be stored in a database engine so that the overall performance of database system with the real time data can be evaluated. On life fetching the data from Twitter Servers, the following database fields can be accessed and stored to the database table so that the overall performance in the storing as well as retrieval can be done. The proposed model is an effective and performance aware approach for opinion mining and the analysis of user timeline from assorted social media so that a common platform or application do not repeatedly require the sign on. Using this approach, the user can be identified on the heterogeneous media platforms and identity can be mined. Once the identity of user is mined, the further creation and activation of new account will not be required. Using this approach, the performance, complexity and time can be optimized a lot with huge optimization factors. The work begins with the experimentation done on the fetched tweets according to the categories. In addition thorough analysis of different tweets is done using deep mining and association of tweets and tokens. Further the proposed technique is being compared with the existing techniques. The proposed model is an performance aware approach for opinion mining and the analysis of user timeline from assorted social media so that a common platform or application do not repeatedly require the sign on. Using this approach, the user can be identified on the heterogeneous media platforms and identity can be mined. Once the identity of user is mined, the further creation and activation of new account will not be required. Using this approach, the performance, complexity and time can be optimized a lot with huge optimization factors. This work is having unique methodology of rule mining and association rules generation so that performance and accuracy aware predictions using escalated text mining can be done with higher degree of optimization in the results and predictive analysis.

Keywords : Evaluation of Tweets, User Mining of Twitter, Sentiment Mining

Introduction

Big Data Analytics is one of the key areas of research with assorted approaches in data science and predictive analysis. A number of scenarios exist where enormous data is logged every day and needs deep evaluation for research and development. In Medical Science, there are enormous examples where

processing, analysis and predictions from huge amount of data is required regularly. As per the reports from *First Post*, the data of more than 50 Peta Bytes are generated from each hospital of 500 beds in USA. In another research, it is found that one gram of DNA is equivalent to 215 petabytes in digital form. In another scenario of digital communication, the number of smart wearable gadgets increased from 26 millions in year 2014 to more than 100 millions in year 2016.

A prominent neuroscientist *Ann-Shyn Chian* from Taiwan presented in a research that more than 1 GB per cell of brain will be required even for a very small creature on this earth. For imaging of more than 80 billion neurons in human brain, it will take around 17 million years. Now, the volume, velocity, variety of medical data can be imagined with these data figures.

Creature	**No. of Neurons in Brain / Nervous System**
Fly	1,35,000
Cockroach	1,000,000
Ant	2,50,000
Honey Bee	9,60,000
Cat	760,000,000
Monkey	3,246,000,000
Macaque	6,376,000,000
Human	86,000,000,000

Here, the key question comes on the evaluation of huge amount of data with enormously growing speed. To preprocess, analyze, evaluate and predict on such big data based applications, there is need to use high performance computing frameworks and libraries so that processing power of computers can be utilized with maximum throughput and performance.

Free and Open Source Big Data Processing Tools

- Apache Storm
- Apache HADOOP
- Lumify
- HPCC Systems
- Apache Samoa
- ElasticSearch
- RapidMiner
- R-Programming
- Scribe
- NoSQL Databases

Twitter might be described as a real-time, highly social microblogging service that allows users to post short status updates, called *tweets*, that appear on

timelines. Tweets may include one or more entities in their 140 characters of content and reference one or more places that map to locations in the real world. An understanding of users, tweets, and timelines is particularly essential to effective use of Twitter's API, so a brief introduction to these fundamental Twitter Platform objects is in order before we interact with the API to fetch some data. We've largely discussed Twitter users and Twitter's asymmetric following model for relationships thus far, so this section briefly introduces tweets and timelines in order to round out a general understanding of the Twitter platform .

Tweets are the essence of Twitter, and while they are notionally thought of as the 140 characters of text content associated with a user's status update, there's really quite a bit more metadata there than meets the eye. In addition to the textual content of a tweet itself, tweets come bundled with two additional pieces of metadata that are of particular note: *entities* and *places*. Tweet entities are essentially the user mentions, hashtags, URLs, and media that may be associated with a tweet, and places are locations in the real world that may be attached to a tweet. Note that a place may be the actual location in which a tweet was authored, but it might also be a reference to the place described in a tweet."

Text mining is the application of natural language processing techniques and analytical methods to text data in order to derive relevant information. Text mining is getting a lot attention these last years, due to an exponential increase in digital text data from web pages, google's projects such as google books and google ngram, and social media services such as Twitter. Twitter data constitutes a rich source that can be used for capturing information about any topic imaginable . This data can be used in different use cases such as finding trends related to a specific keyword, measuring brand sentiment, and gathering feedback about new products and services.

Key attributes of Tweets are the following:

- text: the text of the tweet itself
- created_at: the date of creation
- favorite_count, retweet_count: the number of favourites and retweets
- favorited, retweeted: boolean stating whether the authenticated user (you) have favourited or retweeted this tweet
- lang: acronym for the language (e.g. "en" for english)

- id: the tweet identifier
- place, coordinates, geo: geo-location information if available
- user: the author's full profile
- entities: list of entities like URLs, @-mentions, hashtags and symbols
- in_reply_to_user_id: user identifier if the tweet is a reply to a specific user
- in_reply_to_status_id: status identifier id the tweet is a reply to a specific status

All the *_id fields also have a *_id_str counterpart, where the same information is stored as a string rather than a big int (to avoid overflow problems). We can imagine how these data already allow for some interesting analysis: we can check who is most favourited/retweeted, who's discussing with who, what are the most popular hashtags and so on. Most of the goodness we're looking for, i.e. the content of a tweet, is anyway embedded in the text, and that's where we're starting our analysis. The analysis can be started by breaking the text down into words. Tokenisation is one of the most basic, yet most important, steps in text analysis. The purpose of tokenisation is to split a stream of text into smaller units called tokens, usually words or phrases. While this is a well understood problem with several out-of-the-box solutions from popular libraries, Twitter data pose some challenges because of the nature of the language .

Novelty in the Proposed Research

- New classification algorithmic flow / approach is devised and implemented using prominent programming languages Java and Python
- The proposed work is having effectiveness in fetching live tweets on any combinations of the keywords or hashtags
- The proposed work is effective in terms of getting the user timeline and followers' information which is useful in the prediction and current analysis.
- The proposed work and implementation is effective in decision making and remarkable future predictions.

There is huge scope of research and development using Python scripts and specialized APIs for assorted applications including cyber security, data mining, Internet of Things, cloud simulation, grid implementation and many others. Python is one of the effective programming languages that can process and handle any type of data

International Journal of Computing and Business Research (IJCBR)
ISSN (Online) : 2229-6166
International Manuscript ID : 22296166V8I1201802
Volume 8 Issue 1 January - February 2018

stream. In the proposed work, the live fetching of social media messages shall be integrated with assorted advance data mining tools for effective prediction and modeling.

i. The social media is traditionally used for personal groups, talks and distribution of the thoughts, emotions and related aspects.

ii. Currently, the news and media agencies are using social media for generation of breaking and crisis related news after fetching the tweets from global locations.

iii. In this project work, an effective approach is proposed which is being used by the media and government agencies to find out the most interesting patterns from live tweets

iv. In addition, these live tweets can be used for debate information including finding out any particular person, event, news or related patterns.

v. Twitter and other social media apps can be used and implemented for detailed investigation and prediction on any event or news.

Database structure for mytwitter

Column	**Type**
Id	int(11)
Username	varchar(255)
Twittertimestamp	varchar(255)
Twittertext	Text
Screenname	varchar(255)
Followers	varchar(255)
Friends	varchar(255)
Listed	varchar(255)

Existing Base Work	Proposed Approach
50	70

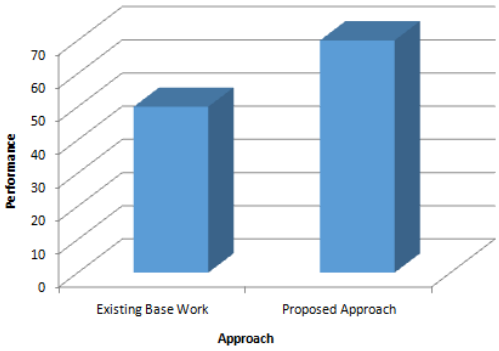

Figure 1. Classical and Proposed approach (Performance in Percentage)

Existing Base Work	Proposed Approach
90	60

Comparison of Existing Base Work and Proposed Approach

Existing Base	**Proposed Approach**

International Journal of Computing and Business Research (IJCBR)
ISSN (Online) : 2229-6166
International Manuscript ID : 22296166V8I1201802
Volume 8 Issue 1 January - February 2018

Work (Overall Effectiveness)	(Overall Effectiveness)
50	60
60	88
70	89
50	69

Table underlines the comparative results in terms of effectiveness. The effectiveness in taken and logged in terms of percentage and its evident from the results that the results of proposed approach is better than the classical approach.

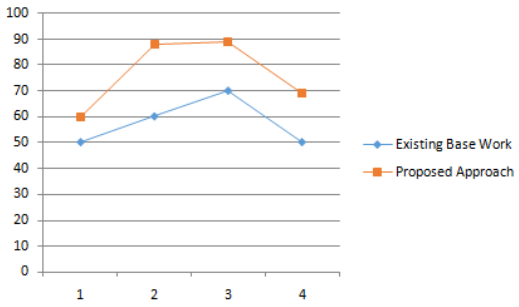

Figure 2. Effective Comparison of Classical and Proposed Algorithm

It is evident from the above mentioned figures and graphical results that the proposed approach is hugely effectual in terms of assorted parameters including performance, effectiveness and other related dimensions.

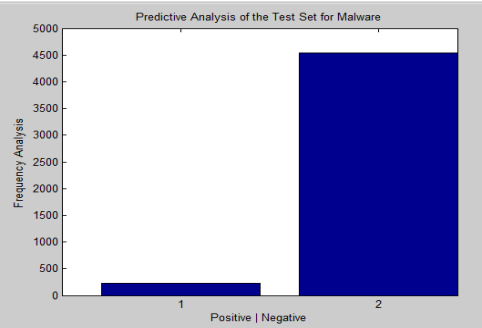

Figure 3. Predicted Positive and Negative Probabilities for Sample dataset

PREDICTION USING SAMPLE DATASET- 2:

```
mypredictions = sim(net, sample2);
sentiment_social_media_positive=0
sentiment_social_media_negative=0
for i=1:4632
if (mypredictions(i)>=0.5)
sentiment_social_media_positive=sentiment_social_media_positive+1;
   disp('Sentiment_social_media')
else
sentiment_social_media_negative=sentiment_social_media_negative+1;
disp('Not Sentiment_social_media')
end
end
disp('Predictions of Sentiment_social_media - Positive')
sentiment_social_media_positive
```

```
disp('Predictions of Sentiment_social_media - Negative')
sentiment_social_media_negative
data=[sentiment_social_media_positive sentiment_social_media_negative]
bar(data)
title('Predictive Analysis of the Test Set for Sentiment_social_media')
xlabel('Positive | Negative')
ylabel('Frequency Analysis')
```

Results from the Base Work

Size of Dataset	Accuracy - Previous Technique	Accuracy - Dynamic Density Based Clustering
50	77	94
100	78	95
200	79	95.8
300	87	96.8
400	89	95.8
1000	91	95.7
5000	92.55	95.7
10000	95.677	97

It is evident from the results of classical approach in increasing number of records in the dataset that the highest accuracy achieved is 97%. This work is making use of dynamic clustering based approach for sentiment_social_media detection and predictive analysis and the scope of future work is focusing on the use of ANN which is used in our proposed work.

Comparison of Accuracy (%) between Classical and Proposed Approach

Scenario	CLASSICAL Density Based Clustering Approach	PROPOSED ANN Based Approach
1	94	99.7
2	95	99.7
3	95.8	99.9
4	96.8	99.9
5	95.8	99.9
6	95.7	99.9
7	95.7	99.9
8	97	100

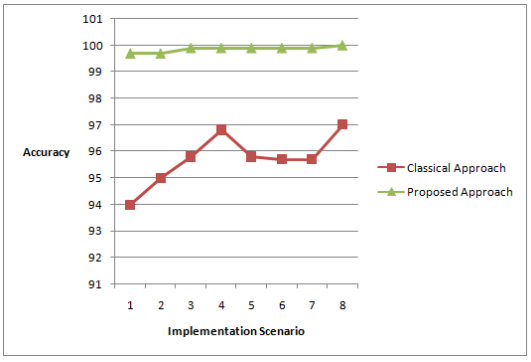

Figure 4. Graphical View of comparison on Accuracy (%) Parameter between Classical and Proposed Approach

In the graphical and tabular comparison, it is evident that with the increasing order of datasets and neurons, the performance of proposed approach is effective and moving towards 100% as compared to maximum 97% accuracy level in the earlier approach. The proposed approach is better in terms of faster execution and minimum error rate which is generally required in the fault tolerant and security specific domains.

The classical work done on the domain of alert files analysis and predictions are associated with classical data mining approaches which include clustering, association rule mining, classification or visualization.

Conclusion

Sentiment analysis, also referred to as Opinion Mining, implies extracting opinions, emotions and sentiments in text. One of the most common applications of sentiment analysis is to track attitudes and feelings on the web, especially for tacking products, services, brands or even people. The main idea is to determine whether they are viewed positively or negatively by the viewers or users on the social media. Twitter is a popular micro blogging service where users create status messages (called "tweets"). These tweets sometimes express opinions about different topics. The work builds an automatic sentiment (positive or neutral or negative) extractor from a tweet. This is very useful because it allows feedback to be aggregated without manual intervention. Using this analyzer, Consumers can use sentiment analysis to research products or services before making a purchase. Marketers can use this to research public opinion of their company and products, or to analyze customer satisfaction. Organizations can also use this to gather critical feedback about problems in newly released products. Fetching the live social media or related dimension sentiment analysis is under research from a long time for detailed analysis and prediction of the events with respect to the social cause. In this research work, the live extraction of timeline from social media platforms are implemented so that a common as well shared dataset can be prepared for future login and predictive analysis of the user behavior. In this work the key focus rely on the fetching of Twitter Timelines with the usage of SDK and API for research and development and real time

dataset can be evaluated for predictive analysis. In this research work, the live extraction of timeline from social media platforms of Twitter and Related Perspectives are implemented so that a common as well shared dataset can be prepared for future login and predictive analysis of the user behavior. In this work the key focus rely on the fetching of Twitter and Facebook Timelines with the usage of SDK and API for research and development and real time dataset can be evaluated for predictive analysis. The recommendation associated with this work is towards the predictive mining on assorted events, celebrities or popularity factors in real time domain. The predictions associated with business including stock market can be done effectually with this implementation. There are number of optimization approaches using which the efficiency, accuracy and performance factors can be improved. The integration of soft computing approaches are prevalent in the research community which provides fuzzy based execution and global optimization from existing results.

References

[1] Bifet, A., & Frank, E.. Sentiment knowledge discovery in twitter streaming data. In *International Conference on Discovery Science*. Springer Berlin Heidelberg, 2010

[2] Bollen, J., Mao, H., & Pepe, A.. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *ICWSM*, 11, 450-453, 2009

[3] Bollen, J., Mao, H., & Pepe, A.. Determining the Public Mood State by Analysis of Microblogging Posts. In *ALIFE* (pp. 667-668), 2010

[4] Asur, S., & Huberman, B. A.. Predicting the future with social media. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 *IEEE/WIC/ACM International Conference on* (Vol. 1, pp. 492-499). IEEE, 2010

[5] Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., & Li, P.. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1397-1405). ACM, 2011

[6] Saif, H., He, Y., & Alani, H.. Semantic sentiment analysis of twitter. In *International Semantic Web*

Conference (pp. 508-524). Springer Berlin Heidelberg, 2012

[7] Leong, C. K., Lee, Y. H., & Mak, W. K.. Mining sentiments in SMS texts for teaching evaluation. *Expert Systems with Applications*, 39(3), 2584-2589, 2012

[8] Wang, H., Cambria, E., Schuller, B., Liu, B., & Havasi, C.. Knowledge-based approaches to concept-level sentiment analysis. *IEEE Intelligent Systems*, 28(2), 12-14, 2013

[9] Dong, H., Shahheidari, S., & Daud, M. N. R. B.. Twitter sentiment mining: A multi domain analysis. In Complex, Intelligent, and Software Intensive Systems (CISIS), 2013 *Seventh International Conference* (pp. 144-149). IEEE, 2013

[10] Cambria, E., Fu, J., Bisio, F., & Poria, S.. AffectiveSpace 2: Enabling Affective Intuition for Concept-Level Sentiment Analysis. In *AAAI* (pp. 508-514), 2013

[11] Kotwal, A., Fulari, P., Jadhav, D., & Kad, R.. Improvement in Sentiment Analysis of Twitter Data Using Hadoop. *Imperial Journal of Interdisciplinary Research*, 2(7), 2014

[12] Poria, S. Cambria, E., Fu, J., Bisio, F., &. AffectiveSpace 2: *Enabling Affective Intuition for Concept-Level Sentiment Analysis*. In AAAI (pp. 508-514), 2015

[13] Guo, Y., Rao, J., Cheng, D., & Zhou, X.. ishuffle: Improving hadoop performance with shuffle-on-write. *IEEE Transactions on Parallel and Distributed Systems, 2016*

[14] Davis, B., Hürlimann, M., Cortis, K., Freitas, A., Handschuh, S., & Fernández, S.. A Twitter Sentiment Gold Standard for the Brexit Referendum. In *Proceedings of the 12th International Conference on Semantic Systems* (pp. 193-196). ACM, 2016

[15] Martínez-Cámara, E., Martín-Valdivia, M. T., Urena-López, L. A., & Montejo-Ráez, A. R.. Sentiment analysis in Twitter. *Natural Language Engineering*, 20(01), 1-28., 2014

[16] Abdul-Mageed, M., Diab, M., & Kübler, S.. SAMAR: Subjectivity and sentiment analysis for Arabic social media. Computer Speech & Language, 28(1), 20-37, 2014

[17] Hassan, S., Yulan, H., Miriam, F., & Harith A.. Contextual Semantics for

Sentiment Analysis of Twitter. *Elsevier*, 2015

[18] Statista. The Statistical Portal for Research and Data Analytics,2017

[19] Berry M, Linoff G. Mastering data mining: The art and science of customer relationship management. John Wiley & Sons, Inc.; 1999 Dec 1.

[20] Srivastava J, Cooley R, Deshpande M, Tan PN. Web usage mining: Discovery and applications of usage patterns from web data. Acm Sigkdd Explorations Newsletter. 2000 Jan 1;1(2):12-23.

[21] Ng RT, Han J. E cient and E ective Clustering Methods for Spatial Data Mining. InProceedings of VLDB 1994 Sep (pp. 144-155).

[22] Chen MS, Han J, Yu PS. Data mining: an overview from a database perspective. IEEE Transactions on Knowledge and data Engineering. 1996 Dec;8(6):866-83.

[23] Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Philip SY, Zhou ZH. Top 10 algorithms in data mining. Knowledge and information systems. 2008 Jan 1;14(1):1-37.

[24] Berkhin P. A survey of clustering data mining techniques. Grouping multidimensional data. 2006 Feb 8;25:71.

[25] Miller HJ, Han J, editors. Geographic data mining and knowledge discovery. CRC Press; 2009 May 27.

[26] Larose DT. Discovering knowledge in data: an introduction to data mining. John Wiley & Sons; 2014 Jun 2.

[27] Van der Aalst WM. Data Mining. InProcess Mining 2011 (pp. 59-91). Springer Berlin Heidelberg.

[28] Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical machine learning tools and techniques. *Morgan Kaufmann*; 2016

[29] Kim JD, Ohta T, Tateisi Y, Tsujii JI. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*. 2003 Jul 3;19(suppl_1):i180-2, 2003

[30] Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau R. Sentiment analysis of twitter data. InProceedings of the *workshop on languages in social media* 2011 Jun 23 (pp. 30-38). Association for Computational Linguistics, 2011

[31] Gräbner D, Zanker M, Fliedl G, Fuchs M. Classification of customer reviews based on sentiment analysis. *IEEE*, 2012

[32] Airoldi E, Bai X, Padman R. Markov blankets and meta-heuristics search: Sentiment extraction from unstructured texts. *International Workshop on Knowledge Discovery on the Web* 2004 Aug 22 (pp. 167-187). Springer, Berlin, Heidelberg, 2004

[33] Li Y, Jang J, Hu X, Ou X. Android malware clustering through malicious payload mining. *arXiv preprint* arXiv:1707.04795. 2017

[34] Ritter A, Etzioni O, Clark S. Open domain event extraction from twitter. InProceedings of the 18th ACM SIGKDD *International Conference On Knowledge Discovery And Data Mining* 2012 Aug 12 (pp. 1104-1112). ACM, 2012

[35] Go A, Huang L, Bhayani R. Twitter sentiment analysis. *Entropy*. 2009

[36] Sarlan A, Nadam C, Basri S. Twitter sentiment analysis. Information Technology and Multimedia (ICIMU), 2014 *International Conference* 2014 Nov 18 (pp. 212-216). IEEE, 2014

[37] Wang H, Can D, Kazemzadeh A, Bar F, Narayanan S. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. *Proceedings of the ACL* 2012 System Demonstrations 2012 Jul 10 (pp. 115-120). Association for Computational Linguistics, 2012

[38] Jose AK, Bhatia N, Krishna S. Twitter sentiment analysis. InSeminar Report, *National Institute of Technology* Calicut, 2010

[39] Pak A, Paroubek P. Twitter as a corpus for sentiment analysis and opinion mining. *InLREc*, 2010

[40] Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision. CS224N *Project Report, Stanford*, 2009

[41] Liu KL, Li WJ, Guo M. Emoticon smoothed language models for twitter sentiment analysis. *AAAI*, 2012

[42] Tatum J, Sanchez JT. Twitter Sentiment Analysis. CS29 Machine Learning course at *Stanford University*, 2013

[43] Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau R. Sentiment analysis of twitter data. *Proceedings of the workshop on languages in social*

media 2011 (pp. 30-38). Association for Computational Linguistics, 2011

[44] Baqapuri AI. Twitter Sentiment Analysis. *arXiv preprint* arXiv:1509.04219. 2015 Sep 14, 2015

[45] Saif H, He Y, Alani H. Alleviating data sparsity for twitter sentiment analysis. *CEUR Workshop Proceedings* (CEUR-WS. org), 2013

[46] Saif H, Fernandez M, He Y, Alani H. Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the *STS-Gold*, 2015

[47] Saif H, He Y, Alani H. Semantic sentiment analysis of twitter. *The Semantic Web–ISWC 2012*. 2012:508-24, 2012

[48] Jiang L, Yu M, Zhou M, Liu X, Zhao T. Target-dependent twitter sentiment classification. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*-Volume 1 2011 Jun 19 (pp. 151-160). Association for Computational Linguistics, 2011

[49] Mittal A, Goel A. Stock prediction using twitter sentiment analysis. *Stanford University, CS229*, 2012

[50] Poursepanj H, Weissbock J, Inkpen D. uOttawa: System description for SemEval 2013 Task 2 Sentiment Analysis in Twitter. SemEval@ *NAACL-HLT* 2013 Jun 14 (pp. 380-383), 2013

[51] Hassan A, Abbasi A, Zeng D. Twitter sentiment analysis: A bootstrap ensemble framework. Social Computing (SocialCom), 2013 *International Conference* on 2013 Sep 8 (pp. 357-364). IEEE, 2013

[52] Severyn A, Moschitti A. Twitter sentiment analysis with deep convolutional neural networks. Proceedings of the 38th *International ACM SIGIR Conference on Research and Development in Information Retrieval* 2015 Aug 9 (pp. 959-962). ACM, 2015

[53] Wang X, Wei F, Liu X, Zhou M, Zhang M. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. *Proceedings of the 20th ACM international conference on Information and knowledge management* 2011 Oct 24 (pp. 1031-1040). ACM, 2011

[54] Rosenthal S, Ritter A, Nakov P, Stoyanov V. SemEval-2014 Task 9: Sentiment Analysis in Twitter. SemEval@ COLING Aug 23 (pp. 73-80), 2014

[55] Cui A, Zhang M, Liu Y, Ma S. Emotion tokens: Bridging the gap among multilingual twitter sentiment analysis. Information retrieval technology. *IEEE* 2011:238-49.