

Analysis of Users' Interest Based on Tweets

Nimita Mangal¹, Rajdeep Niyogi^{1(✉)}, and Alfredo Milani²

¹ Department of Computer Science and Engineering,
Indian Institute of Technology Roorkee, Roorkee 247667, India
nimitamangal@gmail.com, rajdpfec@iitr.ac.in

² Department of Mathematics and Computer Science,
University of Perugia, Perugia 06123, Italy
milani@unipg.it

Abstract. Analysis of tweets would help in designing smart recommendation systems. Analysis of twitter messages is an interesting research area. Sentiment analysis of tweets has been done in some works. Another line of work is the classification of tweets into different categories. However, there are few works that have considered both sentiment analysis and classification to find out users' interest. In this paper, we propose an approach that combines both sentiment analysis and classification. Thus we are able to extract the topic in which users are interested. We have implemented our algorithm using five lakhs of tweets and around one thousand of users. The results are quite encouraging.

Keywords: Sentiment analysis · Twitter user · Social media

1 Introduction

Twitter is one of the popular online social blog where many celebrities post tweets for their fans and also post something related to an event. Twitter is a microblogging service. It is so called by this name because it enables users to send and read a short text message which is known as "tweet". There are 316 million monthly active users on twitter and 500 million tweets are posted per day. We can use these tweets for analyzing the interest of users and get to know the trends going on at any place. Such analysis may help in designing a smart recommendation system.

Several works have been done in the field of social networking, namely, classification of gender, classification of the topic, sentiment analysis of twitter users based on tweets, event detection, community detection, etc. Most of the work on recommendation system is based on network topology. A user's knowledge with social sites could be remarkably improved if other information like demographic attributes and user's personal interest and the interest of other users are considered. Such information allows users to follow a post or a user according to her topic of interest and the user may join a particular community of their own interest.

Moreover, a user may be interested to get recommendations based on her areas of interest. The recommendation first requires to know the user behavior. A person gets information about any event through newspaper, television, social sites or with the people around them. Now, if a person is interested in that event than she may tweet on

twitter about the event positively or negatively according to her viewpoint. To get this negative and positive viewpoint of the user, sentiment analysis of tweets is necessary. The topic to which a particular tweet belongs is done by categorization and through this we get to know the topic in which the user is interested. By applying both the techniques we can provide better recommendations to users.

In this paper, we make an attempt to come up with a method for analyzing the interest of users based on sentiments (positive, negative or neutral) and the topic to which tweets are related to get the correct positive or negative interest of users. We are particularly interested in users and their tweets to help them to give a better recommendation which they need according to their current interest.

Tweets are collected using the Twitter4j api in Java. Sentiment analysis has been done using Stanford core NLP integrated framework. A core NLP tool pipeline code is run on tweets. Sentiment score is computed based on the words composes and longer phrase. Classification of tweets to which topic it is related has been done using the matching of words with a topic.

The paper is organized as follows: Sect. 2 describes the related work. Section 3 describes our method for analyzing tweets. Section 4 describes the implementation details and results obtained by our method. Conclusion and future work are given in Sect. 5.

2 Related Work

Different methods are proposed for sentiment analysis, finding sentiments in words, sentences, sentiments in topics. Some of these approaches use machine learning, pattern based and natural language processing. Hybrid classifiers are designed in [1] to get better sentiment results. Sentiment analysis of twitter data is studied in [11] and it introduces POS-specific earlier polarity feature and explore the use of tree kernel. Experiments were performed on three models [11]: feature based model uses hundred features only and have the same accuracy as that of unigram model that uses ten thousand features. Kernel tree based model first tokenize the tweet into a tree by separating punctuation mark, exclamatory mark, negation word and emoticon and prior calculate the polarity of word using word-net dictionary. The unigram model is used as a baseline for the experiments.

Two approaches (machine learning and lexical approach) are suggested for sentiment analysis in [3]. First, the machine learning approach which first convert each text into a list of words, consecutive word pairs and consecutive word triplets and then based upon some human coded set of texts 'learn' which of these features tends to associate with sentiment scores to classify the new cases. Second, the lexical approach, uses some grammatical structure of language and some list of words with sentiment scores and polarities is used. The accuracy of both approaches depends on the training set and the score, which is already provided for most of the words.

Sentiment tree bank approach is suggested for sentiment analysis in [2]. The recursive neural network approach computes parent node vectors in bottom-up fashion and use a composition function g and node vector is featuring for that node. An approach

for computing sentiment score of short, informal text and sentence that contain phrases within it is suggested in [12].

Many recommender systems provide recommendation using the information based on user profile. A method for user recommendation is suggested in [4] and the method is based on sentiment volume objectivity. User profiling is done and similarity measure is computed between users (similarity measures based on place, sentiments of tweets). A method for friend recommendation is suggested in [5] which uses collaborative filtering and graph structure. Semantic user modeling has been done based on twitter posts in [6]. They suggested a formula for users similarity which is based on topic discussed by the users. A method to predict which political party a twitter user belongs is suggested in [7]. Their approach is based on certain characteristic of parties like activity, influence, structure and interaction, context and sentiment and then user classification has been done based on Bayesian classification.

Analysis over user intentions has been done in [8] that are associated at a community level and show how users with similar intentions connect with each other. The task of user classification in social media using machine learning framework is addressed in [9]. User profile features such as followers, friends, username, user-location are collected to know about a user. Tweets of user are collected for judging the behavior of users and to classify users of same types.

Two methods for classification of the Twitter trending topic are proposed in [10] first, based on textual information and the other based on the network structure. In text based model all the hyperlinks are removed from the tweet and then a tokenizer removes stop words and delimited character. Since there is a limitation of 140 characters in a tweet, people use acronyms for words and so a vocabulary is used that has the full form of these words (e.g. BR is used to represent best regard). The network based approach uses a similarity model to find out the trending topic say X. It searches for five topics that are similar to the topic X and finds out the similarity index. Text categorization method is proposed in [14] that uses support vector machines and gives proof both theoretically and logically that svm is well suited for text classification.

Most of the above works are related to sentiments, recommender systems, and trending topic. However, these works do not discuss about a user's interest on the topic being discussed by the users. Our approach is different from others as we compute the interests of a particular user and the users of a certain location by taking their sentiments (positively or negatively inclined) towards certain topics.

If we do only sentiment analysis on tweets, it gives the sentiments of users, whether she tweets positive or negative for any event happened on social media. Based on this, we do not get any idea about a user's interest. If we do only topic categorization then we get to know about the topic on which a user tweeted. It does not provide us the information that the user is positively interested on a certain topic or not. Hence, doing sentiment analysis and topic categorization separately, do not provide us the result for the user's interest. Thus, in this paper, we combine these two techniques that gives better results.

3 Proposed Methodology

Figure 1 shows the basic flow diagram of our method. First, we have collected the tweets of the user for knowing the interest according to her tweet. Sentiment analysis has been done on the tweets that are collected to know the inclination of users, whether she is positively indicated his sentiments over a particular topic or not.

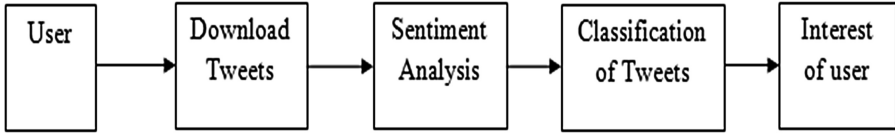


Fig. 1. Overview of our method.

We have used matching of words for classification of tweets to categorize it under a certain label (like sports, politics, entertainment, technology, hospitality, etc.). Finally, interest of a user is obtained that shows the positive inclination of the user towards a certain topic.

3.1 Data Collection

We have collected data for three different problems using Twitter 4j API and implementation results on this data are given in Sect. 4. The first problem deals with the user interest and for this we have collected 2,31,750 tweets of 1,150 users and show their behavior in the form of pie-chart. Tweets for different users were collected for different spans of time period for comparing their interest in different time intervals. The second problem deals with different cities of India in which we have collected around 2,02,578 tweets of 5 different cities of India and show the interested topic at that location. Tweets were collected by taking the value of latitude and longitude of the city. The third problem deals with the comparison of tweets of two countries (India and America). Tweets of the users that act as a bot (bot user is a user whose tweet done automatically by machines and not by persons) like news channel (bbcnews, indiatoday, etc.) are collected. The reason to choose bot user is to get more news about the country and to obtain interest for the country.

3.2 Sentiment Analysis of Tweet

Sentiment analysis is done, using the Stanford coreNLP method. This method is appropriate for short text. One drawback of this method, it is not considering emoticons value and acronym value. To solve this problem, first we check for emoticons and acronyms in the tweet. If it is present we compute the sentiment score accordingly for both. NLP provides some analyzing tools and has some implemented module that tag the words in a sentence, whether they are name of place, people, etc. or belong to noun,

verb, and adjective. These analyzing tools include the parser, sentiment analysis, named entity recognizer, open information extraction tools, etc.

We refine the tweet by removing all hashes, @ and extra spaces to make it more readable plain text. A static init method is called that set the properties to get to know what action is needed for an incoming text. In our case we set four properties, tokenize, ssplit, parse and sentiment. Tokenize property breaks the tweet into tokens. The tokenizer saves the offsets of each token from where it starts and ends. Ssplit property splits a sequence of tokens into sentences. Parse property generates the parse tree, based on some grammatical structure and language information to distinguish between phrases, subject and predicate in a sentence. Sentiment property is used to compute the sentiment score of a tweet, a binarized tree form for a tweet based on positivity and negativity. After the init method findsentiment method is called that first make a labeled tree for a given tweet and based on tree find the sentiment score in the range of 0–4. Higher the value of the score represents the positive sentiment of a tweet.

3.3 Classification of Tweet

After sentiment analysis, tweets are classified according to topic to which it is related. The open NLP package is used for classification [13]. This package provides us a tagger file for tagging of sentences. MaxentTagger is a class used for tagging each word in a tweet with its corresponding form, whether it is an adverb, noun, adjective, etc. There are 36 taggers and each word in a tweet belongs to one of these taggers. After tagging a tweet word tagger pair is formed.

Each word tagger pair is compared with ten different categories of topics like entertainment, technology, politics, etc. For comparing word with the topic we are using wordnet similarity module that implements a variety of semantic similarity and relatedness measures which is based on information found in the lexical database WordNet. For using this WordNet similarity we are having WS4J API. For more accurate results we compare these words with the synonyms of topic for example, if we want to compare any word with technology then we compare word with technology, network, industry, etc. A method getSimilarity is available which compares this word with these topics and calculates some relatedness scores and gives a similarity score.

4 Implementation and Results

Below Fig. 2 shows the implemented flow diagram for our system. We have provided an interface to the user in which user provides a screenname (unique name given to each user on twitter) of the user and our backend system calls the download procedure that downloads the tweets of that user and after this sentiment analysis module is called which finds out the sentiment for each tweet and then each tweet is fall under one category positive or negative or neutral. After this classification module is called that runs for the positive and neutral sentiment tweet and gives percentage according to topic to which it belong. It is possible that one tweet belongs to more than one category. The final result for the user about the interested topic is shown in the form of a pie chart.

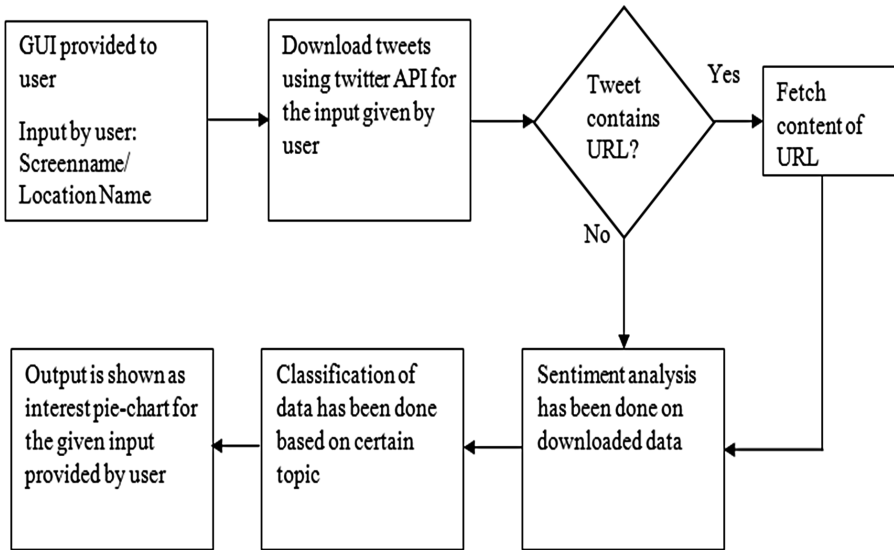


Fig. 2. Implemented flow diagram of our system.

If a tweet contains url then first the content of the link is fetched using some code for content fetching of the url and then on obtaining the plain text we apply sentiment analysis and classification algorithm.

4.1 User Interest

Using the user interface we have obtained the following results. Figure 3 shows the interest pie chart for the tweets done by Shreya Goshal, a popular singer in India, from 1-02-16 to 29-02-16 and this result shows that major topic in which she is interested is entertainment for this period of time. The values shown in the pie chart is in percentage and the entertainment culture topic is having the highest value of the interest that is 49.345 % and second interested topic is a hospitality recreation with 25.199 % value of interest.

Figure 4 shows the interest pie chart for the tweets done by Narendra Modi, Prime Minister of India, and this result shows that major topic in which he is interested are politics and social issues with 22.132 % value of interest. The second interested topic is business finance with 15.996 % value of interest.

4.2 Location Interest

We have collected tweets of different cities of India and show interested topic. Figure 5 represents the bar graph whose x-axis represents the topic name and the y-axis represents the value of interest in percentage. Through figure it has been seen that in Hyderabad most of the people do tweet that belongs to the business finance field with

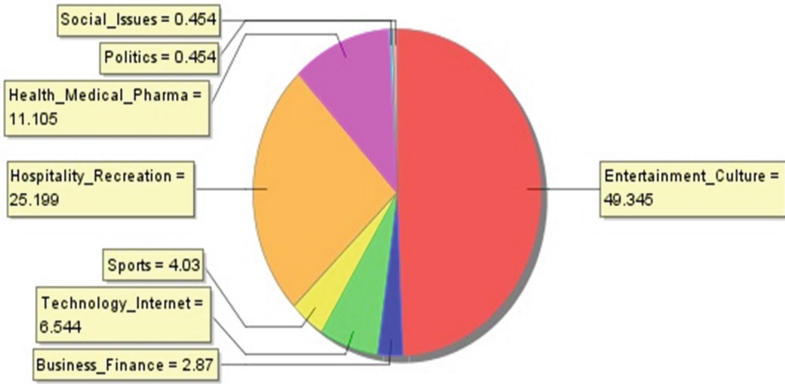


Fig. 3. Tweets done by Shreya Ghoshal with positive interest. (Color figure online)

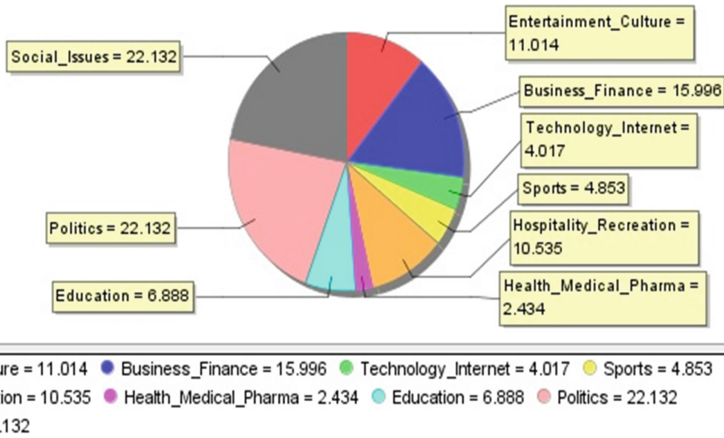


Fig. 4. Tweets done by Narendra Modi with positive interest. (Color figure online)

38.0413 % value of interest and second interested topic is technology with 12.9435 % value of interest.

Table 1 represents the value in percentage of the interest for certain topics that are listed in table for most of the famous cities of India.

Figure 6 shows the comparison results for different topics for five different cities of India. X-axis of the above graph represents the topic name and the y-axis represents the value of interest over certain topics in percentage. Bangalore is one of the cities where most of the multi-national companies are located and most of the business activities take place. The results are showing that among the five cities of India, tweets from

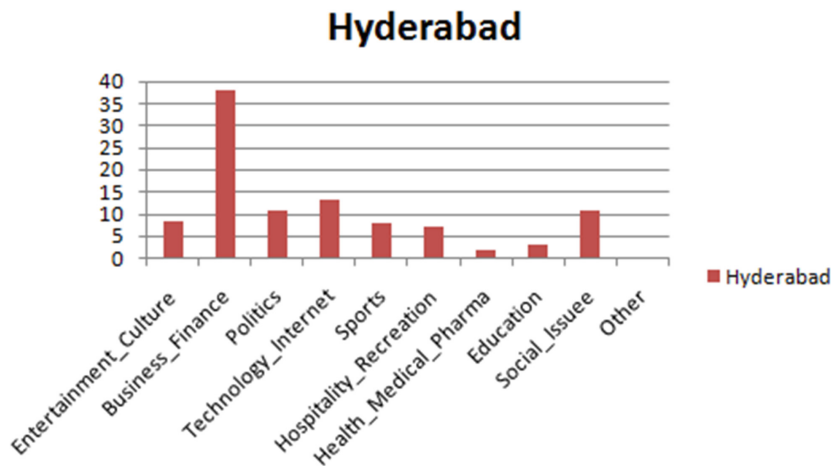


Fig. 5. A topic to which tweets belong to Hyderabad city.

Table 1. Comparison data for different cities.

City \ Topic	Hyderabad	Delhi	Bangalore	Chandigarh	Mumbai
Entertainment_Culture	8.0628	5.6234	3.9814	27.3675	22.517
Business_Finance	38.0413	22.0722	49.6758	3.3993	22.0722
Politics	10.805	22.5177	10.7765	12.2889	5.6234
Technology_Internet	12.9435	14.3932	16.2475	3.3759	14.3932
Sports	7.9552	4.6632	2.4733	27.367	4.6632
Hospitality_Recreation	6.9499	6.0001	4.0945	12.9485	6.0001
Health_Medical_Pharma	1.4342	0.9744	0.3051	2.5275	0.9744
Education	2.9985	1.2268	1.6645	0.3142	1.2268
Social_Issuee	10.805	22.517	10.7765	10.4095	22.5177
Other	0.0041	0.0107	0.0045	7.62E-04	0.0107

Bangalore are highly related to business-finance. Delhi, the capital of India, is a political hub where many politicians and youth that belong to some non- governmental organization (NGO) reside. From the results we infer that among the five cities, tweets from Delhi are mostly related to politics and social issues. It is useful to have such data, information because it provides us the trend of users at certain locations. Based on the trends of the twitter data, new products may be launched in certain locations.

4.3 Comparison of Data of Different Countries

Table 2 represents the value in percentage of the interest for certain topics that are listed in a table for two countries India and America. We have obtained these values by

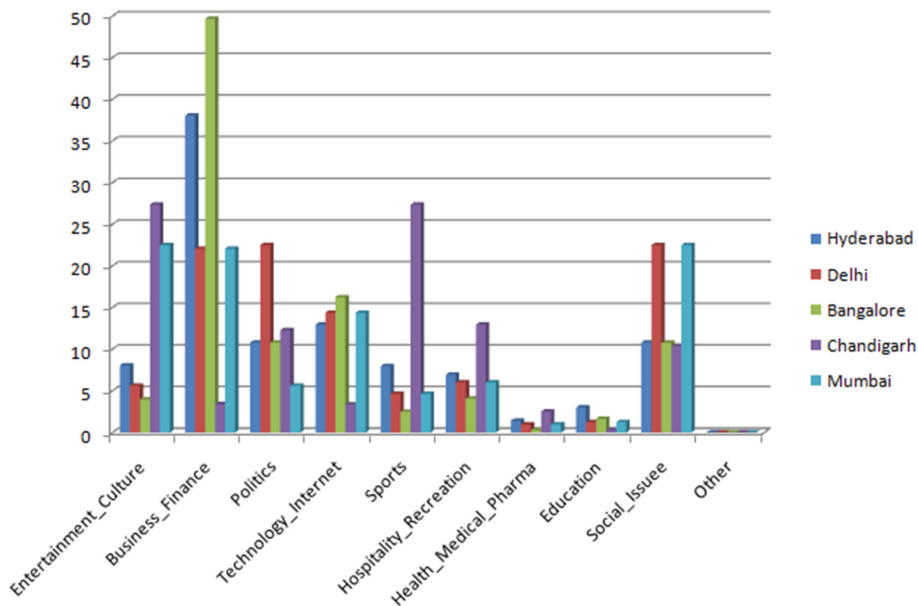


Fig. 6. Shows the comparison of different cities of India. (Color figure online)

collecting the tweets of different bot users of India and America and here we try to get the overall interest of Indian and American users and our analysis shows that most of the tweet done by Indian users are related to politics and social issues with 18.291 % value of interest and then other interested topic is hospitality recreation with 14.393 % value of interest.

Table 2. Comparison data for countries, India and America.

<div>Country</div> <div>Topic</div>	India	America
Entertainment_Culture	10.625	15.9
Business_Finance	9.5233	8.3479
Politics	18.291	5.8073
Technology_Internet	5.5157	14.754
Sports	12.85	14.464
Hospitality_Recreation	14.393	23.986
Health_Medical_Pharma	5.6224	7.3929
Education	4.8878	3.5413
Social_Issues	18.291	5.8073
Other	6.90E-05	

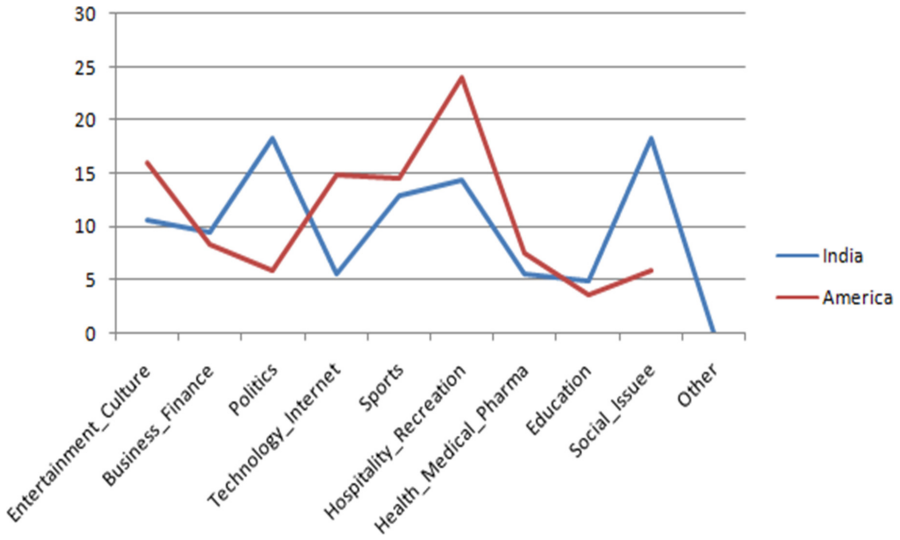


Fig. 7. Shows the comparison between tweets of India and America.

The tweets done by American users are mostly related to hospitality recreation with 23.986 % value of interest and then second major interested topic is entertainment with 15.9 % value of interest. Figure 7 shows the comparison of Indian and American users graphically based on the tweets collected for both the countries.

4.4 Discussion

Table 3 shows the comparison of the proposed approach with other works. Sentiment analysis suggested in [3] is not suitable for short text. Sentiment analysis suggested in [2] is suitable for short text, but they did not consider the sentiment score for the emoticon and acronym. We consider sentiments of emoticon and acronym in our approach. Thus, we obtain better results in tweets.

Table 3. Comparison with other works.

Paper \ Features	Sentiment Analysis	Classification	Users' Interest
R. Prabowol [1]	Yes	-	-
R. Socher [2]	Yes	-	-
M. Thelwall [3]	Yes	-	-
K. Lee [10]	-	Yes	-
This Paper	Yes	Yes	Yes

For classification, a network based approach is suggested in [10]; we have done classification by comparing the words of tweet with different categories. Since both approaches are different and so they cannot be compared. Users' interest is not computed in these works.

5 Conclusion and Future Work

In this paper, we have suggested an approach to analyze the user interest based on her tweets. In our approach, we combine sentiment analysis and classification of tweets. We consider the sentiments of emoticon and acronym. We have implemented our algorithm using five lakhs of tweets and around one thousand of users. We obtain promising results. As part of our future work we would like to develop a recommendation system based on the techniques suggested in this paper.

Acknowledgement. The authors thank the anonymous reviewers for their valuable suggestions that have helped in improving the paper.

References

1. Prabowo, R., Thelwall, M.: Sentiment analysis: a combined approach. *J. Informetrics* **3**(2), 143–157 (2009)
2. Socher, R., Perelygin, A., Wu, J.Y., Chuang, J., Manning, C.D., Ng, A.Y., Potts, C.: Recursive deep models for semantic compositionality over a sentiment Treebank. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642 (2013)
3. Thelwall, M.: Heart and soul: sentiment strength detection in the social web with SentiStrength. In: *Proceedings of the CyberEmotions*, pp. 1–14 (2013)
4. Gurini, D.F., Gasparetti, F., Micarelli, A., Sansonetti, G.: A sentiment-based approach to Twitter user recommendation. In: *Proceedings of 5th ACM RecSys workshop on Recommender Systems and the Social Web*, June 2013
5. Agarwal, V., Bharadwaj, K.K.: A collaborative filtering framework for friends recommendation in social networks based on interaction intensity and adaptive user similarity. *J. Soc. Netw. Anal. Min.* **3**, 359–379 (2013)
6. Abel, F., Gao, Q., Houben, G.-J., Tao, K.: Semantic enrichment of Twitter posts for user profile construction on the social web. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., Leenheer, P., Pan, J. (eds.) *ESWC 2011, Part II. LNCS*, vol. 6644, pp. 375–389. Springer, Heidelberg (2011)
7. Boutet, A., Kim, H., Yoneki, E.: What's in Twitter, I know what parties are popular and who you are supporting now! *J. Soc. Netw. Anal. Min.* **3**(4), 1379–1391 (2013)
8. Java, A., Song, X., Finin, T., Tseng, B.: Why we Twitter: understanding microblogging usage and communities. In: *Proceedings of 9th WebKDD and 1st SNA-KDD Workshop*, SanJose, California, USA, pp. 56–65, August 2007
9. Pennacchiotti, M., Popescu, A.: A machine learning approach to Twitter user classification. In: *Proceedings of the Fifth ICWSM*, pp. 281–288 (2011)

10. Lee, K., Palsetia, D., Narayanan, R., Patwary, M.M.A., Agrawal, A., Choudhary, A.: Twitter trending topic classification. In: Proceedings of 11th IEEE International Conference on Data Mining Workshops, pp. 251–258, December 2011
11. Kiritchenko, S., Zhu, X., Mohammad, S.M.: Sentiment analysis of short informal texts. *J. Artif. Intell. Res.* **50**(1), 723–762 (2014). USA
12. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of Twitter data. In: Proceeding of LSM 2011 Workshop on Languages in Social Media, pp. 30–38. Association for Computational Linguistics (2011)
13. Manning, D., Christopher, M., Surdeanu, J., Bauer, J., Finkel, S., Bethard, J., McClosky, D.: The stanford CoreNLP natural language processing toolkit. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55–60 (2014)
14. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: Proceedings of the 10th European Conference on Machine Learning, London, pp. 137–142 (1998)
15. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with Twitter: what 140 characters reveal about political sentiment. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (2010)

Computational Science and Its Applications – ICCSA
2016

16th International Conference, Beijing, China, July 4-7,
2016, Proceedings, Part V

Gervasi, O.; Murgante, B.; Misra, S.; Rocha, A.M.A.C.;
Torre, C.M.; Tanir, D.; Apduhan, B.O.; Stankova, E.;
Wang, S. (Eds.)

2016, XXVII, 636 p. 207 illus., Softcover

ISBN: 978-3-319-42091-2