Performing Interest Mining on Tweets of Twitter Users for Recommending Other Users with Similar Interests

Richa Sharma, Shashank Uniyal and Vaishali Gera

Abstract With an upsurge in the popularity of microblogging sites, Twitter has emerged as a huge source of assorted information. People often use Twitter to post their ideas and beliefs about the prevailing issues, feedbacks about products they use and opinions on the topics which appeal to them. Therefore, Twitter is considered to be one of the most appropriate virtual environments for information retrieval through data extraction as well as for analysis and drawing out inferences. This paper proposes a system that maintains a database of the Twitter users, fetches their areas of interests and accordingly recommends them the lists of other users with similar interests whom they may like to follow. The prototype of the system is developed in R and has been evaluated on various datasets. The results are promising and portray decent levels of accuracy, i.e., the proposed system is able to discover the correct area of interests of the users and accordingly make appropriate recommendations.

Keywords Twitter · Tweets · Information retrieval · Interest mining · R

1 Introduction

Due to increased use of social media applications, interest mining is gaining prominence as a hot topic of research. Interest mining involves techniques to extract information regarding the interests of people from texts, images, music playlists, etc. Users post their views on products and services used by them, opinions about political and religious issues or simply present some factual information related to their

R. Sharma · S. Uniyal · V. Gera (✉)
Department of Computer Science, Keshav Mahavidyalaya, University of Delhi,
New Delhi, Delhi, India
e-mail: vaishaligera95@gmail.com

R. Sharma
e-mail: rsharma@cs.du.ac.in

S. Uniyal
e-mail: uniyalshashank94@gmail.com

B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,
https://doi.org/10.1007/978-981-13-1708-8_55

interests. Such social media applications include blogs, bookmarks, communities, files, forums, microblogs, profile tags, wikis and so forth. Out of these, mircoblogs and profile tags most accurately reflect a person's area of expertise or interest [4].

Twitter is one such microblogging site with more than 313 million monthly active users from around the world [3]. This volume of Twitter users tweeting regularly on varied topics makes a rich repository of data available on Twitter for analysis and research purposes. Since Twitter data are abundant and freely available, researchers see it as a valuable source of input for their research in various subfields of data mining, [14] for example, sentiment analysis [7] and text mining [19]. Besides being popular among users, Twitter restricts the users to frame meaningful tweets within the limit of 140 characters, making the tweets easier to parse.

The aim of this research is to determine the areas of interest of a Twitter user on the basis of what the user posts frequently and accordingly suggest him people with similar interests he can follow. Through this, we bring together people with similar interests. The idea is to generate a list containing people sharing similar interests; this list is self-evolving such that as soon as the system finds the areas of interest of a user it makes a new entry for it.

The mentioned approach matches root words present in a particular tweet with a predefined list, and based on the number of matches the genre of the tweet is determined. For example, consider the following tweet by the cricket expert Harsha Bhogle:

> *So enjoyed watching @**ImZaheer bowl**. That first **inswinger to Rahane** was a classic. Wonder if there is another **IPL** left in him…*

Here, the terms—*ImZaheer, bowl, inswinger, Rahane, IPL*—are associated with cricket and hence can be categorized to be related to cricket and if more such tweets are found in his account, then it can be inferred that Harsha Bhogle is a cricket enthusiast.

The work done under this research can be divided into 3 sections:

(1) Applying parsing techniques on extracted tweets of Twitter users to find their interest areas and store this information in a database.
(2) Use the database having the information about interests of previous users to suggest every next user the list of people he can follow.
(3) Examining the accuracy of the algorithm.

Among the different software packages that can be used to analyze Twitter, R offers a wide variety of libraries and packages that meet the requirements of this research. R is open source and provides a large integrated collection of tools for data analysis. R is designed to interface well with other technologies that included programming languages and databases [5, 13].

For this research, Twitter API was used to collect a corpus of text posts from 16 Twitter users to users in accordance with the genre of their tweets mainly into two categories—(1) politics and (2) cricket (these two being the areas of interest catered in this research). For this, it was required to create two dictionaries containing the terminologies related to these fields.

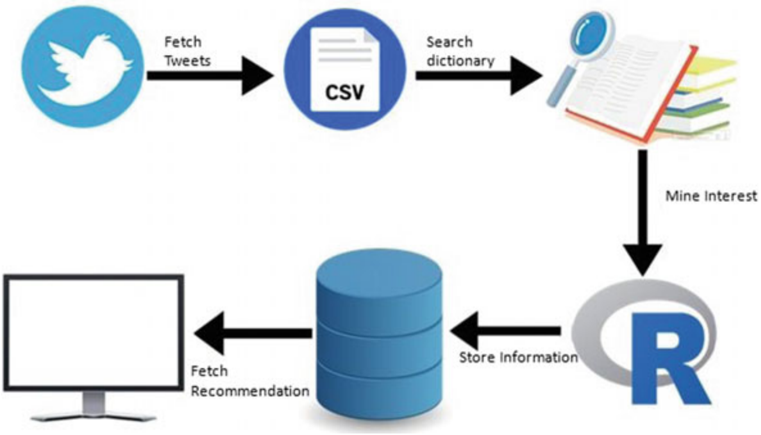

Fig. 1 Interest mining framework

Figure 1 illustrates the framework of the followed approach. First the tweets corresponding to a particular Twitter user are fetched and stored in a.csv file. Then, using various R libraries every tweet is split into individual words and these words are then compared with the predefined dictionaries to categorize the people on the basis of their interests and this information is stored in a MySQL database. Finally, in accordance with areas of interest of a user, a list of people is suggested whom he can follow. An entry of the current user is also made in the database such that he could also be recommended to other people making this system self-evolving.

The organization of the paper is as follows: The next section presents the related research work. Section 3 puts forth the proposed methodology. Section 4 discusses the results followed by challenges faced in doing this research and the work to be carried in future in Sect. 5. Section 6 concludes the paper.

2 Related Work

Traditionally, Twitter feeds have been used as a corpus for sentiment analysis and opinion mining as Twitter is used by people to express opinion about different topics. Twitter contains huge number of text posts which come from celebrities, company representatives, world leaders and general people. Thus, the data of Twitter become valuable for marketing and social studies. Prior work done in this field is related to classification of tweets as positive, negative or neutral. In [10], the author used "TreeTagger" for POS tagging and observed the patterns in distributions among positive, negative and neutral sets and concluded that emoticons and facts are stated by the use of syntactic structures. Read in [12] used emoticons and formed a training set for sentiment classification. For this purpose, the author used "usenet" newsgroups to get emoticons from texts. The dataset was divided into “positive” (happy emoticons) and “negative” (sad or angry emoticons) samples for application of machine learning techniques.

Twitter data has also proved its worth for evaluation of performance of different machine learning algorithms. In [1], the authors used emoticons as noisy labels and showed that machine learning algorithms (Naive Bayes, Maximum Entropy, and SVM) with certain preprocessing steps have 80% accuracy when trained with emoticon data.

In [2], authors have used Twitter to measure the popularity of a user and his influence on Twitter using 3 measures of influence—in degree, retweets and mentions. Through their research they found that popularity is not gained spontaneously but through continuous efforts.

Recently, mining the interests and areas of expertise of a Twitter user has gained prominence. Research scholars have used Twitter data (text, photographs) to extract the areas of interest of a person. In [4], the evaluation done by the authors compared the usefulness of eight different social media applications for mining expertise and interests. The results suggest that socialization sources such as people’s tag and blogs are more accurate for extracting the areas of interest/expertise in comparison to the collaborating sources, such as files and wikis. In [11], Qiu and Cho through their research tried to observe patterns in users’ past search histories to know their interests. Wang et al. [16], Wen and Lin [17] projected to deduce user interests from users’ social connections and interactions. Li et al. [8] used the information about places visited by people to mine their interests. In [6], Kim et al. categorized user interests by reading level and topic distributions. In [18], the authors studied the problem of interest mining from personal photographs. They proposed an approach of user image latent space model to model the user’s interest and image content. In [9], the authors suggest an approach named “*twopics*”, which characterizes users’ topics of interest, by recognizing the entities that appear very frequently in a tweet. The tweet is parsed for its entities which are disambiguated first and are then discovered (power becomes power play). The discovered entities are then used to determine the topics of interest. Their system was able to achieve 52.33% accuracy.

This paper proposes a system to find areas of interest of twitter users from what they post on Twitter and accordingly suggest them other Twitter users with similar interests whom they can follow. The application implements preprocessing of tweets of users to find their interest and then recommends to them other users with similar interests.

3 Proposed Methodology

The proposed methodology involves the stages as shown in Fig. 2.

The input of the Application is user Tweets. User Tweets are being fetched through Twitter API and twitteR library of R. The rest of the process is described in the following phases.

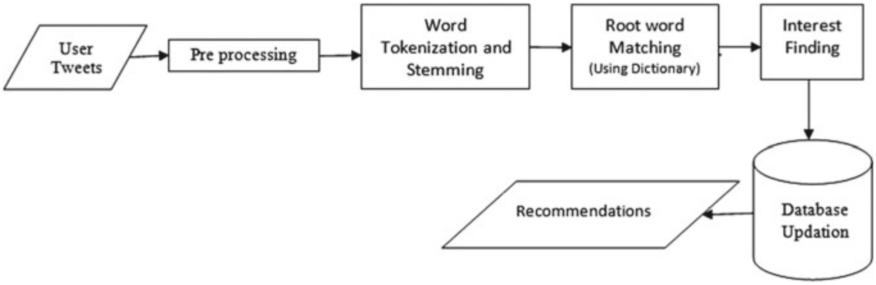

Fig. 2 Work flow diagram

A. *Preprocessing*: Preprocessing phase has to be done on tweets to clean and prepare them for classification with better accuracy. This cleanup is done by R's regex-driven global substitute, gsub(). Preprocessing is done by performing following operations:

1. **Stop words removal**: Since stop words (a, about, further, every, also, is) do not possess any relevant information, therefore to make searching process easy, they must be removed.
2. **Punctuation removal**: Punctuation characters (! " # $ % & ' () * +, − . / : ;<= >? @ [\] ^ _ ′ {′} ~) are removed from each individual tweet.
3. **Control words removal**: Since content of control words (\n, \r) determine action rather than meaning, they must be removed.
4. **Digits removal**: Digits are also removed to make a tweet concise having valuable information.

For example, in the following tweet:

No School Bag Day Will Break Rote Learning &amp; Foster All Round Development. Thumbs Up to the Idea @myogiadityanath.,

"no, will, &, ; , all, . , up, to, the, @" will be removed; hence, the output after the preprocessing steps will be:

school bag day break rote learning amp foster round development thumbs idea myyogiadityanath.

B. *Word tokenization and stemming*: For further processing, each Tweet is broken into individual words. These individual words are replaced with their root words using Porter's algorithm for matching with the dictionary [15]. The output of the above preprocessed tweet will be:

"*school*" "*bag*" "*day*" "*break*" "*rote*" "*learn*" "*amp*" "*foster*" "*round*" "*develop*" "*thumb*" "*idea*" "*yogiadityanath*".

C. *Root word matching*: The root words obtained from the previous step are compared with the dictionary containing the politics and cricketing terms. This gives a count for number of matches which is prerequisite for finding the areas of interest.

D. *Interest finding*: The number of tweets which do have some words that match to the terms in the dictionary are calculated and checked if it exceeds a threshold value to infer the areas of interest.

E. *Database updating*: After having known the interest of a user, an entry of the current user and his interest is made into the database.

Finally, on the basis of a person's interest a list of people who share the same interest is recommended to him.

4 Results

Unlike some other microblogging services, the data posted by different twitter users is freely available. This data can be used for creating standardized datasets for various purposes. For this research, Twitter feeds from 16 different Twitter users were used as the corpus. Among the 16 users, 10 were experts in cricket, while 6 held interest in politics. The algorithm was applied on this data to evaluate the accuracy measures.

Tables 1 and 2 depict statistical data for politics and cricket as areas of interest. The value for expected count was calculated manually. "Expected" count is the actual number of tweets belonging to either area of interest from the total fetched tweets, whereas the "measured" count is the number of tweets belonging to either interest as detected by the application. These two values are further used to measure the accuracy of the system.

The formula used to measure the accuracy of the classification process where the person's interests are being fetched is given as:

$$\text{Accuracy} = \frac{\text{Measured value}}{Expected\ Value} \times 100 \tag{1}$$

Table 1 depicts the accuracy measure for cricket. The results are encouraging and went up to 80% in many cases but at the same time dipped to 60% in some other cases.

From Table 2, it can be observed that the numbers for accuracy achieved for politics are better than in case of cricket. The accuracy was more than 90%, in fact it even went up to 100% in many cases.

The rows which do not have any values for accuracy are the cases where measured value exceeds the expected value, i.e., the number of tweets calculated by our application for that particular interest is more than the actual number of tweets for that interest.

Table 1 Statistical data for cricket-based tweets

User name	Expected count	Measured count	Accuracy for cricket (%)
Bhogleharsha	164	117	71.3
Cricketwallah	114	77	67.5
ArvindKejriwal	2	9	–
ShashiTharoor	9	8	88.8
Sanjaymanjrekar	19	16	84.2
Rgcricket	44	38	86.3
RajatSharmaLive	0	0	100
VijayGoelBJP	2	4	–
Sardesairajdeep	5	13	–
SeerviBharath	88	76	86.3
Mohanstatsman	64	37	57.8
Virendersehwag	17	20	–
Kp24	7	6,8	85.7
Narendramodi	0	0	100
Gauravkapur	35	32	91.4
Kartikmurli	4	4	100

Table 2 Statistical data for politics-based tweets

User name	Expected count	Measured count	Accuracy for politics (%)
Bhogleharsha	0	9	–
Cricketwallah	11	27	–
ArvindKejriwal	113	89	78.7
ShashiTharoor	75	72	96
Sanjaymanjrekar	0	2	–
Rgcricket	1	3	–
RajatSharmaLive	30	29	96.6
VijayGoelBJP	75	77	–
Sardesairajdeep	46	43	93.4
SeerviBharath	0	0	100
Mohanstatsman	3	4	–
Virendersehwag	2	6	–
Kp24	2	2	100
Narendramodi	30	27	90
Gauravkapur	0	5	–
Kartikmurli	0	0	100

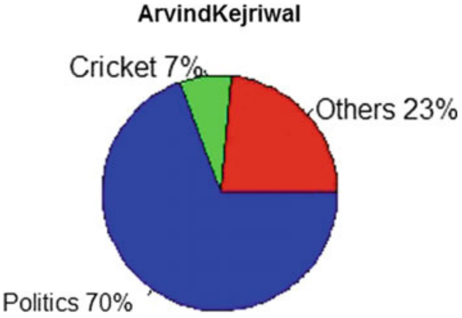

Fig. 3 Distribution of *ArvindKejriwal* tweets

For example, in case of *sardesairajdeep*, the value for accuracy for cricket is calculated as:

$$\frac{13}{5} \times 100 = 260$$

The value exceeds 100% and is thus ambiguous because there are certain words which find place in both the dictionaries (for example, the word "*power*" as "*power—play*" in cricket and simply "*power*" in politics) as a result such words affect the count for measured value.

Figure 3 illustrates the distribution of tweets for *ArvindKejriwal* which clearly shows his interest in politics.

Recall, Precision and F-score were used as metrics for evaluation of the application.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{2}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{3}$$

$$\text{F-score} = \frac{2 * Recall * Precison}{Recall + Precison} \tag{4}$$

Calculation of these is done using confusion matrix. In the confusion matrix, the "correct" cells are:

True Negative (**TN**): Case was negative and predicted negative, i.e., a user who did not have interest in a particular area was correctly identified as not having interest in that area.

True Positive (**TP**): Case was positive and predicted positive, i.e., a user who did have interest in a particular area was correctly identified as having interest in that area.

And the "error" cells are:

Table 3 Confusion matrix for cricket-based tweets

	Interested	Not interested
Interested	6(TP)	3(FN)
Not interested	1(FP)	6(TN)

Table 4 Confusion matrix for politics-based tweets

	Interested	Not interested
Interested	4(TP)	2(FN)
Not interested	0(FP)	10(TN)

False Negative (**FN**): Case was positive but predicted negative, i.e., a user who did have interest in a particular area was incorrectly identified as not having interest in that area.
False Positive (**FP**): Case was negative but predicted positive, i.e., a user who did not have interest in a particular area was incorrectly identified as having interest in that area.

Tables 3 and 4 show the confusion matrices for both areas of interest. They were calculated separately to know the levels of accuracy for both cases.

$$\text{Recall} = \frac{6}{6+3} = 0.66$$

$$\text{Precision} = \frac{6}{6+1} = 0.85$$

$$\text{F-score} = \frac{2 * 0.66 * 0.85}{0.66 + 0.85} = 0.74$$

$$\text{Recall} = \frac{4}{4+2} = 0.66$$

$$\text{Precision} = \frac{4}{4+0} = 1$$

$$\text{F-score} = \frac{2 * 0.66 * 1}{0.66 + 1} = 0.79$$

On evaluation the application gave better results in case of politics than cricket. This can be attributed to the fact that linguistics for politics is much less diverse than in case of cricket. There is a specific trend that can be observed in most politics genre tweets. For example, users mentioned names of political leaders and political parties either by twitter handle name or by hash tag. However, in cricket no such trend could be observed as every expert had his own creative way of expressing his views.

The cases that resulted in ambiguous values for accuracy were more in politics than cricket. This is because of the intersecting words in dictionaries of both areas of interest which resulted in ambiguity. For example, root word "*power*" is used as "*power play*" in cricket but only "*power*" in politics.

5 Challenges and Future Work

The application is a basic prototype, yet it generates a lot of encouraging results. It accurately retrieves the areas of interest of users and makes appropriate recommendations accordingly.

However, there are certain challenges associated with the application. The predefined dictionaries maintained for every interest (or field) need to be updated frequently and should be made as specific as possible. Even after listing almost all relevant terms specific to a particular field, there still remain words that can be found in more than two dictionaries leading to conflicting results. Since different people have different style of writing; we cannot define our dictionary in accordance with the choice of every single person. For example, Harsha Bhogle refers to the cricket team Mumbai Indians as *#mipaltan*, while Aakash Chopra refers to them as *#mi*. Therefore, automatic analysis of such diverse and ambiguous tweets poses a challenge [10].

Apart from this, there were certain limitations while performing Twitter analysis using R—firstly, the number of retrieved tweets was less than the number of requested tweets; secondly, the older tweets could not be retrieved.

In future, we plan to expand the system to improve upon the results by incorporating self-updating dictionaries, disambiguation via context, inclusion of third-party tools for better processing and integration with machine learning techniques.

6 Conclusion

Through this paper, we researched on how Twitter may prove to be a powerful source of data that can be analyzed to give out purposeful information. Each user on Twitter wants to follow people having similar areas of interest to stay updated on any information regarding the common area of interest, to formalize opinion and to maintain better social relationships.

This paper summarized the results of our application, whose objective is to recommend a Twitter user people he can follow according to his interest. After having fetched the areas of interest of different Twitter users from their tweets, we store their details in a database, thereby making suggestions to every user regarding people he can follow according to his interest, thus clustering together the people with similar interests.

References

1. Alec, G., Lei, H., Bhayani, R.: Twitter sentiment analysis. Final Projects from CS224N for Spring 2008/2009 at The Stanford Natural Language Processing Group (2009)
2. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.P.: Measuring user influence in twitter: the million follower fallacy. In: International AAAI Conference on Weblogs and Social Media (2010)
3. Company|About (Twitter). Retrieved from https://about.twitter.com/company 26 May 2017
4. Guy, I., Avraham, U., Carmel, D., Ur, S., Jacovi, M., Ronen, I.: Mining expertise and interests from social media. In: International World Wide Web Conference, Rio de Janiero, Brazil (2013)
5. Hennessy, A.: Sentiment Analysis of Twitter Using Knowledge Based and Machine Learning Techniques (2014)
6. Kim, J.Y., Collins-Thompson, K., Bennett, P.N., Dumais, S.T.: Characterizing web content, user interests, and search behavior. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, pp. 213–222. ACM, Seattle (2012)
7. Kiruthika, M., Woonna, S., Giri, P.: Sentiment analysis of twitter data. Int. J. Innov. Eng. Technol. **6**(4), 264–273 (2016)
8. Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., Ma, W.-Y.: Mining user similarity based on routine activities. In: Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information System, p. 34. ACM, Irvine (2008)
9. Matthew, M., Macskassy, S.A.: Discovering users' topics of interest. In: Proceedings of Workshop of Analytics of Noisy and Unstructured Text Data(AND). ACM, Toronto, Ontario, Canada (2010)
10. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: LREC (2010)
11. Qiu, F., Cho, J.: Automatic identification of user interest. In: Proceedings of the 15th International Conference on World Wide Web, pp. 727–736. ACM, Edinburg (2006)
12. Read, J.: Using emoticons to reduce dependency in machine learning techniques for sentiment classification. The Association for Computer Linguistics (2005)
13. Seefeld, K., Linder, E.: Statistical Using R with Biological Examples. University of New Hampshire, Duhram (2007)
14. Tarlekar, A.K.P.K.: Sentiment analysis of twitter data from political domain using machine learning techniques. Int. J. Innov. Res. Comput. Commun. Eng. **3**(6), 5590–5597 (2015)
15. The Porter Stemming Algorithm. Retrieved 26 May 2017, from Tartarus: https://tartarus.org/martin/PorterStemmer/ Jan 2006
16. Wang, T., Liu, H., He, J., Du, X.: Mining user interests from information sharing behaviors in social media. In: Advances in Knowledge Discovery and Data Mining. Springer, Berlin, pp. 85–98 (2013)
17. Wen, Z., Lin, C.-Y.: On the quality of inferring interests from social neighbors. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, Washington, DC, pp. 373–382 (2010)
18. Xie, P., Pei, Y., Xing, Y.X.: Mining User Interests from Personal Photos. Association for the Advancement of Artificial, Pittsburgh (2015)
19. Zhao, Y.: Text Mining with R—Twitter Data Analysis, Melbourne (2014)