Business recommendation based on collaborative filtering and feature engineering – aproposed approach

Prakash P. Rokade, Aruna Kumari D.
Department of Computer Science &Engineering, KLEF Deemed University Vaddeswaram, India

Article Info

Article history:

ReceivedJan 9, 2018
Revised Nov 28, 2018
Accepted Mar4, 2019

Keywords:

Collaborative filtering
Feature extraction
Machine learning
Sentiment analysis

ABSTRACT

Business decisions for any service or product depend on sentiments by people. We get these sentiments or rating on social websites like twitter, kaggle. The mood of people towards any event, service and product are expressed in these sentiments or rating. The text of sentiment contains different linguistic features of sentence. A sentiment sentence also contains other features which are playing a vital role in deciding the polarity of sentiments. If features selection is proper one can extract better sentiments for decision making. A directed preprocessing will feed filtered input to any machine learning approach. Feature based collaborative filtering can be used for better sentiment analysis. Better use of parts of speech (POS) followed by guided preprocessing and evaluation will minimize error for sentiment polarity and hence the better recommendation to the user for business analytics can be attained.

Corresponding Author:

Prakash P. Rokade,
Department of Computer Science & Engineering,
KLEF Deemed UniversityVaddeswaram,
Nalanda,Garimanagari,Near Mahila Mahavidyalaya,
Kopargaon, Dist.-Ahmednagar, Maharashtra, India.
Email: prakashrokade2005@gmail.com

1. INTRODUCTION

Sentiment analysis (SA) of blogs is playing a vital role for business decisions to plan a good business strategy. SA is an artificial intelligence strategy that quantifies the sentiment as positive, negative or neutral [1]. Sentiments are expressed at Document-level, Sentence-level, and Aspect-level [2]. SA has many applications in various fields like ranking products, services and merchants, predicting share price, predicting movie popularity, recommendation using business intelligence. SA aims to provide the right knowledge to the right person at the right time [3].

Current algorithms are being used for a single group of users or products, which ignore the impact for the other groups [4].There, may be few fake posts which are posted by fake users, competitors. So it is a challenge to filter the posts which are not specific to the feature of a product or service. Traditional SA algorithms do not consider the fact that as time passes, the value of data decreases for making decisions. The data considered for short tenure will decrease the quality of recommendation or decision. Bugs in bugs out problem still remain there [5].

Clustering followed by collaborative filtering has proposed a remarkable solution to resolve these issues [5].In the first step, we preprocess the input sentiments and identify the features of the product or service described in sentiments. Using clustering, likely blogs are selected and then feed to collaborative filtering algorithm to fill missing gaps of rating for some features [6].One objective of the proposed recommendation system is to enhance traditional content-based filtering by building user profile based on the static information that represent the likeliness of users to the features of the items or service [7].

2. LITERATURE SURVEY

A remarkable work is carried out in the research area of sentiment classification. The main focus of this work is on classifying larger pieces of text, like reviews of product or event [8]. Tweets are different from reviews as they have different purpose. Reviews are summary of author's thoughts. Tweets are limited to 140 characters of text.Tweets represent general mood of people through various reactions based on experience or as an impression for news articles [9]. Hu and Liu have given a technique for Feature Based Summarization system (FBS) of customer reviews of products. It also generates sentiment based summary as either positive or negative opinion using adjective words in reviews [10]. Chaovalit and Zhou compared supervised and unsupervised algorithm for classification and got 83.54% of accuracy for supervised method and 77% of accuracy for unsupervised method [11]. Pang O Keefe and Koprinska have given technique to select features using attribute weights and applied Navie Bayes and SVM classifiers for classification of moods [12, 13]. Linguistic features are used to detect the twitter sentiment using hash tagged data set (HASH) and emoticon data set. Results are evaluated by using unigrams and bigrams [14, 15].

The study by Hassan shows that parts of speech features are not playing good role in sentiment analysis for micro-blogging domain. Author introduces classification method for query term sentiment analysis. Here classifier and feature extractor are considered as two different components [16]. Each token is assigned a sentiment score called total sentiment index. Using classification algorithm the sentiments are classified as positive or negative polarity sentiments [17]. Political future can be analyzed real time monitoring and analyzing public conversation on social sites [18]. Feature vectors and tagged content of corpus can be used to make model by using machine learning approach. This model is used to classify or categories untagged corpus of text document [19]. For language consistency twitter is more informal. Emoticons are used express the opinion. Many tweets are ambiguous and these are maximizing the opinion for readers; but deflect the opinion to a machine learning algorithm [20]. Sentiment classification algorithm (SCA) and SVM are used to evaluate the performance of the approach used accuracy, recall, precision are some parameters on which sentiment analysis performance is evaluated [21].

3. PROPOSED APPROACH

3.1. Mathematical model

LetS bethemodel which describesthe extraction,preprocessing,lebling and evaluating the sentiments.

S= {Tw, Pt, Sl, Se}

where

Tw = Twitter sentiments.
Pt =PreprocessingofTweets
Sl =Labling the sentiments as positive, negative or neutral
Sl = {Pv, Nv, Ne}

- Pv= {P1, P2,…, Pn}=Positive Class
- Nv= {N1,N2,…,Nn}=Negative Class
- Ne= {Ne1,Ne2,…,Nen}=Neutral

Se =Sentiment evaluation

3.2. Research design

A proposed research design for sentiment analysis using collaborative filtering and feature engineering is given in Figure 1.

3.2.1. Data collection

A correct input may leads us to get a correct output. Sentiment data is available on twitter website or from kaggle dataset.

3.2.2. Data preprocessing

a. Case normalization

The tweets are available in combined case that is it may contain upper and lower case characters. In case normalization the entire document or sentence is converted in to lower case pattern generally.

b. Tokenization

A document is split in to sentences. Sentences may be divided in to words. By removing certain characters like punctuation marks, remaining words are now tokens.

c. Stop ward removal

A set of stop words list is provided to remove them from sentiments. The frequently used stop words are ‘a’,’an’,’the’,’shall’,’will’,’that’,’am’,’is’,’are’,etc..

d. Root stemming

In this process derived words are reduced to their stem. For example ‘careful’, ‘careless’, ‘carefully’ are reduced to ‘care’.

e. Transforming the words

A set of defined rules are used to transform the word to a specific form. For example a word clarifies can be replaced by clarify. The Table 1 describes how the words with suffixes are converted to equivalent stem after removal of suffixes.The words with suffixes in clumn 1 are converted to equivalent srtem in column 2.

Table 1. Word with their equivalent stem

Word with their equivalent stem	
Words	Stem
Equality, Equally	Equal
Engineering,Engineer,Engineered	Engineer
Manually,Manual,Man	Man

f. Removal of handles like # etc.

Users include Twitter usernames in their tweets in order to direct their messages. A de facto standard is to include the@ symbol before the username (e.g.@alecmgo). An equivalence class token (USERNAME) replaces all words that start with @ symbol.

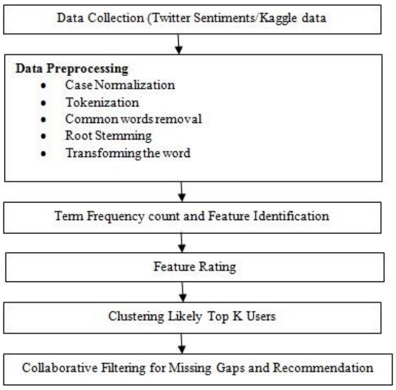
Data Collection (Twitter Sentiments/Kaggle data
Data Preprocessing
Case Normalization
Tokenization
Common words removal
Root Stemming
Transforming the word
Term Frequency count and Feature Identification
Feature Rating
Clustering Likely Top K Users
Collaborative Filtering for Missing Gaps and Recommendation

Figure 1. Flow of proposed sentiment analysis approach

3.2.3. Term frequency count and feature extraction

After doing preprocessing a list of adjectives in the dictionary is matched with every reaming word in the data set to find out adjectives and thus the features, along with these adjectives.

3.2.4. Feature rating

We will provide a list of adjectives along with a crisp value say 0 to 5 saying that 0 stand for the worst, 5 stands for the best and so on.Thus we can provide the rating for the features if the user has commented on.The uncommented feature will not have any rating, rather it will be empty rating as shown in Table 2.

Table 2. Adjective list with rating

Sr. No.	Rating(Crisp Value)	Proposed adjective list
1	0	worst,very very bad
2	1	bad,not good
3	2	Ok
4	3	Good
5	4	very good
6	5	best,excellent,marvaolous,fabulous

3.2.5. Clutering the top k users

We need to find similar users based on their interest for the features of product or service. Here we are interested to get top k users having the similar taste for their impressions.We can provide threshold value to optimize the result. While clustering using an appropriate clustering algorithm, say k nearest neighbour.

In the Table 3 shown user 1, 3, 4 are having similar taste of interarest for features. Likewise out of P users top k users we are finding. These top k users are now the representatives of the original data set we have considered as an input. The top k users have not rated for all features. But these top k users have commented on similar features very closely. The missing gaps of rating for some features by these k users will be overcome in collaborative filtering.

Table 3. User rating for different features

User	Feature				
	F1	F2	F3	F4	F5
1	5	4	4		3
2	3	1	2	5	3
3	4	4	4		3
4	5	3	5		4
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
P (Finite No.)	3	2	2	5	5

3.2.6. Collaborative filtering for recommendation

Collaboration means recommendation of item or service based on feature rated in user's choice. Filtering is separation of similar entities based on user's likes or dislikes. The motivation for collaborative filtering comes from the idea that one person can get best recommendation for any business say B, from another person who has the same interest in B already. Collaborative filtering methods are used for monitoring data such as financial data, sentiment blogs for product or services, an electronic commerce and web applications. Table 4 shown explains working of collaborative filtering. Consider movie rating is given for 5 features f1 to f5. Rating for features are in the form of 1 to5.1 stands for dislike and 5 stands for most like.

Table 4. Customer rating for features of movie

Customer	Feature				
	F1	F2	F3	F4	F5
1	5	3	4	4	?
2	3	1	2	3	3
3	4	3	4	3	5
4	3	3	1	5	4
5	1	5	5	2	1

Step 1**:** Ignore the missing reading column and calculate the average of remaining rows.
Average of row 1=(5+3+4+4)/4=4
Average of row 2=(3+1+2+3)/4=2.25
Average of row 3=(4+3+4+3)/4=3.5
Average of row 4=(3+3+1+5)/4=3
Average of row 5= (1+5+5+2)/4=3.25

Step 2: Choose 2 rows whose similarity is to be calculated using given formula.

$$Sim(Ci,Cj)=[\Sigma(rip-riavg)*(rj-rjavg)^2\,]/[\sqrt{\Sigma(rip-riavg)^2}\sqrt{\Sigma(rjp-rjavg)^2}]$$

where,
Sim (Ci, Cj) =Similarity between customer i and j.
rip=Particular rating of customer i.
rjp= Particular rating of customer j
riavg=Average rating of customer i
rjavg=Average rating of customer j.

By putting the values in above table into formula, we will get
Sim (C1, C2) =0.85
Sim (C1, C3) =0.7
Sim (C1, C4) =0
Sim (C1, C5) =0.79

Above results clearly state that customer 1 and customer 2 has highest similarity in their ratings. We may conclude that, rating for feature 5 for customer 1 will be same as given by customer 2. So, it will be 3 for customer 1.

Step3: In this step we can find out column average for all customers for all features. The Table 5 exaplains the column average for different features.As the colun average is between 1 to 5, we can set threshold as per our demand to comment on the quality of a feature for any product or service.

Table 5. Column average for features

Customer	Feature				
	F1	F2	F3	F4	F5
1	5	3	4	4	3
2	3	1	2	3	3
3	4	3	4	3	5
4	3	3	1	5	4
5	1	5	5	2	1
Column Average	3.2	3	3.2	3.4	3.2

Now one can use above statistics with some threshold for every feature for feature based recommendation of the movie.

4. CONCLUSION

We have thoroughly studied the proposed approach using collaborative filtering and feature engineering for business recommendation. The preprocessing on input data set will definately improves the quality of the corpus. We will get a proper set of features using frequently occurred adjectives. Clustering algorithm like k nearest neighbour will provide us top k similar users which can give the recommendation for any product of service using collaborative filtering.We can provide threshold value for individual feature so that product or service can be recommended based on that specific feature only. For the proposed approach in this paper, we will provide threshold value to all features considering as a system, which will give us the recommendation for any product or service.

In the future, one can directly consider the Kaggle data set, which provides the rating of any product or service by m number of users for f number of features. It will reduce the role of preprocessing and we can compare the machine learning techniques for better outcomes.

REFERENCES

[1] P. S. Priya and T. V.S. Rao, “Analysing Event-Related Sentiments on Social Media with Neural Networks,” *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol/issue: 7(3), pp. 119-124, 2018.

[2] M. A. Fauzi, et al., “Improving Sentiment Analysis of Short Informal Indonesian Product Reviews using Synonym Based Feature Expansion,”*TELKOMNIKA Telecommunication Computing Electronics and Control*, vol/issue: 16(3), pp.1345-1350, 2018.

[3] Z. Z. Gao, et al., “Time-Weighted Uncertain Nearest Neighbor Collaborative Filtering Algorithm,” *TELKOMNIKA Indonesian Journal of Electrical Engineering*, vol/issue: 12(8), pp. 6393-6402, 2014.

[4] M. W. Chughtai, et al., “Goal-based Hybrid Filtering for User-to-user Personalized Recommendation,”*International Journal of Electrical and Computer Engineering (IJECE)*, vol/issue: 3(3), pp. 329-336, 2013.

[5] P. Arora, et al., “An Approach for Big Data to Evolve the Auspicious Information from Cross-Domains,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol/issue: 7(2), pp. 967-974, 2017.

[6] M. R. Ma’arif and A. Mulyanto, “Improving Recommender System Based on Item’s Structural Information in Affinity Network,” *Proceeding of International Conference on Electrical Engineering, Computer Science and Informatics (EECSI 2014), Yogyakarta, Indonesia*, 2014.

[7] A. El-Korany and S. M. Khatab, “Ontology-based Social Recommender System,” *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol/issue: 1(3), pp. 127-138, 2012.

[8] B. Pang, et al., “Thumbs up? Sentiment classifcation using machine learning techniques,”*Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 79, 2002.

[9] P. Turney, “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews,”*Proceedings of the Association for Computational Linguistics*, 2002.

[10] M. Hu and B. Liu, “Mining and Summarizing Customer Reviews,” *Proceedings of the 10th ACM SIGKDD, International Conference on Knowledge Discovery and Data Mining*, 2004.

[11] P. Chaovalit and L. Zhou, “Movie Review Mining: A Comparison between Supervised andUnsupervised Classification Approaches,”*System Sciences, HICSS'05, Proceedings of the 38th Annual Hawaii International Conference on IEEE*, pp. 112c- 112c, 2005.

[12] T. O‟Keefe and I. Koprinska, “Feature Selection and Weighting in Sentiment Analysis,”*Proceedings of the 14th Australasian Document Computing Symposium, Sydney, Australia*, 2009.

[13] A. Pak and P. Paroubek, “Twitter as a corpus for sentiment analysis and opinion mining,” *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, pp. 1320-1326, 2010.

[14] E. Koulompis, et al., “Twitter Sentiment Analysis: The Good the Bad and the OMG!,”*Proceeding of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[15] F. M. F. Wong, et al., “Why Watching Movie Tweets Won’t Tell the Whole Story?,” Arxiv preprint arXiv:1203.4642, pp. 6, 2012.

[16] H. Saif, et al., “Semantic Sentiment Analysis of Twitter,”*Proceedings of the 11th International Semantic Web Conference*, 2012.

[17] Gann W. J. K., et al., “Twitter analytics for insider trading fraud detection system,”*Presented at second ASE international conference on Big Data*, 2014.

[18] Jensen M. J., et al., “Introduction,” E.Anduiza, M. Jensen, & L. Jorba (Eds.), “Digital media and political engagement worldwide: A comparative study,” New York, NY, Cambridge University Press, pp. 1-15, 2012.

[19] A. A. Kothari andW. D. Patel, “A Novel Approach towards Context Based Recommendations Using Support Vector Machine Methodology,” *Procedia Computer Science*, vol. 57, pp. 1171-1178, 2015.

[20] A. Tripathy, et al., “Classification of Sentimental Reviews Using Machine Learning Techniques,” *3rd International Conference on Recent Trends in Computing*, 2015.

[21] V. Sahayak, et al., “Sentiment Analysis on Twitter Data,”*International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, vol/issue: 2(1), 2015.

BIOGRAPHIES OF AUTHORS

Prakash P.Rokade has received hisB.E.degree in Computer from Pune University, Maharashtra; India in 2005.He has received his M.Tech. degree in Computer Engineering from Bharti Vidyapeerth, Pune, Maharashtra, India in 2011 and presently pursuing his Ph.D. in Computer Science andEngineering from Koneru Lakshmaiah Education Foundation, formerly K L University, Vaddeswaram , Andhra Pradesh, India.His research interest includes Sentiment Analysis, Opinion Mining, and Machine Learning.

Aruna Kumari D has received her Ph.D. degree in Computer Science and Engineering from the K L University, Vaddeswaram, Andhra Pradesh, India. Currently, She is Professor Koneru Lakshmaiah Education Foundation, formerly K L University. Her teaching and research areas includes Data Mining, Machine Learning and has published more than 50 papers in many National, International journals.She is honoured by DST Young Scientist Award (Govt. of India).