International Journal of Advanced Science and Technology
Vol. 29, No. 7s, (2020), pp. 1375-1386

ANALYSIS OF DIGITAL MARKETING DATASET USING MACHINE LEARNING ALGORITHM

Dr.T.Avudaiappan K. Abirami, M. Dharani Lakshmi,R.B.Karunyaa, ,
[1,2,3 & 4] Department of Computer Science & Engineering, K. Ramakrishnan College Of Technology, Trichy
Email: avudaiappanmecse@gmail.com

Abstract

Recent years have seen opinion analyzes on Twitter becoming a common trend for science. Most existing Twitter sentiment analysis solutions essential to understand only organized Twitter message information. The performance of tweets or not good at all time because some tweets or not having one obvious meaning. Research indicate that the emotion transmission patterns in Twitter have connections to the polarities of emotional Twitter posts.This paper focus on how to analysis of digital marketing data set using machine learning algorithm. The diffusion of sentiments by studying a fact or situation that is observed called reversal sentiment. Then the look at the interrelationships between twitter's textual awareness and patterns of feeling diffusion and suggest an iterative algorithm to conclude the polarities in twitter posts. This study is to help and improve an interpretation of twitter's feelings.We suggested learning algorithms for machines running SVM and Random Forest. We compare traditional algorithms such as the Nave bayes machine learning algorithm to make successful layout. Measurement of evaluation was taken using accuracy, precision and F1.The Social Framework was developed for business purposes and checked with end-users for effective implementation. Machine learning algorithm has found the best algorithm, and work is done to block inappropriate comments in twitter

Keywords: *Tweets, Sentiment Analysis, sentiment diffusion.*

Objectives

Using machine learning algorithms we try to find the best machine learning model for classifying twitter data.

1. Introduction

In recent years, the novel appeal and growing popularity of social media has generated an accumulation of vast quantities of data for the people to contribute their outlook, prospect and perspective. This mountain of data emerging is widely called Big Data. The analysis of tweets in this area gives a enormous potential techniques of data mining research in order to achieve a more reliable detection of hidden knowledge in big data.

Twitter is a social networking worldwide, has influenced and changed the way individuals or organizations get the information they're interested in.

Users use mobile phone are computer to send and to read message to tell their followers what they think, what they are doing around them. Reply and repost can also be dons with other specific users. User can also interact with other users. The feeling polarities of users conveyed in twitter messages have become a recent trends in researchbecause of its wide. Ranging applications for example, the analysis of twitter user polarities on political partierand candidates, a range of approaches have been developed to offer policy election strategies. Business organizations is a another best examples to track people feelings about their brands and products.

The main aim of analyzing twitter data is to classify the people opinion in to positive, negative and neutral twitter sentiment exploit the traditional method of analyzing text feeling directly. nevertheless, twitter message are open brief and vague, separating from other types of text such as news reports and book post. Furthermore, due to their casual form, twitter message involve more repeatation, word hippo, and modal practices.

ISSN: 2005-4238 IJAST
Copyright ⓒ 2020 SERSC

As a result, the predict of Twitter messages' feeling polarities, the performance of traditional text sentiment analysis drops drastically.Many novel methods for analysis of Tweets have been developed to solve this problem. This is divided into two types: fully supervised methods, and methods that are far supervised.
The completely supervised methods strive to learn classifiers of emotion the problem that finds in fully supervised methods is that manually building sentiment lexicons and labeled the data is time-consuming and labor-intensive, and therefore the feeling lexicons and labeled data are guarantee less performance. Additionally, fully managed solutions are usually focused on hand-crafted applications, and it remains a challenging challenge to develop effective applications. Remotely supervised methods can avoid labor-intensive manual annotation, due to the noise in the labels, their performance is not satisfy. The preprocessing technique avoids the problem of noisy labels for an interpretation of the emotions. The researchers conformed not all dataset and algorithm have efficient method of preprocessing the dataset of one algorithm differ from algorithm of other dataset results in decrease in performance.
Essentially, it's the mechanism in our project of deciding if a piece of writing is positive or negative. It is also called Material Polarity. Companies may be aware of how their customers feel about their company by reviewing twitter messages, user reviews and consumer feedback. They can also track particular subjects and get valuable insights into how people speak about them. So when a person uses offensive words when tweeting the same person who posted them will be blocked and will not be able to tweet again.
The best method of machine learning algorithm has been found and the best method has been found from the output of the machine learning algorithm, and the output is evaluated after testing withcomments in the database if the posted tweet is positive, negative or neutral.

2. EXISTING SYSTEM

Accuracy is less then 71%.Lack of decision making when non English word came in toas input attribute. The good decision will not be taken about the investment if our sentence score is lowThe large data set should not be handled some unexpected results occurs. More depends on datasetApriori algorithm fails to handle large datasets and as a result can generate faulty resultsAny person can tweet which cannot be blocked in twitter so there are chances of misusing or posting unwanted words.

3. PROBLEM DEFINITION

The micro-blogging sensitivity analysis is a new exploration topic, so there is a lot of resources available in this field. In the terms of reviews of user, documents, web blogs / articles and general sentence level of sentimental analysis.These differences from twitter (or) social media application due to the maximum of 150 characters per tweet. So it is the compulsion to the user to convey their opinion in very short form of text supervised learning technique in the machine learning algorithm uses the support vector machine and varve bay gives the best output to identify feelings, but the manual design used for supervised learning approach is high-priced. Some excretion on semi-supervised approaches has been done and there is more ways for improvement.So many research scientist who test the advanced features and classification methods also equate their results with the output of the base line.To choose the finest features and the most efficient classification techniques for specified applications, there is a need for actual and formal differentiation between these output arrived through different features and classification techniques.

4. PROPOSED WORK

We proposed SVM and Random forest machine learning algorithm. For evaluation purpose we implement Naïve bayes.Each of them is classier machine learning algorithms, we are going to implement using Scikit learn machine learn library. In our proposed work From the Output of Random forest method twitter sentiment analysis is done by blocking abusive words when user posts tweets in twitter which is helpful to analyse the information in the tweets where opinions

Copyright ⓒ 2020 SERSC

are either positive or negative, or neutral where it is necessary to block unnecessary comments providing security to users In our proposed work from the output of Random forest model we are converting into application projects and checked with end user and twitter sentiment analysis is done by blocking abusive tweets when user posts tweets in twitter by checking with comments in the database which is helpful to analyse the information in the tweets where opinions are either positive or negative, or neutral where it is necessary to block unnecessary comments providing security to users .

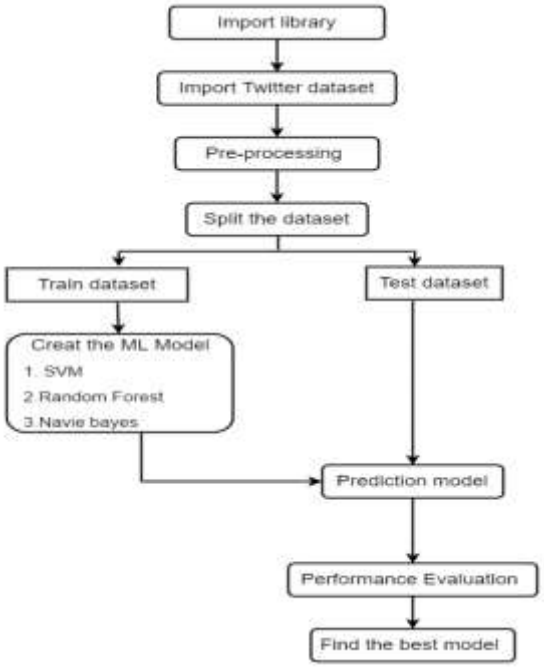


Fig: 4.1 Machine learning Architecture

4.1 Machine learning Architecture

4.2 Application Architecture

Copyright ⓒ 2020 SERSC

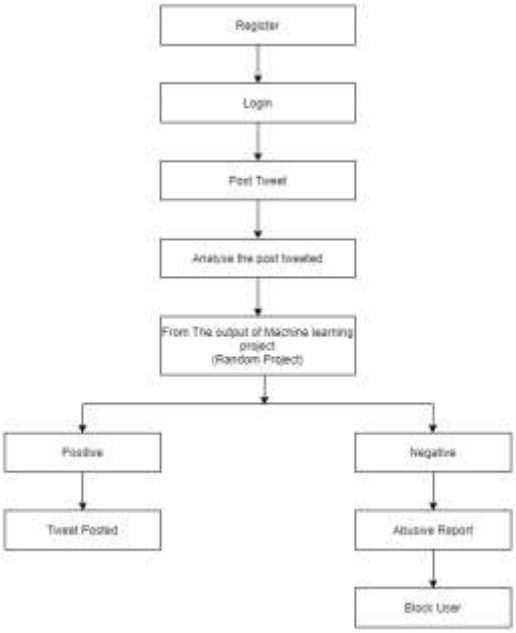


Fig: 4.2 Application Architecture

Why do we choose different Algorithms?

• Because of that we do not know which one provides better results.
• We ensure that different algorithms are tested to develop the algorithm
• We can apply the same train and check data set functions for different algorithms.
• Finally, in comparing other machine learning algorithms, we would conclude pest algorithms.

4.3PROPOSED WORK

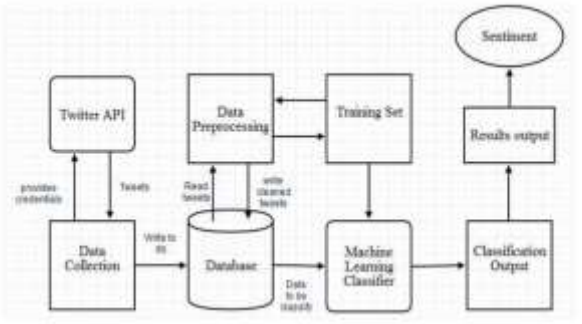


Fig: 4.3 proposed work

5. PROPOSED METHODOLOGY

The project has been separated into eight modules:

- Importing Library and Dataset
- Pre-processing

ISSN: 2005-4238 IJAST
Copyright ⓒ 2020 SERSC

- Data extraction and selection.
- Machine Learning model
- Prediction and Evaluation Model.
- Checking abusive comments

A. Member Module

B. Article Censure Module

C. Administrator Module:

- Posting tweet
- Blocking users

5.1 Importing Library and Dataset

We have to import appropriate library function, they are the listed below

- os
- numpy
- pandas
- seaborn
- matplotlib.pyplot
- splitting test and train dataset
- SVM
- Randomforest
- Naïve bayes

5.2Pre-processing

Tweets retrieved from twitter are a mixture of URL and other non-sentimental data such as "#" hashtags, "@" posts, and "RT" retweets; text information must first be tokenized in order to receive n-features. With standard tokenizers built with standardized and normal text, tweets pose a challenge. The following figure illustrates the various intermediate-processing function phasesIntermediate steps are the set of functions which will be considered by the classifier. In short, any feature that has been introduced is something we think about.

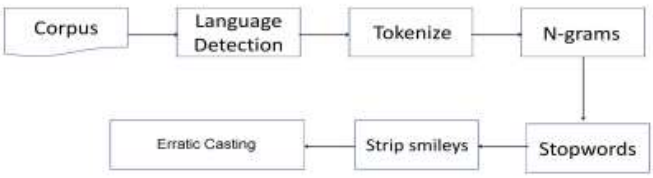


Fig: 5.2 pre-processing

Language detection-Since we are mainly concerned only with English text. All tweets were separated into the data in English and non-English. This is possible with the language detection function of NLTK.Tokenize-To say "The weather is bright and beautiful today" Tokenizers divide strings into substring lists also known as token for a sample input text. Tokenizing the text makes it easy to isolate all needless symbols and punctuations, and filters out only those words that can add meaning

Copyright ⓒ 2020 SERSC

to the emotional polarity score of the text.Creating n-grams- We must create a collection of n-grams from consecutive terms. For example, a phrase ' I don't like fish ' will form two bigrams: ' I do + not, ' ' do + not, ' ' like, ' ' not fish. '. Such a procedure enables the classification to be improved since negation plays a special role in an expression of opinion and sentiment.Stop words- In information retrieval, ignoring very common words like "a," "an," "the" etc. is a common tactic.As their posting does not provide any useful information when classifying a text.Since the query term itself should not be used to assess the post's feeling about it, each query term is replaced by a QUERY keyword. While this makes it somewhat of a stop word, it can still be useful when not using a bag-of-words model and it becomes important to locate the query in relation to other words.

5.3 Feature extraction and selection

The extraction of features is a quite complex concept regarding the translation of raw data into the inputs required by a specific Machine Learning algorithm. Features must represent the data information in a format that best suits the algorithm needs that will be used to solve the problem.

Extraction of features fulfills this requirement: it extracts useful information from the raw data–the apps–by reformatting, combining and converting primary apps into new features until it is complete generates a new set of data that the Machine Learning models can consume to achieve their objectives.

For its part, the selection of features is a clearer task: select some of them, given a set of potential features and discard the rest. The selection of features is either applied to prevent redundancy and/or irrelevance in the features or to obtain a limited number of features to avoid overfitting.

5.4 Machine learning model.

We proposed SVM and Random forest machine learning algorithm. For evaluation purpose we implement Naïve bayes.Each of them is classier machine learning algorithms, we are going to implement using Scikit learn machine learn library

5.5 Prediction and Evaluation Model

Prediction

Prediction' refers to an algorithm's success after it has been educated on a historical dataset and applied to new data when predicting the probability of a particular result, such as whether the picture has a disease or not.Test the model We will assess our model using Specific and Rememberto form that prediction function.

Precision is a fraction of people who actually suffer from pneumonia to all those predicted to have pneumonia by the model.

After implementing hybrid approach that are combining Naïve bayes, Random forest and SVM algorithm, we are going to evaluate each algorithm using evaluation measure. The performance of evaluation measure are Accuracy, specificity and sensitivity. These value taken based on confusion matrix values like True positive, True negative, False positive and False Negative.

5.6 Checking the abusive Comments

We regard the Random Forest Method as the best research process. From the performance of the Random Forest Method we translate this into a proposed system application able 5.6. total abusive comments in tweet

Copyright ⓒ 2020 SERSC

5.6.1 Member Module:

New users are required to register the information in this module. Until this the user is required to access their details by logging in. Users will get into their account by logging in. The login module allows users to sign in with a User Name and Password. You can this module on any Module Tab to allow users to log in to the program. When the administrator has authorized users to create accounts a connection to the Create Account appears in the login tab. The user won't be able to access their account until registering on this platform.

5.6.2 Article Censure Module:

In this module, User can submit any posts on this site to others. When posting, words in the post are compared to the censured words created by admin. Here, any censured words are comes on that post, the user get a warning message for that post and the post should not published to anyone on this site.

5.6.3 Administrator Module:

This module allows the admin to have the overall control of the website. Admin can allow the group permission, authorization, enable articles, send mail to groups, etc. Admin can also configure the design and control the templates via the template module. The pages can be published whenever the admin make ready the page to be published on the site. Admin can see all the feedback sent by the users.

5.7 POSTING THE COMMENTS:

After checking the abusive comments with the database values. If the posted comment is positive, negative or neutral is checked out. If the comment is positive the tweet is posted. Once the tweet is negative the comment is blocked from posting. Hence avoiding unwanted users from posting unwanted comments.

5.8 BLOCKING USERS:

After checking with the database, abusive comments are blocked and those comments will not be posted. It allows the users of twitter to secure the posted tweets.

6. EXPERIMENTAL RESULTS

DATA SET	**KAGGLE**	**DATA1**	**DATA2**
No of tweets language		**231+58=289**	**4090+1023=5113**
No of Negative		**132**	**4022**

Copyright ⓒ 2020 SERSC

words			
No of Positive words		**132**	**519**

Table: 6.1total number of data set

Table: 6.2 Detail of dataset

NO OF DATASET	NO OF TRAIN DATASET	NO OF TEST DATABASE
Samsung	**231**	**58**
Amazon	**4090**	**1023**

6.2 TWITTER ANALYSIS

Result:

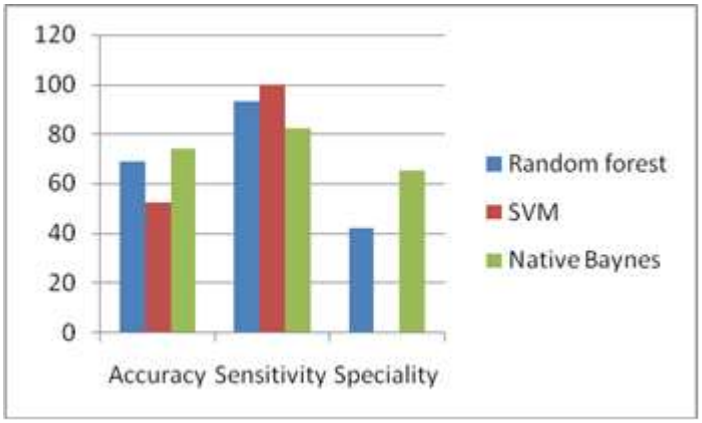


Fig:6.1: Comparison between SVM, Random forest, Native Baynes

Algorithm	Accuracy	Sensitivity	Speciality
SVM	30	100	0.0
Random forest	59	88	33
Native	40	100	14

ISSN: 2005-4238 IJAST
Copyright ⓒ 2020 SERSC

Baynes			

Table 6.3: comparison of accuracy

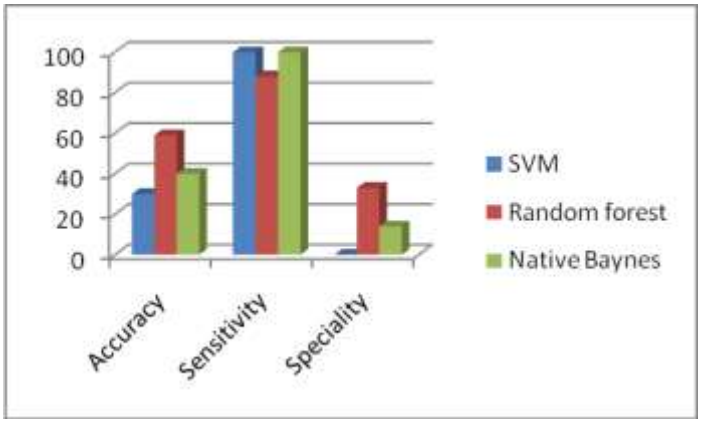
Algorithm	**Accuracy**	**Sensitivity**	**Speciality**
Random forest	**69**	**93**	**42**
SVM	**52**	**100**	**0.0**
Native Baynes	**74**	**82**	**65**

6.3 AMAZON ANALYSIS

Table 6.4: comparison of accuracy

Algorithm	**Accuracy**	**Sensitivity**	**Speciality**
SVM	30	100	0.0
Random forest	59	88	33
Native Baynes	40	100	14

Result:



Copyright ⓒ 2020 SERSC

Fig:6.2: Comparison between SVM, Random forest, Native Baynes

7. LITERATURE REVIEW

7.1 Twitter Sentiment Analysis, 3-Way Classification: Positive, Negative or Neutral? In the year of 2018

Merits:Best method in terms of overall accuracy ratio is MultiClassClassifier (0.711). Close overall accuracy ratios comes from Random Forest(0.707), SVM (0.706)

Proposed Algorithm: Support Vector Machine (SVM)

2Naive Bayes Classification J48 Decision TreeRandom Forest MultiClassClassifier, IterativeClassifierOptimizer
7.2 SENTIMENT ANALYSIS ON TWITTER USING STREAMING APIon 2017 IEEE 7th International Advance Computing Conference

Merits: Real time deployment.

Proposed Algorithm: uni-word naïve bayes' classification

7.3Using Sentimental Analysis in Prediction of Stock Market Investment on 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions).
Merits: The sentimental score with market values were provided to an artificial neural network to predict the future market value
Proposed Algorithm: An Artificial neural network is a computational model which is based on the structure and functions like a biological neural network.

7.4 Sentimental Analysis Using Fuzzy and Naive Bayes By

Ruchi Mehra1 , Mandeep Kaur Bedi2 , Gagandeep Singh3 , Raman Arora4 ,Tannu Bala5 , Sunny Saxena on IEEE 2017 International Conference on Computing Methodologies and Communication.
Merits:They have confirmed that proposed algorithm offer better performance when conducting the classification
process supporting results.

Proposed Algorithm: Fuzzy and Naive Bayes

7.5Sentiment Analysis of Tweets using Machine Learning Approach by Megha Rathi, Aditya Malik, Daksh Varshney, Rachita Sharma, Sarthak Mendiratta on 2018 Eleventh International Conference on Contemporary Computing (IC3).

Merits:The comparative results prove that hybrid model improved the overall classification accuracy and f-measure of sentiment prediction as compared to traditional existing techniques for classification.

Proposed Algorithm:SVM, ADABOOSTED DECISIONTREE

8 Expected outcome

We could get preprocessed data
We could create features using TF-IDF model

Copyright ⓒ 2020 SERSC

After implementing model we should crate prediction using test data
Using predicted model values to calculate performance metrices(Accuracy, sensitivity and specificity)
Based on performance metrics find the best model
Using the best model to predict the output.

9. CONCLUSION

Mining polarities of sentiments expressed in Twitter messages are important though demanding activities. Most of the current Twitter sentiment analysis solutions only find Twitter textual details and cannot attain adequate performance due to the distinctive features of message from Twitter. Although some recent analysis gives that sentiment diffusion patterns have very close relationships with Twitter message sentiment polarities, existing approaches basically focus only on Twitter message textual information, but ignore sensitivity diffusion information. Inspired by recent work on fusion of knowledge from multiple domains, we are taking a initial stepto integrate information on textual and sensitivity diffusion to attain better performance of Twitter sentiment analysis.
To this termination, initially we need to analyze the sentiment diffusion process on twitter by investigate a aspect called sentiment reversal based on trees and diffusion networks. Then we need to build a sentiment reversal prediction model, and we need to construct a novel twitter sentiment classification algorithm know as sentiDiff .
In this SentiDiff, the interrelationships between Twitter's textual information and feeling diffusion patterns are seeing, and a supervised learning system incorporates the textual information-based sentiment classifier and the model called sentimental reversal prediction model. Real-world data set experiments gives that our proposed SentiDiff algorithm with we assist with the state of the art textual information-based sentiment analysis algorithms.
It has been found from study method that the best approach is random forest. By using the output of this Random forest method analysis is performed by comparing with comments in the database. From this offensive comments twitter users can be blocked and covered.

Hence this Web application enhances to protect the restricted words that are post in social networking by providing a warning to the user while submitting their posts in the social networking

REFERENCES

1. D. Paul, F. Li, M. K. Teja, X. Yu, and R. Frost, “Compass:Spatio temporal sentiment analysis of us election what twitter says!” in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2017, pp. 1585–1594.
2. K.-L. Liu, W.-J. Li, and M. Guo, “Emoticon smoothed language models for twitter sentiment analysis,” in Proceedings of the Twenty- Sixth AAAI Conference on Artificial Intelligence, 2012.
3. J. Zhao, L. Dong, J. Wu, and K. Xu, “Moodlens: an emoticon-based sentiment analysis system for chinese tweets,” in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012, pp. 1528–1531.
4. A.Go,R.Bhayani,andL.Huang,“Twittersentimentclassification using distant supervision,” *CS224N Project Report, Stanford*, vol. 1, no. 12,2009.
5. D.-T. Vo and Y. Zhang, “Target-dependent twitter sentiment clas- sification with rich automatic features.” in *IJCAI*, 2015, pp. 1347– 1353.
6. E. Cambria, “Affective computing and sentiment analysis,” *IEEEIntelligent Systems*, vol. 31, no. 2, pp. 102–107,2016.
7. D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, “Learning sentiment-specific word embedding for twitter sentiment classifi- cation,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014, pp.1555–1565.
8. K. Schouten and F. Frasincar, “Survey on aspect-level sentiment analysis,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 3, pp. 813–830,2016.
9. J. Zhao and X. Gui, “Comparison research on text pre-processing methods on twitter sentiment analysis,” *IEEE Access*, vol. 5, pp. 2870–2879,2017.

Copyright ⓒ 2020 SERSC

10. N. Du, Y. Liang, M. Balcan, and L. Song, "Influence function learn- ing in information diffusion networks," in *International Conference on Machine Learning*, 2014, pp.2016–2024.
11. M.Tsytsarau,T.Palpanas,andM.Castellanos,"Dynamicsofnews events and social media reaction," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp.901–910.
12. S. Stieglitz and L. Dang-Xuan, "Emotions and information diffu- sion in social mediasentiment of microblogs and sharing behav- ior," *Journal of Management Information Systems*, vol. 29, no. 4, pp. 217–248,2013.

13. Y.Fu,Y.Ge,Y.Zheng,Z.Yao,Y.Liu,H.Xiong,andJ.Yuan,"Sparse real estate ranking with online user reviews and offline moving behaviors," in *2014 IEEE International Conference*

Copyright ⓒ 2020 SERSC