Location Based Sentiment Analysis of Products or Events over Social Media

Aarati Patil
Computer Science Department
JSPM College of Engineering
Pune, India
Email:aaratipatil.007@gmail.com

Srinivasa Narasimha Kini
Computer Science Department
JSPM College of Engineering
Pune, India
Email:snkini@gmail.com

***Abstract*—** Nowadays social media has become a very momentous and trendy communication medium amongst all online surfers, users and data scientists because of the recent advancements in it. It constituted the study of information diffusion, user communication and user control over social networks. All types of users share their opinions on various aspects of day to day activities every day. Therefore social media web-sites are rich sources of data for opinion mining. Such data can be efficiently used for sentiment analysis. This research aims to analyze location based social media data to compute the popularity of the products/events. And this is achieved by integrating sentiment analysis, location based data analysis and machine learning approach. An application has been developed which captures the real time communication over social media sites and implements sentiment analysis on collected data. This research work uses publicly available and location enabled social media data. Analysis results are used to optimize the decision making.

Keywords-*Social media; opinion Mining; sentiment analysis; machine Learning*

*****

I. INTRODUCTION

In the era of Internet and World Wide Web, netizens or social media users are gradually increasing because of very rich internet connection all over the world for many wikis, social sites like Facebook, twitter, and blogs. Social networks have evolved over the years to the modern-day and are still evolving. The recent advancements in the social networks have putted it at the heart of communication on the web. In today's digital world social media is gaining popularity every day and becoming essential communication medium. By using micro-blogging services, users post messages about their daily life and initiate discussions by sharing personal opinions and emotions on different topics. The topics of discussion on social media are limitless they can vary from something as simple as some products, events to more complex issues like economic issues, problems, interests, culture, politics, religions, diseases, epidemic, food crisis, and famine and so on. The richness of information available on social network provides unparalleled opportunities for data analytics in the context of social networks.

In daily life communication happen over social media sites, people initiate conversation and others open up with their opinions. In conversation some have positive aspects about topic or some have negative aspects or some people are neutral at their opinions. In this way opinions flow over social media and diffusion of opinion can be captured while communication. In our research we are interested in extracting social media messages location wise, preprocessing and Sentiment analysis on the extracted data and provisioning of smarter decisions by computing sentiment score through visualization.

This research is interested in online real time communication happening over social media. Location based social media data is extracted and sentiments of users are identified to obtain location wise popularity of any topic/product. Analysis results are provided so that one can use them for optimizing decision making. In this way Information Retrieval over social network and Smarter Decision Making is achieved through this research[1]. Design architecture and other details are explained in [1]. Subsequent sections will explain related work and how Computing of sentiment score and sentiment analysis is achieved by means of this research. Experimental results are explained in this paper.

II. RELATED WORK

In the survey [2], we got some insightful information that data scientists and social users are showing interest in sentiment analysis of social media data. Different approaches have been implemented to automatically detect sentiment on texts [3, 4]. An active research on Sentiment analysis on selective micro blogging sites has explored in [5, 6].

In [7] they have classified the subjectivity of social media messages based on traditional features with the inclusion of some social site specific clues such as retweets, hash tags, links, uppercase words, emoticons, and exclamation and question marks. Further a Part-Of-Speech (POS) specific prior polarity features and a tree kernel to obviate the need for tedious feature engineering is introduced in [8].

In [9] they have investigated linguistic features utility to detect the sentimental weight for social media messages. They tried to analyze automatic part-of-speech and sentiment lexicons to examine whether it is suitable for sentiment analysis. Through their experiments they found out that the use of POS might not be a good fit for sentiment analysis, especially for micro blogging domain. Also, their experiments concluded that the abbreviations, polarity and emoticons were

[1] This paper is extension of "IRSDM of SocioData: Location Based Analysis of Opinion Diffusion on Social Network", which is published in fifth post graduate conference of computer engineering, CPGCON 2016. System architecture and other details are briefly explained in above mentioned paper.

clearly useful for sentiment analysis of micro blogging services. For sentiment analysis a numbers of methods and lexical resources are available.

In [10] they have proposed a system to combine opinion strength, emotion and polarity indicator to generate improved sentimental analysis of polarity and subjectivity. Manual classification for opinion mining is hardly possible effort, although several methods have been proposed for automatic opinion mining. Supervised classification use polarity estimation and for unsupervised approach lexicon resources are used widely. They have considered two major classifications: Subjectivity and Polarity. The suitability checking for training-dataset is very important and should be measured beforehand. It should be checked whether the training examples are capable of capturing the sentiment diversity of specified domain.

[11] Described a real-time system for analyzing public sentiment about Election Cycle using Social media data. Social media site is considered as a central site for expressing public opinions where people express their political views about candidates and different parties [12]. Also in [13] it is observed that public mood has a significant and immediate effect on the social and political events. Public mood has various dimensions and it highly affects the cultural and economic events [13].

Due to huge, noisy and complex social media data, researchers have influenced visualization techniques to assist users with deep analysis. Much visualization has been implemented to highlight the patterns of information spreading on social media [14, 15].

III. PROBLEM DEFINITION

This research measures the popularity of any event/product by analyzing public statements location wise. This is possible by analyzing messages that contain general sentiment.

- In the first step different locations for collecting social media messages are identified. Targeting location can be identified by users' choice.
- In the second step big data source, which can deliver real-time or relatively new data is selected. This research work is using social media-API as a feeding data source.
- Once the locations are chosen and social media messages are captured from each location, this project will undertake detailed data analysis on. Each single message will be preprocessed and analyzed separately to extract all the useful information. As a result the process will generate more concise output (message string), which is easier and faster to parse
- Afterwards, the smaller filtered message will be parsed and stored into arrays of words to be compared with the sentiment dictionary to score as positive, negative or neutral.
- Once we obtain the sentiment score for all messages collected and stored separately location wise, analysis is made to define the popularity of event or product at different locations.

IV. OUR PROPOSED APPROACH AND SOLUTION ARCHITECTURE

We have opted twitter for implementation as twitter API deliver better insights with easier access to the information extraction. Twitter content is very concise and provides free APIs it is easy to mine using twitter4j and analyze the data.

I. Creating Twitter Application

Once the developer login to the twitter to get access to the java project first they are supposed to create an application. Configuration Builder is one of the classes available in twitter4j jar collection, which is responsible to collect the data from twitter by creating an interface between the application and twitter. Once the application is created successfully it returns access tokens to the developer. Parameters of Authentication required configuring Twitter4j: Consumer Key, Consumer Secret, Access Token and Access Token Secret. The secret tokens are used to collect tweets from the twitter. Twitter return Consumer Key and Consumer Secret for application which acts as credentials for the application and later on to update the settings.

II. Data Collection

Tweets are collected from the system by using the secret tokens of the twitter application. From this text mining, the lists of search strings used to collect the tweets are saved in the database. Tweets basically selected based on the search string, keywords and geo location.

III. Data Pre-processing

The data in the form of tweets are saved to a excel file along with the database. After that, preprocessing of the text is done before using classifier. All the hashtags, @ signs, punctuations, symbols, smiles or special characters are also removed. RT which is used to retweet and all the hyperlinks are removed. Compress the duplicate letters if they appear more than twice.

IV. Computation of Sentiment Score

a) Data Collection

There are very few existing data sets of social media sentiment messages. We collected our own set of data. For the testing data, we collected messages related to search string "Brexit" from the social media API.

Tweet Te	User Name	Retweet (	Favourite	Full Tweet	Create At	Source
brexit	hexeana	44	0	RT bomanirani: On my room balcony posing with the f	06-25-2016	Twitter for Android
brexit	rshang	14	0	RT shyammundhada: The most articulate take on Brex	06-25-2016	Twitter for Android
brexit	coffeekenwo	2	0	RT AndrewTMackay: coffeekenworld CoffeeAcademy	06-25-2016	Twitter Web Client
brexit	KhabarAntar	0	0	Delegates at 2016 Cannes Lions Festival feel the Brexi	06-25-2016	Google
brexit	KhabarAntar	0	0	Brexit: Rich British Indians voted for 'remain'; the less	06-25-2016	Google
brexit	a18life	0	0	"@ReutersBiz: Law firms see short-term opportunity,	06-25-2016	Twitter for Android
brexit	rpalleyi	131	0	RT minhazmerchant: Brexit will help India seal favour	06-25-2016	Twitter for Android
brexit	3NovicesHyd	0	0	3Novices : Brexit: Students debate impact, what this n	06-25-2016	IFTTT
brexit	sowmay_jain	0	0	Brexit Fallout: Don't stay out of market -	06-25-2016	MailChimp
brexit	prampara	33	0	RT gsurya: Britain: Xenophobic liars managed to convi	06-25-2016	Twitter for iPad
brexit	_QuickieNew	0	0	DECODED: How 'Brexit' threatens to undermine U.S.-B	06-25-2016	IFTTT
brexit	flyroundthew	0	0	Viewpoint: Brexit puts UK on new economic path	06-25-2016	IFTTT
brexit	rkickris	0	0	Sorry to see this happening...Brexit	06-25-2016	Facebook
brexit	thesankumar	904	0	RT RJ_Balaji: Brexit EvloPeriyaKevalamTheriyuma	06-25-2016	Twitter for Android
brexit	Harshank23	0	0	entertainment Brexit meets It's Always Sunny in fan-r	06-25-2016	IFTTT
brexit	Ahteshamazr	531	0	RT BDUTT: Can i have some of the charas you're smoki	06-25-2016	Twitter for Android
brexit	chewonstraw	0	1	This whole brexit thing has shown how ignorant peop	06-25-2016	Twitter for Android
brexit	amitmahend	0	0	DECODED: How 'Brexit' threatens to undermine U.S.-B	06-25-2016	dlvr.it
brexit	AnzzDevilzz	0	0	No offence but Hahaha lilyallen Brexit	06-25-2016	Twitter for Android
brexit	pultuskpa	80	0	RT simonmundy: Great explainer of Brexit result, from	06-25-2016	Twitter for Android

Figure 1 Sample Tweet Collection

b) Classification

Several different classifiers are there, but for text classification one can use Naive Bayes. Naive Bayes is a simple model for classification. It is simple and works well on text classification. Multinomial Naive Bayes model assumes each feature is conditional independent to other features given the class.

Algorithm: Naive bayes Classifier

Location based social media data are analyzed from social media messages based on events or search strings. Java based version of naive bayes algorithm is implemented to classify social media messages into positive, negative and neutral classes.

Steps

1 Create a data for the classifier
1.1 Create a list of positive messages
1.2 Create a list of negative messages
1.3 Create a list of neutral messages
1.4 Convert these lists into single list with word array for each message and its type.
2 Design a Classifier
2.1 Extract the word feature list from the list with its frequency count
2.2 Using this words list, create feature extractor which contains the words which will matched with a dictionary created by us indicating what words are contained in the input passed
3 Training the Classifier using training dataset
4 Calculate the probability for the positive, negative and neutral sentiment.
5 Compare this probability to identify the message category as positive, negative or neutral and generate sentiment score.

V. MATHEMATICAL MODEL

We now present a mathematical model for the location based sentiment analysis of twitter data.

Where system can be summarized as,

S= {I, O, F, U}

Where,
I: Input
O: Output
F: Functions
U: User

Input

I= {U, UT, IS, L}

Where,
U = User having a twitter application authentication keys
UT= User Tweets which will be extracted using authentication key
IS = Input String provided as input for extracting event related tweets
L= {L1, L2, L3} = Location set provided for extracting tweets from particular location

Output

O = {UP, PT, SA, VA}

Where below are the output generated from system processing
UP= Retrieved User Profile details
PT= Processed Tweets by removing unwanted urls, emoticons, keywords, stop words etc.
SA= Sentiment Analysis on processed tweets will provide sentiment score
VA= Volume Analysis will provide the analysis based on location wise volume analysis

User

U = {SV, TU, A}

Where,
SV = System Visitor
TU = Tweeter User whose account is used to retrieve tweets related to search string
A= Administrator

Functions

F= {F1, F2, F3, F4, F5}

Where,
1) Function F1: This function to authenticate twitter applications with given set of keys.
2) Function F2: This function retrieves the tweets through secure authentication using OAuth

 ISSN: 2321-8169
50 - 55

3) Function F3: This function process the extracted tweets for pre-processing
4) Function F4: This function performs location based sentiment analysis by classifying text and generating sentiment score.
5) Function F5: This function generates analysis results.

In text classification,
We are given a description t ∈ X of a tweet, where X is the twitter space; and a fixed set of classes C = {c1, c2. . . cJ}. Classes are also called categories or labels. Typically, the twitter space X is some type of high-dimensional space, and the classes are human defined for the needs of an application,
We are given a training set T of labelled tweets <t, c>, where <t, c> ∈ X × C.
Using a *learning algorithm*, we then wish to learn a classifier or *classification function* ω that maps tweets to classes:

$$\omega\!: \boldsymbol{X} \rightarrow \boldsymbol{C} \qquad (1)$$

This type of learning is called supervised learning, in our research we have opted multinomial naive bayes classification algorithm. It is a probabilistic learning method. The probability of a tweet *t* being in class *c* is computed as

$$P\,(c|t) \propto P(c) \prod_{1 \le k \le n_t} P(w_k|c) \qquad (2)$$

Where, $P\ (w_k|c)$ is the conditional probability of term w_k occurring in a tweet of class *c*.
We interpret $P\ (w_k|c)$ as a measure of how much evidence w_k contributes that *c* is the correct class.
P(*c*) is the prior probability of a tweet occurring in class *c*.
If a tweet's terms do not provide clear evidence for one class versus another, we choose the one that has a higher prior probability.
$<w_1, w_2 . \ . \ . \ , w_{nt}>$ are the tokens in *t* that are part of the vocabulary used for classification and *nt* is the number of such tokens in *t*.
.

VI. Experimental Setup and Results

In this section, we have described location based predication model for sentiment analysis and results are obtained after testing it on dataset containing tweets related to event "Brexit".

A. *Dataset*

The dataset for the system is the particular event based location wise extracted tweets. For Location based Sentiment analysis positive, negative or neutral thesaurus will be provided.

B. *Results*

As discussed above the system will generate sentiment analysis. The tweets related to event "Brexit" are downloaded location wise from twitter database and are pre-processed. These pre-processed tweets will be used as input for sentiment analysis. Tweets with positive sentiments have sentiment score 3, tweets with negative sentiments have sentiment score 1 and with neutral sentiments tweets have assigned sentiment score as 2. Further analysis to provide smarter decision is based on sentiment score obtained from sentiment analysis. Below Figure 2 shows the location wise tweet count.

Centralized system: This system can run on any standard machine with operating system Windows7 and upward compatibility. Programming language used JAVA/J2EE 6.

The datasets used in the application were retrieved from twitter using twitter4j API's. Further pre/post processing and data analysis was conducted as shown in results. Screen capture in this section shows result analysis of location based sentiment analysis of event "Brexit". Figure 1 represents sample tweet collection.
Testing dataset details are:
Total number of tweets extracted-15,274
Tweets collected from London -13,597
Tweets collected from New York-1,298
Tweets collected from Mumbai -84

Location	Positive	Negative	Neutral	Total	SearchKey
Mumbai	7	53	24	84	brexit
Newyork	63	1055	180	1298	brexit
London	823	10262	2512	13597	brexit

Figure 2 Location wise tweet count

Location wise tweet analysis is presented in Figure 2.

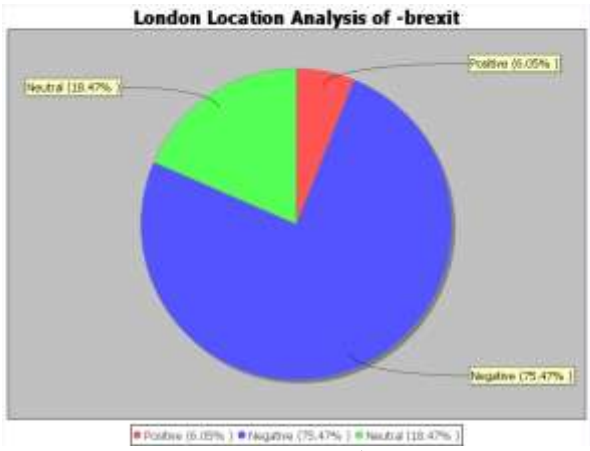

Figure 3 Location based sentiment analysis of tweet dataset at London

Figure 3, 4, 5 visualizes the location wise popularity of the event "Brexit" in the form of pie chart.

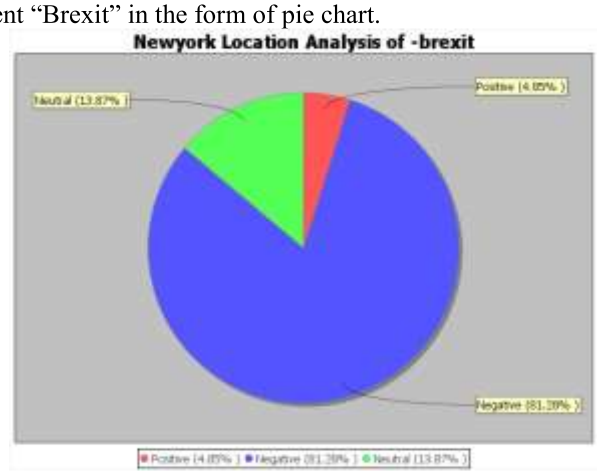

Figure 4 Location based sentiment analysis of tweet dataset at Newyork

ISSN: 2321-8169
50 - 55

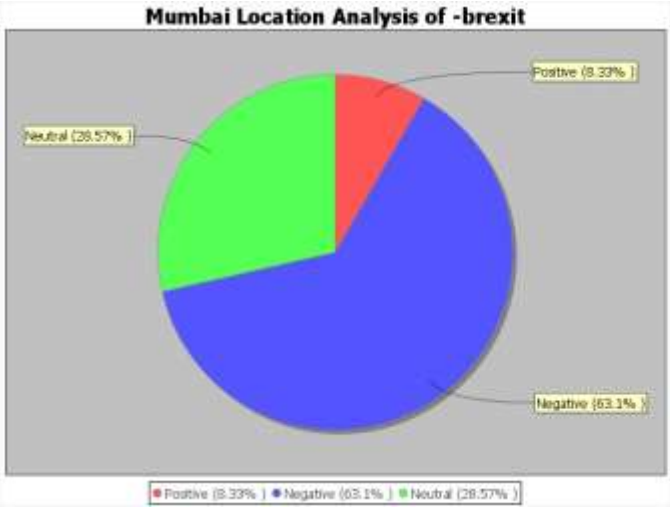

Figure 5 Location based sentiment analysis of tweet dataset at Mumbai

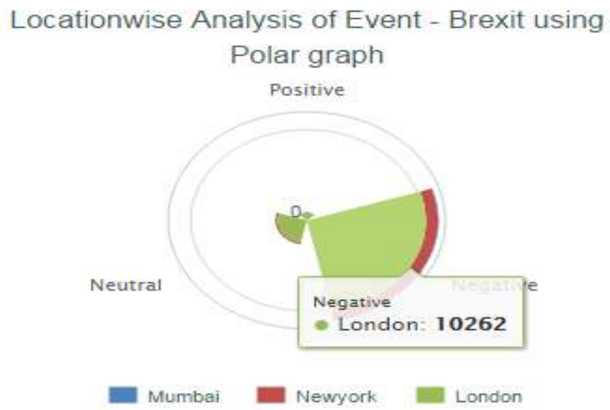

Figure 6 Location based analysis of event "Brexit" using Polar graph

We have analyzed results in the form of pie chart using our built-in visualization engine, apart from that we integrated visualization tool like Jedox to use advance level of visualization. Fig. 6 and 7 present the different views of sentiments of people about event "Brexit" from different locations.

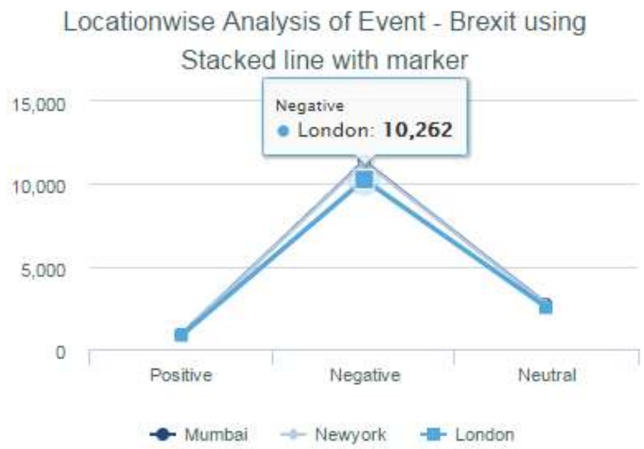

Figure 7 Location based sentiment analysis of event "Brexit" using Staked line with marker

VII. CONCLUSION AND FUTURE SCOPE

The purpose of this work was to scale location base popularity for different products or events and use the output for smarter decision making in real-life. Multiple dimension of business aspects have been investigated, analyzed and visualized. An automation of sentiment analysis has been obtained on data collected from social network.

Location based scaling can be applied to different analysis to control the decision making in product marketing, sentiment analysis, event detection, trend identification, election forecasting, health-related information retrieval and epidemic identification. Through a step by step process, it has been proved that the research is capable of performing location based scaling on different products and events very efficiently. Hence this research is successful for performing its proposed goal of location based popularity scaling.

In near future we are planning to evaluate the larger selection of machine learning algorithms on the data for sentiment analysis. Finally, if sentiment analysis can be subjected to psychology field, we can see in a distant future computer science and engineering field will be in a position to quantify the concepts thus helping one to provide the brain mapping with significant contributions to the understanding of psychology and human thought process.

REFERENCES

[1] Aarati Patil, Srinivasa Narasimha Kini, IRSDM of SocioData: Location Based Analysis of Opinion Diffusion on Social Network, fifth post graduate conference of computer engineering, cpgcon 2016.

[2] Aarati Patil, Srinivasa Narasimha Kini, Evolution of Social Media from the era of Information Retrieval, International Journal of Science and Research (IJSR), 4(12):2326-2331, 2015

[3] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification Using Machine Learning Techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)

[4] Wiebe, J. and E. Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. Computational Linguistics and Intelligent Text Processing, pages 486497

[5] Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford (2009)

[6] Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of LREC.

[7] Barbosa, L., Feng, J.: Robust sentiment detection on Twitter from biased and noisy data. In: Proceedings of COLING. pp. 3644 (2010)

[8] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of Twitter data. In: Proc. ACL 2011Workshop on Languages in Social Media. pp. 3038 (2011)

[9] K. Efthymios, W. Theresa, and J. Moore, Twitter Sentiment Analysis: The Good the Bad and the OMG! in Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain on 2011, pp. 538 541

[10] Felipe , M. Marcelo, and P. Barbara, Combining strengths, emotions and polarities for boosting Twitter sentiment analysis, in Second International Workshop on Issues of

Sentiment Discovery and Opinion Mining, ACM New York, 2013, pp. 202-207

[11] W. Haoet , C. Dogan, K. Abe, B. Franois, and N. Shrikanth , A system for realtime Twitter sentiment analysis of 2012 U.S. presidential election cycle, in 12th Proceedings of the ACL System Demonstrations, Stroudsburg, USA, 2012, pp. 115-120

[12] B. Johan, M. Huina, and P. Alberto, Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena, in Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, July, 2011, pp. 451-453

[13] T. Jiliang, Y. Chang, and L. Huan, Mining Social Media with Social Theories: A Survey, in ACM SIGKDD Explorations Newsletter, New York, 2013, pp. 20-29

[14] Jian Zhao, Nan Cao, ZhenWen, Yale Song, Yu-Ru Lin, and Christopher Collins, FluxFlow: Visual Analysis of Anomalous Information Speading on Social Media, IEEE transactions on visualization and comfputer graphics, 20(12): 1773-1782 , 2014

[15] Yingcai Wu, Member, IEEE, Shixia Liu, Senior Member, IEEE, Kai Yan, Mengchen Liu, FangzhaoWu, OpinionFlow: Visual Analysis of Opinion Diffusion on Social Media, IEEE transactions on visualization and computer graphics, 20(12): 1763-1772, 2014