International Journal of Computing and Corporate Research
ISSN (Online) : 2249-054X
Volume 6 Issue 3 May - June 2016
International Manuscript ID : 2249054XV6I3052016-78

EFFECTIVE ALGORITHMIC APPROACHES FOR ANALYSIS AND VISUALISATION OF SOCIAL MEDIA

Hermannpreet Kaur
M.Tech. Research Scholar
Computer Science and Engineering
Punjabi University
Patiala, Punjab, India

Dr. Neeraj Sharma
Punjabi University
Patiala, Punjab, India

Dr. Kawaljeet Singh
Director
University Computer Centre
Punjabi University
Patiala, Punjab, India

ABSTRACT

Big data analytics is the process of examining large data sets containing a variety of data types -- i.e., big data -- to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. The analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations and other business benefits. The primary goal of big data analytics is to help companies make more informed business decisions by enabling data scientists, predictive modelers and other analytics professionals to analyze large volumes of transaction data, as well as other forms of data that may be untapped by conventional business intelligence (BI) programs. That could include Web server logs and

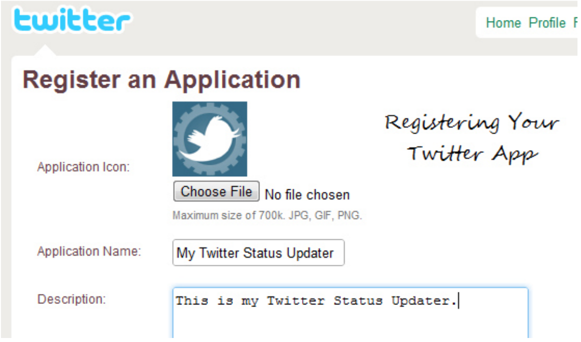
Internet clickstream data, social media content and social network activity reports, text from customer emails and survey responses, mobile-phone call detail records and machine data captured by sensors connected to the Internet of Things. Some people exclusively associate big data with semi-structured and unstructured data of that sort, but consulting firms like Gartner Inc. and Forrester Research Inc. also consider transactions and other structured data to be valid components of big data analytics applications. In this research work, the live tweets from social media are fetched and analyzed using Java and Python based platforms and performance evaluation is done on assorted parameters. This research work in addition focus towards the similarity score of assorted and prominent online shopping platforms based on their related followers. In this research work, a social media is worked out and developed so that the social community with respect to academics and research can be associated effectively.

INTRODUCTION

Twitter describes itself as a "real-time information network," a network that primarily connects one person to many in short, 140-character messages. Using it requires only a free Twitter account and a computer with Internet access (or a phone with SMS capability).

While Twitter has been around since 2006, it really came to prominence when a user posted the first picture of a plane floating on the Hudson River with its passengers standing on the wings. Initially, Twitter was used to tell others what one was doing or thinking. Celebrities, notably Ashton Kutcher, helped it garner public attention for that kind of use. Politicians then used it to tweet their thoughts and make calls for action. Businesses soon followed with introductions of their products, coupons or special Twitter promotions, and integration with other social media messages. Jack–in-the-Box integrated it with a Super Bowl video and Facebook page where Jack was hit by a bus, his condition reported as tweets along with an insider story about a takeover. It has also been used for posting job announcements and for backchanneling during meetings. And today companies are constantly monitoring it for public and customer relations purposes. Clearly, Twitter is an evolving and growing communication medium for a variety of social messaging purposes.

In the same way, the twitter live feeds can be fetched using Python APIs. Using twitter developer account, the new app can be created and then the Python Script is mapped with the Twitter App



Register an Application

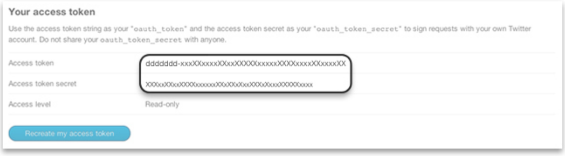
Application Icon:

Maximum size of 700k. JPG, GIF, PNG.

Application Name: My Twitter Status Updater

Description: This is my Twitter Status Updater.

Figure 1 – Creating New App in Twitter



Your access token

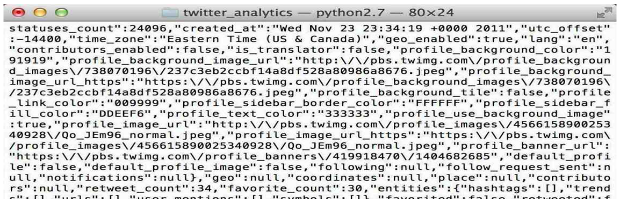
Use the access token string as your "oauth_token" and the access token secret as your "oauth_token_secret" to sign requests with your own Twitter account. Do not share your oauth_token_secret with anyone.

Access token: dddddd-xxxXXxxxxXXxxXXXXxxxxxXXXXxxxxXXxxxxXX

Access token secret: XXXxxXXxxXXXXxxxxxxXXxXXxXxxXXXxXxxxXXXXXxxxx

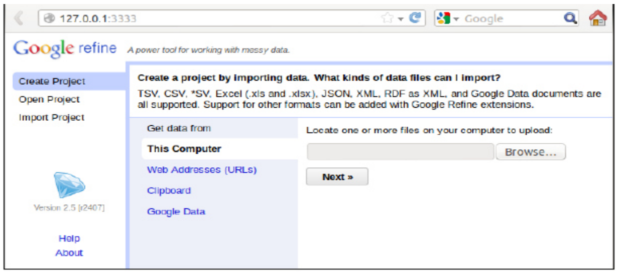
Access level: Read-only

Figure 2 – Generation of Authentication Tokens from Twitter



```
statuses_count":24096,"created_at":"Wed Nov 23 23:34:19 +0000 2011","utc_offset"
:-14400,"time_zone":"Eastern Time (US & Canada)","geo_enabled":true,"lang":"en",
"contributors_enabled":false,"is_translator":false,"profile_background_color":"1
91919","profile_background_image_url":"http:\/\/pbs.twimg.com\/profile_backgroun
d_images\/738070196\/237c3eb2ccbf14a8df528a80986a8676.jpeg","profile_background_
image_url_https":"https:\/\/pbs.twimg.com\/profile_background_images\/738070196\
/237c3eb2ccbf14a8df528a80986a8676.jpeg","profile_background_tile":false,"profile
_link_color":"009999","profile_sidebar_border_color":"FFFFFF","profile_sidebar_f
ill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image"
:true,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/456615890025З
40928\/Qo_JEm96_normal.jpeg","profile_image_url_https":"https:\/\/pbs.twimg.com\
/profile_images\/456615890025340928\/Qo_JEm96_normal.jpeg","profile_banner_url":
"https:\/\/pbs.twimg.com\/profile_banners\/419918470\/1404682685","default_profi
le":false,"default_profile_image":false,"following":null,"follow_request_sent":n
ull,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributo
rs":null,"retweet_count":34,"favorite_count":30,"entities":{"hashtags":[],"trend
s":[],"urls":[],"user_mentions":[],"symbols":[]},"favorited":false,"retweeted":f
```

Figure 3 – Fetching Live Tweets from Twitter in JSON Format



Create a project by importing data. What kinds of data files can I import?

TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats can be added with Google Refine extensions.

Locate one or more files on your computer to upload:

Figure 4 – Parsing of JSON using Google Refine

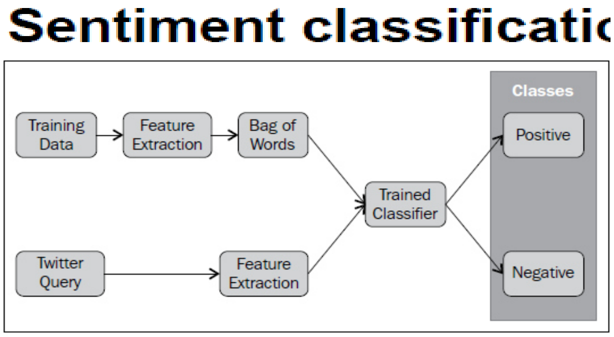


Figure 5 – Generation of Sentiments in Different Classes

As per the international statistical reports from Statista, there are around 1 million new user registrations on Whatsapp. Besides this, 700 million active users present on Whatsapp. Around 30 Billion messages are sent and 34 billion messages are received everyday. If we analyze the statistics of Twitter, 350 Million Tweets daily and more than 500 Million Accounts. There is the huge and rapid growth in the unstructured data every moment. The production and generation of data is predicted to be 44 times in 2020 as compared to the data in 2009. All these figures and statistical data are amazing and growing in exponential pattern. Such data is unstructured in nature which means the data of different and heterogeneous formats. This concept is classically known as Big Data. The deep investigation of intelligence and meaningful patterns from Big Data is known as Big Data Analytics. A number of researchers and scientists are working in this domain of Big Data using assorted technologies and tools. There are number of approaches by which the live data can be obtained for research and development. One of these approaches is getting data from Open Data Portals. The open data portals provide authentic data sets for research and development in multiple domains. The data sets can be downloaded from these portals in multiple formats including XML, CSV, JSON and many others. In this research work, the sentiment data analysis and prediction shall be done on the big data fetched using Python Scripts. Using prediction tools and algorithms, the effective and accurate results shall be measured.

REAL WORLD APPLICATIONS

- Banking and Finance
- Marketing and Business Policies
- Consumer Behavior
- Market Basket Analysis
- Military and Defense
- Log Files Analysis
- Forensic Investigation
- Bio-Statistics and Bio-Informatics
- Web Usage Mining
- Medical Data Mining
- Expert Systems
- Knowledge Discovery

LITERATURE SURVEY

For completion, justification and solving the problem definition, a number of research papers, magazines, journals and online links are investigated in details.

A number of research scholars and scientists has written a number of research papers and found excellent results. This section underlines all those research papers and their extracts

Bollen (2010) - Behavioral economics tells us that emotions can profoundly affect individual behavior and decision-making. Does this also apply to societies at large, i.e. can societies experience mood states that affect their collective decision making? By extension is the public mood correlated or even predictive of economic indicators? Here we investigate whether measurements of collective mood states derived from large-scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA) over time. This work analyze the text content of daily Twitter feeds by two mood tracking tools, namely OpinionFinder that measures positive vs. negative mood and Google-Profile of Mood States (GPOMS) that measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy). This paper cross-validate the resulting mood time series by comparing their ability to detect the public's response to the presidential election and Thanksgiving day in 2008. A Granger causality analysis and a Self-Organizing Fuzzy Neural Network are then used to investigate the hypothesis that public mood states, as measured by the OpinionFinder and GPOMS mood time series, are predictive of changes in DJIA closing values. The results in this paper indicate that the accuracy of DJIA predictions can be significantly improved by the inclusion of specific public mood dimensions but not others. The authors find an accuracy of 87.6% in predicting the daily up and down changes in the closing values of the DJIA and a reduction of the Mean Average Percentage Error by more than 6%.

Bollen (2009) - Microblogging is a form of online communication by which users broadcast brief text updates, also known as tweets, to the public or a selected circle of contacts. A variegated mosaic of microblogging uses has emerged since the launch of Twitter in 2006: daily chatter, conversation, information sharing, and news commentary, among others. Regardless of their content and intended use, tweets often convey pertinent information about their author's mood status. As such, tweets can be regarded as temporally-authentic microscopic instantiations of public mood state. In this article, we perform a sentiment analysis of all public tweets broadcasted by Twitter users between August 1 and December 20, 2008. For every day in the timeline, we extract six dimensions of mood (tension, depression, anger, vigor, fatigue, confusion) using an extended version of the POMS, a well-established psychometric instrument. The authors in this paper compare the results to the values recorded by stock market and crude oil price indices and major events in media and popular culture, such as the U.S. Presidential Election of November 4, 2008 and Thanksgiving Day. This work finds that events in the social, political, cultural and economic sphere do have a significant, immediate on the various dimensions of public mood. The authors speculate that large scale analyses of mood can provide a solid platform to model collective emotive trends in terms of their predictive value with regards to existing social as well as economic indicators.

Asur (2010)– Sentiment Analysis is important part for social networking and content sharing. And yet, the content that is generated from these websites remains largely untapped. In this paper, the authors demonstrate how social media content can be used to predict real-world outcomes. In particular, this work uses the chatter from Twitter.com to forecast box-office revenues for movies. This paper shows that a simple model built from the rate at which tweets are created about particular topics can outperform market-based predictors. This work further demonstrates how sentiments extracted from Twitter can be further utilized to improve the forecasting power of social media.

Tan (2011)– The authors show that information about social relationships can be used to improve user-level sentiment analysis. The main motivation behind the approach is that users that are somehow “connected” may be more likely to hold similar opinions; therefore, relationship information can complement what we can extract about a user’s viewpoints from their utterances. Employing Twitter as a source for our experimental data, and working within a semi-supervised framework, we propose models that are induced either from the Twitter follower/followee network or from the network in Twitter formed by users referring to each other using “@” mentions. The proposed transductive learning results reveal that incorporating social-network information can indeed lead to statistically significant sentiment classification improvements over the performance of an approach based on Support Vector Machines having access only to textual features.

Saif (2012) - Sentiment analysis over Twitter offer organisations a fast and effective way to monitor the publics’ feelings towards their brand, business, directors, etc. A wide range of features and methods for training sentiment classifiers for Twitter datasets have been researched in recent years with varying results. In this paper, we introduce a novel approach of adding semantics as additional features into the training set for sentiment analysis. For each extracted entity (e.g. iPhone) from tweets, we add its semantic concept (e.g. “Apple product”) as an additional feature, and measure the correlation of the representative concept with negative/positive sentiment. The authors apply this approach to predict sentiment for three different Twitter datasets. The results show an average increase of F harmonic accuracy score for identifying both negative and positive sentiment of around 6.5% and 4.8% over the baselines of unigrams and part-of-speech features respectively. We also compare against an approach based on sentiment-bearing topic analysis, and find that semantic features produce better Recall and F score when classifying negative sentiment, and better Precision with lower Recall and F score in positive sentiment classification.

Davidov (2010) - Automated identification of diverse sentiment types can be beneficial for many NLP systems such as review summarization and public media analysis. In some of these systems there is an option of assigning a sentiment value to a single sentence or a very short text. In this paper the authors proposes a supervised sentiment classification framework which is based on data from Twitter, a popular microblogging service. By utilizing 50 Twitter tags and 15 smileys as sentiment labels, this framework avoids the need for labor intensive manual annotation, allowing identification and classification of diverse sentiment types of short texts. The authors evaluate the contribution of different feature types for sentiment classification and show that the proposed framework successfully identifies sentiment types of untagged sentences. The quality of the sentiment identification was also confirmed by human judges. This paper also explores dependencies and overlap between different sentiment types represented by smileys and Twitter hashtags.

Saif (2012) - Twitter has brought much attention recently as a hot research topic in the domain of sentiment analysis. Training sentiment classifiers from tweets data often faces the data sparsity problem partly due to the large variety of short and irregular forms introduced to tweets because of the 140-character limit. In this work the authors proposes using two different sets of features to alleviate the data sparseness problem. One is the semantic feature set where this work extracts semantically hidden concepts from tweets and then incorporate them into classifier training through interpolation. Another is the sentiment-topic feature set where we extract latent topics and the associated topic sentiment from tweets, then augment the original feature space with these sentiment-topics. Experimental results on the Stanford Twitter Sentiment Dataset show that both feature sets outperform the baseline model using unigrams only. Moreover, using semantic features rivals the previously reported best result. Using sentiment topic features achieves 86.3% sentiment classification accuracy, which outperforms existing approaches.

Bifet (2009) - Micro-blogs are a challenging new source of information for data mining techniques. Twitter is a micro-blogging service built to discover what is happening at any moment in time, anywhere in the world. Twitter messages are short, and generated constantly, and well suited for knowledge discovery using data stream mining. The authors briefly discuss the challenges that Twitter data streams pose, focusing on classification problems, and then consider these streams for opinion mining and sentiment analysis. To deal with streaming unbalanced classes, this work propose a sliding window Kappa statistic for evaluation in time-changing data streams. Using this statistic this work performs a study on Twitter data using learning algorithms for data streams.

Social networking, blogging and online forums have turned the web into a vast repository of comments on many topics, generating a potential source of information for social science research (Thelwall, Wouters, & Fry, 2008). The availability of large scale electronic social data from the web and elsewhere is already transforming social research (Savage & Burrows, 2007). The social web is also being commercially exploited for goals such as automatically extracting customer opinions about products or brands. An application could build a large database of web sources (Bansal & Koudas, 2007; Gruhl, Chavet, Gibson, Meyer, & Pattanayak, 2004), use information retrieval techniques to identify potentially relevant texts, then extract information about target products or brands, such as which aspects are disliked (Gamon, Aue, Corston-Oliver, & Ringger, 2005; Jansen, Zhang, Sobel, & Chowdury, 2009). From a social sciences perspective, similar methods could potentially give insights into public opinion about a wide range of topics and are unobtrusive, avoiding human subjects research issues (Bassett & O'Riordan, 2002; Enyon, Schroeder, & Fry, 2009; Hookway, 2008; White, 2002).

The sheer size of the social web has also made possible a new type of informal literature-based discovery (for literature based discovery, see: Bruza & Weeber, 2008; Swanson, Smalheiser, & Bookstein, 2001): the ability to *automatically* detect events of interest, perhaps within pre-defined broad topics, by scanning large quantities of web data. For instance, one project used time series analyses of (mainly) blogs to identify emerging public fears about science (Thelwall & Prabowo, 2007) and businesses can use similar techniques to quickly discover customer concerns. Emerging important events are typically signalled by sharp increases in the frequency of relevant terms. These bursts of interest are important to study because of their role in detecting new events as well as for the importance of the events discovered.

One key unknown is the role of sentiment in the emergence of important events because of the increasing recognition of the importance of emotion in awareness, recall and judgement of information (Fox, 2008, p. 242-244, 165-167, 183; Kinsinger &

Schacter, 2008) as well as motivation associated with information behaviour (Case, 2002, p. 71-72; Nahl, 2006, 2007a).

PROPOSED MODEL

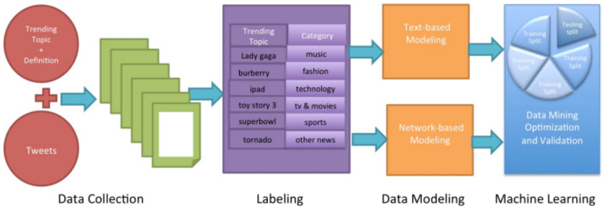


Figure 6 – Machine Learning and Data Modeling Process

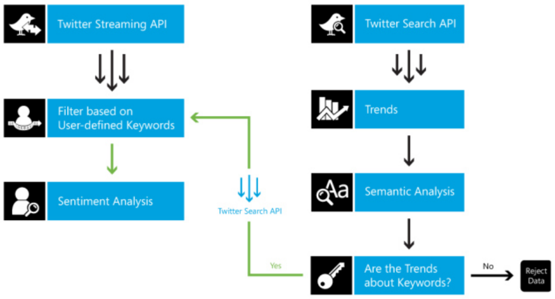


Figure 7 – Sentiment Analysis

HARDWARE REQUIREMENTS

The minimum requirements needed to perform operations are

- Intel Pentium Processor at 2 GHz or Higher
- RAM 256MB or more
- Hard disk capacity 10GB or more

The software required to perform the implementation are

- Windows or Linux Operating System (Ubuntu, Fedora)
- Advance Java
- Twitter APIs
- AJAX
- MySQL Database Engine
- Notepad++
- Twitter4J
- Eclipse IDE
- GEPHI
- JSON (JavaScript Object Notation)
- WEKA - Data Mining and Machine Learning Tool

RESEARCH OBJECTIVES

- To study various research issues related to social web analytics and efficient data mining techniques for analyzing and visualizing the social data . This social data includes both user's profile information as well as user opinions and sentiments
- To collect and analyze users information from twitter's account and use it to login or connect with other websites.
- The research follows to understand and study how useful it is to deploy users information from twitter account. Answering this question is the basic objective of my research.
- To perform basic frequency analysis , tweet-retweet analysis and sentiment analysis on twitter data

Simulation Attempt	Java	Python
1	0.232353	0.138383
2	0.599444	0.395944
3	0.78393939	0.588822
4	0.1939393	0.0788982
5	0.56789	0.323444

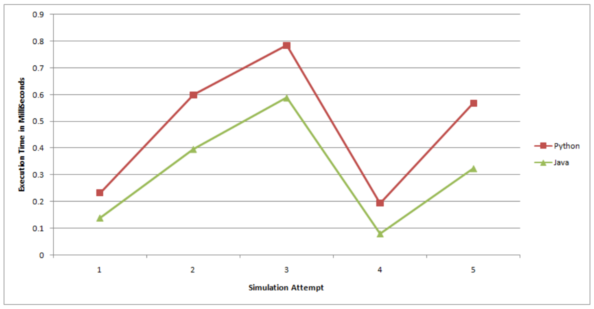


Figure 8 – Comparison of performance between Java and Python Platforms

FOLLOWERS ANALYSIS OF PROMINENT SHOPPING PORTALS

Amazon	Ebay	Snapdeal
304203	165277	299155

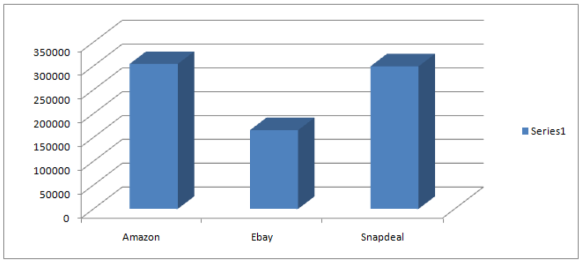


Figure 9 – Comparison of followers

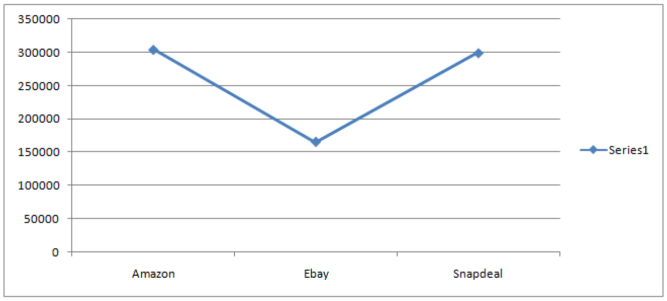
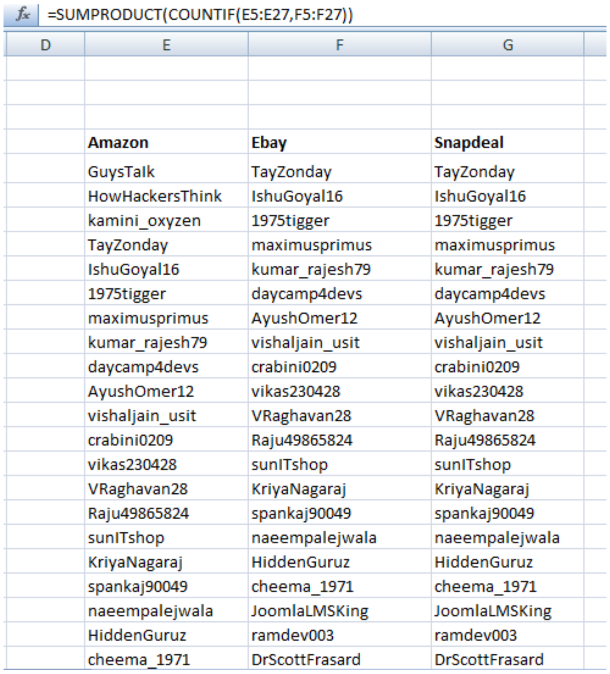


Figure 10 – Comparison of followers of shopping portals



=SUMPRODUCT(COUNTIF(E5:E27,F5:F27))

Amazon	Ebay	Snapdeal
GuysTalk	TayZonday	TayZonday
HowHackersThink	IshuGoyal16	IshuGoyal16
kamini_oxyzen	1975tigger	1975tigger
TayZonday	maximusprimus	maximusprimus
IshuGoyal16	kumar_rajesh79	kumar_rajesh79
1975tigger	daycamp4devs	daycamp4devs
maximusprimus	AyushOmer12	AyushOmer12
kumar_rajesh79	vishaljain_usit	vishaljain_usit
daycamp4devs	crabini0209	crabini0209
AyushOmer12	vikas230428	vikas230428
vishaljain_usit	VRaghavan28	VRaghavan28
crabini0209	Raju49865824	Raju49865824
vikas230428	sunITshop	sunITshop
VRaghavan28	KriyaNagaraj	KriyaNagaraj
Raju49865824	spankaj90049	spankaj90049
sunITshop	naeempalejwala	naeempalejwala
KriyaNagaraj	HiddenGuruz	HiddenGuruz
spankaj90049	cheema_1971	cheema_1971
naeempalejwala	JoomlaLMSKing	JoomlaLMSKing
HiddenGuruz	ramdev003	ramdev003
cheema_1971	DrScottFrasard	DrScottFrasard

Figure 11 – Similarity Index

Similarity Score (Amazon-Ebay)	Similarity Score (Amazon-Snapdeal)	Similarity Score (Ebay-Snapdeal)
60	48	59

CONCLUSION

Fetching the live social media or related dimension sentiment analysis is under research from a long time for detailed analysis and prediction of the events with respect to the social cause. There is huge scope of research and development using Java scripts and specialized APIs for assorted applications including cyber security, data mining, Internet of Things, cloud simulation, grid implementation and many others. Java is one of the effective programming languages that can process and handle any type of data stream. As per the international statistical reports from Statista, there are around 1 million new user registrations on Whatsapp. Besides this, 700 million active users present on Whatsapp. Around 30 Billion messages are sent and 34 billion messages are received everyday. If we analyze the statistics of Twitter, 350 Million Tweets daily and more than 500 Million Accounts. There is the huge and rapid growth in the unstructured data every moment. The production and generation of data is predicted to be 44 times in 2020 as compared to the data in 2009. All these figures and statistical data are amazing and growing in exponential pattern. Such data is unstructured in nature which means the data of different and heterogeneous formats. This concept is classically known as Big Data. The deep investigation of intelligence and meaningful patterns from Big Data is known as Big Data Analytics. A number of researchers and scientists are working in this domain of Big Data using assorted technologies and tools. There are number of approaches by which the live data can be obtained for research and development. One of these approaches is getting data from Open Data Portals. The open data portals provide authentic data sets for research and development in multiple domains. The data sets can be downloaded from these portals in multiple formats including XML, CSV, JSON and many others.

REFERENCES

[1] Allan, J., Papka, R., & Lavrenko, V. (1998). On-line new event detection and tracking In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 37-45). New York, NY: ACM Press.

[2] Archak, N., Ghose, A., & Ipeirotis, P. G. (2007). Show me the money!: Deriving the pricing power of product features by mining consumer reviews. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 56-65). New York, NY: ACM Press.

[3] Balog, K., Mishne, G., & Rijke, M. d. (2006). Why are they excited? Identifying and explaining spikes in blog mood levels. 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL 2006), Retrieved July 8, 2010 from: http://staff.science.uva.nl/~mdr/Publications/Files/eacl2006-moodsignals.pdf.

[4] Bansal, N., & Koudas, N. (2007). BlogScope: a system for online analysis of high volume text streams. In Proceedings of

the 33rd international conference on Very large data bases (pp. 1410-1413). New York, NY: ACM Press.

[5] Bassett, E. H., & O'Riordan, K. (2002). Ethics of Internet research: Contesting the human subjects research model Ethics and Information Technology, 4(3), 233-247.

[6] Bifet, A., & Frank, E. (2010). Sentiment knowledge discovery in Twitter streaming data. In Proc 13th International Conference on Discovery Science (pp. 1-15). Berlin: Springer.

[7] Blumler, J. G., & Katz, E. (1974). The uses of mass communications: Current perspectives on gratifications research. Beverly Hills, CA: Sage.

[8] Bollen, J., Pepe, A., & Mao, H. (2009). Modeling public mood and emotion: Twitter sentiment and socioeconomic phenomena. arXiv.org, arXiv:0911.1583v0911 [cs.CY] 0919 Nov 2009.

[9] boyd, d., Golder, S., & Lotan, G. (2009). Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. Proceedings of HICSS-43, Retrieved November 12, 2009 from: http://www.danah.org/papers/TweetTweetRetweet.pdf.

[10] Bruza, P., & Weeber, M. (2008). Literature-based discovery. Berlin: Springer.

[11] Case, D. O. (2002). Looking for information: A survey of research on information seeking, needs, and behavior. San Diego, CA: Academic Press.

[12] Cataldi, M., Caro, L. D., & Schifanella, C. (2010). Emerging topic detection on Twitter based on temporal and social terms evaluation In Proceedings of the Tenth International Workshop on Multimedia Data Mining table of contents (pp. A4). New York, NY: ACM Press.

[13] Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. P. (2010). Measuring user influence in Twitter: The million follower fallacy. In Proceedings of the International AAAI Conference on Weblogs and Social Media (pp. Retrieved August 9, 2010 from: http://an.kaist.ac.kr/~mycha/docs/icwsm2010_cha.pdf).

[14] Choudhury, M. D., Sundaram, H., John, A., & Seligmann, D. D. (2008). Can blog communication dynamics be correlated with stock market activity? In Proceedings of the Nineteenth ACM conference on Hypertext and Hypermedia (pp. 55-60). New York, NY: ACM Press.

[15] Diakopoulos, N. A., & Shamma, D. A. (2010). Characterizing debate performance via aggregated twitter sentiment. In Proceedings of the 28th international conference on Human factors in computing systems (pp. 1195-1198). New York, NY: ACM Press.