GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

SENTIMENT ANALYSIS OF THE TWITTER DATA

A Report for the Evaluation 3 of Project 2

Submitted by

SURBHI SINGH

(1613101765 / 16SCSE101281)

in partial fulfillment for the award of the degree

of

Bachelor of Technology

IN

Computer Science and Engineering

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

Under the Supervision of

Mr. Padhmanabhan P

Professor

APRIL / MAY- 2020

TABLE OF CONTENTS

ABSTRACT

Social media platforms are continuously gaining more attention in the current era. The various opinions, private opinions and public opinions about various subjects are expressed and are spread continuously through so many social media platforms. Twitter is one such platform that has gained immense popularity. Twitter provides organizations with a speedy and an effective way to analyze their customers' perspective towards the products and services that are critical to gain success in the marketplace. Building a model for tweet sentiment analysis has proved to be a valid approach which can be easily made to measure customers' perceptive computationally. This paper is a report on the design of model sentiment analysis, extracting a tremendous amount of tweets. The concept of Prototyping has been used in this development. Results are the classification of customers' perception via their tweets into negative and positive, which is depicted through a pie chart. However, it has been planned for development of a web application, but since Django's work is limited and can be worked on a Linux server.

CHAPTER 1
INTRODUCTION

According to [1], a lot of people are using social network sites to precise their emotions, opinion and disclose about their daily lives. However, people write anything such as social activities or any touch upon products. Through the online communities provide an interactive forum where consumers inform and influence others.Moreover, social media provides a chance for businesses by giving a platform to attach with their customers like social media to advertise or speak to customers for connecting with the customer's perspective of products and services.
In contrast, consumers have all the ability when it involves what consumers want to determine and the way consumers respond. With this, the company's success & failure is publicly shared and you find yourself with word of mouth. However, the social network can change the behavior and higher cognitive process of consumers, for example, [2] mentions that 87% of internet users are influenced in their purchase and decision by customer's review. So that, if organization can catch up faster on what their customers think, it might be more beneficial to prepare to react on time and are available up with an honest strategy to compete their competitors.

A. Problem Statement

Despite the provision of software to extract data regarding a person's sentiment on a selected product or service, organizations and other data workers still face issues regarding the information extraction.

- Sentiment Analysis of Web Based Applications specializes in Single Tweet Only.

With the ascent of the world Wide Web, people are using social media like Twitter which generates big volumes of opinion texts within the type of tweets which is available for the sentiment analysis [3]. This translates to a huge volume of data from a personality's viewpoint which makes it difficult to extract sentences, read them, analyze tweet by tweet, summarize them and organize them into an understandable format in a very timely manner [3].

- Difficulty of Sentiment Analysis with inappropriate English

Informal language refers to the utilization of colloquialisms and slang in communication, employing the conventions of spoken language [4] like 'would not' and 'wouldn't'. Not all systems are ready to detect sentiment from use of informal language and this might hanker the analysis and decision making process. Emoticons are a picturing of human facial expressions [5], which within the absence of communication and prosody serve to draw a receiver's attention to the tenor or temper of a sender's nominal verbal communication, improving and changing its interpretation [6]. For example, indicates a contented state of mind. Systems currently in place do not have sufficient data to allow them to draw feelings out of the emoticons. As humans often inter-communicate emoticons to properly express what they cannot put into words [6]. Not having the flexibility to analyze this puts the organization at a loss. Short-form is widely used even with short message service (SMS). The usage of short-form are visiting

be used more frequently on Twitter so on help to scale back the characters used. This is because Twitter has put limit on its characters to 140 [7]. For example, ‘Tba’ refers to an announcement**.**

B. Objective

The objectives of the study are first, to check the sentiment analysis in microblogging which visible to investigate feedback from a customer of an organization’s product; and second, is to develop a program for customers’ review on a product which allows a corporation or individual to sentiment and analyzes an enormous amount of tweets into a useful format.

CHAPTER 2
METHODOLOGY

The project is now being divided into the following two phases. In the first phase, literature review is conducted, which is followed by development of the system in the second phase. Literature review involves study of a variety of sentiment analysis methods and techniques that is currently being used. In the second phase, application's requirement and functionality are defined before the start of its development. Also, the architecture and the design interface of the program and how it is going to interact are also looked for. For the development of Twitter-Sentiment-Analysis application, a wide range of tools are utilized, such as Python Shell 2.7.2, Anaconda, Jupyter-Notebooks, Tensorflow API.

CHAPTER 3
LITERATURE REVIEW

A. Opinion Mining

Opinion mining refers to a very broad area of the field of natural language processing, computational linguistics, text mining which involves computational study of sentiment, opinion and emotion expressed in the text [8]. Although, attitude or view based on an emotion in spite of any reason is quite often referred to as sentiments [8]. It lends to a technique for or sentiment analysis opinion mining. [9] has stated that the method of opinion mining/sentiment analysis has a wide range of application domains including research, education, law, accounting, entertainment, politics, technology, marketing and so on. In earlier days many social media platforms have given web users the avenue for opening up to express and share their thoughts and opinions [10].

B. *Twitter*

Twitter is a popular real time micro blogging platform service that allows users to share short information known as tweets which are limited to 140 characters [2,3], [11]. Users write these micro blogs to express their opinion about various topics that are related to their daily lives. Twitter is an ideal social media platform for the extraction of the general public's perspective on specific issues [9,10]. A collection of tweets is used as the primary corpus or training set for our model for sentiment analysis, which refers to the use of natural language processing or opinion mining[1].

Twitter has more than 500 million active users and a million messages per day, it has quickly become an invaluable asset for the organizations to keep vigil on their reputation and brands by extracting and analyzing the sentiments of the tweets by the people about their product, service in the market and even about their competitors [12]. [2] has highlighted that, from the opinions social media has generated and with the exponential growth of the world wide web, super volumes of opinion texts in the form of reviews, tweets, blogs and any discussion groups and forums are available for analysis, thus making the internet the fastest, most promising and easily accessible medium for sentiment analysis and opinion mining.

C. Micro blogging with E-commerce

A micro blogging platform such as Twitter is not so likely as a conventional blogging platform just because single posts are shorter [13]. Twitter has limited for a small number of words that are designed for the speedy transmission of information or exchange of opinions [7]. However, small businesses or large organizations are initiating the potential of micro blogging as an e-commerce marketing tool [3]. Though, micro blogging platforms have been developed a few years’ time for promoting foreign trade websites by using a foreign micro blogging platform as Twitter marketing [3].

The instant of sharing, interactive, community-oriented features are opening an e-commerce, launched a new bright spot which it can be shown that micro blogging platform has enabled companies do brand image, product sales channel, improvised product sales, talk with consumer for a good interaction and various other business related activities are involved [2,3] [14].

[13] said that in fact, the company's manufacturing such products have started to poll these micro blogs to get a sense of general sentiment about a product. Many a times these companies try to study users' reactions and reply to users via micro blogs [14].

D. *Social Media*

[15] has defined social media as the group of web-based applications that are created on the ideological and technological foundations of Web 2.0 and is allowed to build and exchange user generated contents. In a discussion of the Internet World Start, [16] has identified that a trend of world wide web users is increasing and they are continuing to spend more time with social media out of the total time spent on their mobile devices and social media in the U.S.across PC increased by more than 37 percent to 121 billion minutes in 2012, which is far more than 88 billion minutes in 2011. On the other hand, businesses use social networking sites to find and communicate with clients, business can be demonstrated as damage productivity caused by social networking [17]. As social media can be used to post information so easily to the public, it can cause harm to the private information to spread out in the social world [11].

On the contrary, in [18] it is discussed that the benefits of participating in social media platforms have gone beyond simply social information sharing to building an organization's brand reputation and bringing in lots of career opportunities and monetary income sources. In addition, [15], [35] mentioned that social media platforms are also being used for advertisement by various companies for promotion, professionals for searching, recruitment of candidates, learning online and electronic commerce. Electronic-commerce or E-commerce means the purchase and sale of products and services online which can be made possible via social media, such as Twitter which is a lot more convenient due to the 24x7 availability, ease of customer service and also global reach [19].

Among the reasons why business tends to use more social media is for getting insight into consumer behavioral tendencies, market intelligence and present an opportunity to learn about customer review and perceptions.

E. *Twitter Sentiment Analysis*

The sentiment may be found within the comments or tweet to supply useful indicators for several different purposes [20]. Also, [12] and [36] stated that a sentiment may be categorized into two groups, which is negative and positive words. Sentiment analysis could be a linguistic communication processing technique to quantify an expressed opinion or sentiment within a range of tweets [8].

Sentiment analysis refers to the final method to extract polarity and subjectivity from semantic orientation which refers to the strength of words and polarity text or phrases [19]. There has two main approaches for extracting sentiment automatically which are the lexicon-based approach and machine-learning-based approach [19-23].

1. Lexicon-based Approach

Lexicon-based methods make use of predefined list of words where each word is related to a selected sentiment [21]. The lexicon methods vary in keeping with the context in which they were created and involve calculating orientation for a document from the semantic orientation of texts or phrases within the documents [19]. Besides, [24] also states that a lexicon sentiment is to detect word-carrying opinion within the corpus and to predict opinion expressed within the text. [20] has shown the lexicon methods which have a basic paradigm which are:

i. Preprocess each tweet, post by remove punctuation
ii. Initialize a complete polarity score (s) equal 0 -> s=0
iii. Check if token is present in an exceedingly dictionary, then
If token is positive, s are positive (+) If token is negative, s are negative (-)
iv. Look at the full polarity score of tweet post If s > threshold, tweet post as positive If s < threshold, tweet post as negative

However, [21] highlighted one advantage of learning-based method, is that it's the power to adapt and make trained models for specific purposes and contexts. In contrast, an availability of labeled data and hence the low applicability of the method of latest data which is cause labeling data could be costly or perhaps prohibitive for a few tasks [21].

2. Machine-learning-based Approach

Machine learning methods often rely upon supervised classification techniques where sentiment detection is framed as a binary target problem which are positive and negative [24]. This approach requires labeled data to teach classifiers [21]. This approach, it becomes apparent that aspects of the local context of a word must be taken into consideration like negative (e.g. Not beautiful) and intensification (e.g. Very beautiful) [19]. However, [20] showed a basic paradigm for create a feature vector is:

i. Apply a part of speech tagger popularly known as the POS tagger to tweet post
ii. Collect all the adjective for entire tweet posts
iii. Make a preferred word set composed of the best N adjectives
   - Number of positive words
   - Number of negative words
   - Presence, absence or frequency of each word

[19] showed some example of switch negation, negation simply to reverse the polarity of the lexicon: changing beautiful (+3) into not beautiful (-3). More examples: She isn't terrific (6-5=1) but not terrible (-6+5=-1) either.

In this case, the negation of a strongly negative or positive value reflects a mixed perspective which is correctly captured in the shifted value. However, [21] has mentioned the limitation of machine-learning-based approach to be more suitable for Twitter than the lexical based method.

Furthermore, [20] stated that machine learning methods can generate a tough and fast number of the foremost regularly happening popular words which assigned an integer value on behalf of the frequency of the word within the Twitter.

F. ***Techniques of Sentiment Analysis***

The semantic concepts of the entities are extracted from tweets and can be used to measure the final correlation of a group of entities with any given sentiment's polarity [12]. Polarity refers to the most basic form of a sentence, that is if a text or sentence is classified as positive or negative [25]. However, the sentiment analysis model has techniques in assigning polarity values such as:

1. Natural Language Processing (NLP)

The NLP techniques are based on machine learning/deep learning and especially statistical learning which normally uses a general learning algorithm combined with a large sample, i.e. a huge corpus of data such as the Brown Corpus to learn the rules [26]. Sentiment analysis has been a part of Natural Language Processing denoted NLP, at various levels of granularity. Starting from a document level classification [27], it is continued to be handled at the sentence level [28] and furthermore at the phrase level [13]. NLP is a field in computer science which involves making computers derive semantic meaning from the human language and is put as a way of interacting with the real world. It can not only understand human languages but also the systems are designed to interact with humans via human languages.

2. Case-Based Reasoning (CBR)

Case-Based Reasoning (CBR) is a technique that is available to implement sentiment analysis on Texts. CBR is known for its functionality of recalling past successfully solved problems and using these solutions in order to solve the current closely related problems. [25] Identified some of the advantages of using CBR that CBR does not require an explicit domain model and so elicitation becomes a task of gathering care histories and CBR systems can continuously learn by acquiring knowledge as new cases. This and the application of database techniques make the maintenance of large columns of information easier [25].

3. Artificial Neural Network (ANN)

[13] mentioned that Artificial Neural Networks (ANN) or more popularly known as neural networks is a mathematical technique that connects group of artificial neurons stacked in layers. It will process information using the connections approach and activation functions on the value of their learned parameters to computation. ANN is widely used in finding the relationships between input data and output labels or to find hidden patterns in the input data [25].

4. Support Vector Machine (SVM)

Support Vector Machine is used to detect the sentiments of tweets [23]. [10] together with [37] has stated that SVMs are able to extract and analyze and obtain up to 70%-81.3% of accuracy on the test set. [29] has collected training data from three separate Twitter sentiment detection websites which have used some pre-built sentiment lexicons and labeled each of the tweets as either positive or negative. On using SVM and training on these noisy labeled data, they managed to obtain 81.3% in sentiment classification accuracy.

G. *Application Programming Interface(API)*

Alchemy is an API that performs much better than the other APIs in terms of quality and the quantity of extracted entities [14]. As time has passed the Python Twitter Application Programming Interface (API) is created for collecting tweets [30]. This API can automatically calculate the frequency of messages being tweeted again and again every 100 seconds, sort the top 200 messages based on their tweeting frequency, and store them in their designated databases [12]. The Python Twitter API only includes Twitter messages for the past six days, the collected data needs to be stored in different databases [14].

H. *Python*

Python was founded by Guido Van Rossum in Netherland, 1989 which was published in 1991[31]. Python is also an artificial language that's available and solves a computer problem which is providing a straightforward way to write out a solution [31]. [32] Mentioned that Python are often called as a scripting language. Moreover, [32] and [32] also supported that basically Python is also a just description of language because they are often one written and run on many platforms. Additionally, [34] mentioned that Python is also a language that's great for writing a prototype because Python may be a smaller amount of time consuming and dealing prototype provided, in contrast with other programming languages.

Many researchers are saying that Python is efficient, especially for a fancy project, as [33] has mentioned that Python is suitable to start up social networks or media steaming projects which most always are web-based which is driving an infinite amount of data. [34] gave the rationale that because Python can handle and manage the memory used. Python creates a generator that allows an iterative process of things, one item at a time and permit program to grab source data one item at a time to pass each through the whole processing chain.

CHAPTER 4
RESULTS AND DISCUSSIONS

A. Twitter Retrieved

To associate with the Twitter API, a developer needs to agree with terms and conditions of the Twitter platform which has been provided to get authorization of access to data. The output from this process will be saved in a JSON file. The reason is that JSON (JavaScript Object Notation) is a much lighter data-change format which is easier for humans to read and write. Moreover, stated that, JSON is much simpler for machines to generate and to parse. JSON is a text format that is totally language independent, but uses a convention that is known to programmers of the C family of languages, including Python and many other languages. However, output size depends mainly on the time for retrieving tweets from Twitter. Nevertheless, the output of the analyzer will be categorized into two forms, of which one is encoded and the other is not encoded. According to the security issue for accessing data, some of the output will be visible in an ID form such as string ID Sentiment Analysis. The tweets will be assigned with the value of each word, together with categorization into positive and negative words, according to the lexicon dictionary. The result will also be shown in txt, csv and html files.

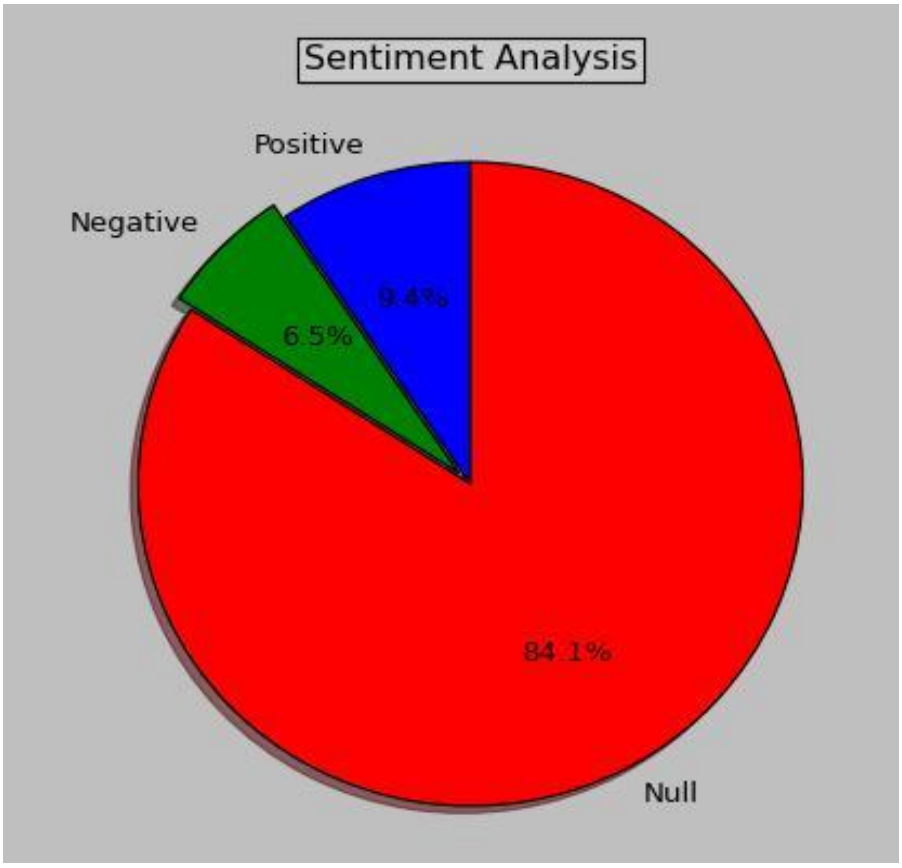

Fig. 1 Pie chart showing percentage of positive, negative and null tags

As shown in Fig. 1, the pie chart represents the percentage of the positive, negative and null sentiment tags in separate colors.

B. Sentiment Analysis

Tweets from the JSON file are assigned with a value for each word by matching from the lexicon dictionary. There is a limitation of words in the lexicon dictionary as it cannot contain all the words in the universe and is not able to assign a value to every single word from tweets. However, as a scientific function of python, which is able to analyze a sentiment of each tweet into positive or negative for getting the result.

C. Information Presented

The result is displayed in the form of a pie chart which is representing a percentage of positive, negative and null sentiment tags. For null tag is representing the hash tags that were assigned zero value. However, this program is able to list the top ten positive and the top ten negative hash tags.

CHAPTER 5
CONCLUSION AND RECOMMENDATION

Twitter sentiment analysis is developed to research customers' perspectives toward the critical to success within the marketplace. The program is employing a machine- based learning approach which is more accurate for analyzing a sentiment; along with tongue processing techniques are used As a result, programs are categorized sentiment into positive and negative, which is represented in a very chart and html page although, the program has been planned to be developed as an online application, thanks to limitations of Django which may only work on Linux server or LAMP. Thus, it can't be realized. Therefore, further enhancement of this element is usually recommended in future study.

REFERENCES

[1] M.Rambocas, and J. Gama, "MarketingResearch:*TheRoleof SentimentAnalysis"*. The 5th SNA-KDD Workshop'11. Universityof Porto, 2013.

[2] A. K. Jose, N. Bhatia, and S. Krishna, "TwitterSentimentAnalysis". *NationalInstituteof TechnologyCalicut*,2010.

[3] P. Lai, "ExtractingStrongSentimentTrendfromTwitter". Stanford University, 2012.

[4] Y. Zhou, and Y. Fan, " A Sociolinguistic Study of American Slang," *Theory and Practice in Language Studies*, *3*(12), 2209–2213, 2013. doi:10.4304/tpls.3.12.2209-2213

[5] M. Comesaña, A. P.Soares, M.Perea, A.P. Piñeiro, I. Fraga, and A. Pinheiro, " Author ' s personal copy Computers in Human Behavior ERP correlates of masked affective priming with emoticons," C*omputers in Human Behavior*, *29*, 588–595, 2013.

[6] A.H.Huang, D.C. Yen, & X. Zhang, "Exploring the effects of emoticons," *Information & Management*, 45(7), 466–473, 2008.

[7] D. Boyd, S. Golder, & G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter," *System Sciences (HICSS), 2010 …*.
Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5428313

[8] T. Carpenter, and T. Way, "Tracking Sentiment Analysis through Twitter,". *ACM computer survey*. Villanova:VillanovaUniversity, 2010.

[9] D. Osimo, and F. Mureddu, "Research Challenge on Opinion Mining and Sentiment Analysis," *Proceeding of the 12th conference of Fruct association*, 2010, United Kingdom.

[10] A. Pak,and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," *Special Issue of International Journal of Computer Application*, France:Universitede Paris-Sud, 2010.

[11] S.Lohmann, M. Burch, H. Schmauder and D. Weiskopf, "Visual Analysis of Microblog Content Using Time-Varying Co-occurrence Highlighting in Tag Clouds," *Annual conference of VISVISUS. Germany:* University of Stuttgart, 2012.

[12] H. Saif, Y.He, and H. Alani, "SemanticSentimentAnalysisof Twitter," *Proceeding of the Workshop on Information Extraction and Entity Analytics on Social Media Data*. United Kingdom: Knowledge Media Institute, 2011.

[13] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R.Passonneau, "Sentiment Analysis of Twitter Data," Annual International Conferences. New York:Columbia University, 2012.

[14] J. Zhang, Y. Qu, J. Cody and Y. Wu, " A case study of Microblogging in the Enterprise: *Use, Value, and Related Issues*," Proceeding of the workshop on Web 2.0., 2010.

[15] G. Kalia, "A Research Paper on Social Madia: *An Innovative Educational Too"*, Vol.1, pp. 43-50, Chitkara University, 2013.

[16] Internet World Start, "Usage and Population Statistic", Retrieved 10 15, 2013 from: http://www.internetworldstats.com/stats.htm

[17] A.M. Kaplan, and M, Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," France: *Paris*, 2010.

[18] Q. Tang, B. Gu, and A.B. Whinston, "Content Contribution in Social Media: *The case of YouTube"*, 2nd conference of social media. Hawaii: *Maui, 2012.*

[19] M.Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, " Lexicon-Based Methods for Sentiment Analysis," *Association for Computational Linguistics*, 2011.

[20] M. Annett, and G. Kondrak, "A Comparison of Sentiment Analysis Techniques: *Polarizing Movie Blogs," Conference on web search and web data mining (WSDM*). University of Alberia: Department of Computing Science, 2009.

[21] P. Goncalves, F. Benevenuto, M. Araujo and M. Cha, "Comparing and Combining Sentiment Analysis Methods", 2013.

[22] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter Sentiment Analysis:*The Good the Bad and theOMG!"*, (Vol.5). International AAAI, 2011.

[23] S. Sharma, "Application of Support Vector Machines for Damage detection in Structure," *Journal of Machine Learning Research, 2008.*

[24] A.Sharma, and S. Dey, "Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis," *Association for the advancement of Artificial Intelligence*, 2012.

[25] J. Spencer and G. Uchyigit, "Sentiment or: Sentiment Analysis of Twitter Data," *Second Joint Conference on Lexicon and Computational Semantics*. Brighton:University of Brighton, 2008.

[26] A. Blom and S. Thorsen, "Automatic Twitter replies with Python," *International conference "Dialog 2012"*.

[27] B. Pang, and L. Lee, "Opinion mining and sentiment analysis," 2nd workshop on making sense of Microposts. Ithaca: Cornell University. Vol.2(1), 2008.

[28] M. Hu, and B. Liu, "Mining and summarizing customer reviews," 2004.

[29] P. Nakov, Z. Kozareva, A. Ritter, S. Rosenthal, V. Stoyanov, T. Wilson, Sem Eval-2013 Task2:*Sentiment AnalysisinTwitter* (Vol.2,pp. 312-320 ,2013.

[29] J. Wu, J., Wang, & L. Liu, "Kernel-Based Method for Automated Walking Patterns Recognition Using Klnematics Data". 5th Workshop on Natural Language Processing. China: *Xi'an Jiaotong University*. 2006.

[30] T. D. Smedt, and W. Daelemans, "Pattern for Python," Proceeding of COLING. Belgium: University of Antwerp, 2012.

[31] A. Sweigart, "Invent your own computer games with Python. 2nd edition, 2012. Retrieved from http://inventwithpython.com/

[32] C. Seberino, "Python. *Faster and easier software development*," Annual Conference. California: *San Diego*, 2012.

[33] A.Lukaszewski, “MySQL for Python. *Integrate the flexibility of Python and the power of MySQL to boost the productivity of your applications,”* UK: *Birningham*. Packt Publishing Ltd, 2010.
[34] V. Nareyko, “Why python is perfect for startups,” Retrieved 01 10, 2014 from: http://opensource.com/business/13/12/why-python-perfect-startups
[35] A. Hawkins, “There is more to becoming a thought leader than giving yourself the title”. Retrieved 10 18, 2013. from: http://www.thesocialmediashow.co.uk/author/admin/
[36] R. Prabowo, and M. Thelwall, “Sentiment Analysis:*A Combined Approach,” International World Wide Web Conference Committee (IW3C2), 2009*. UnitedKingdom:Universityof Wolverhamption.
[37] H. Saif, Y. He and H. Alani, “Alleviating Data Scarcity for Twitter Sentiment Analysis”. Association for Computational Linguistics, 2012.