Twitter Sentiment Analysis

Aliza Sarlan[1], Chayanit Nadam[2], Shuib Basri[3]
Computer Information Science
Universiti Teknologi PETRONAS
Perak, Malaysia
aliza_sarlan@petronas.com.my; chayanit171@gmail.com;shuib_basri@petronas.com.my

***Abstract*—Social media have received more attention nowadays. Public and private opinion about a wide variety of subjects are expressed and spread continually via numerous social media. Twitter is one of the social media that is gaining popularity. Twitter offers organizations a fast and effective way to analyze customers' perspectives toward the critical to success in the market place. Developing a program for sentiment analysis is an approach to be used to computationally measure customers' perceptions. This paper reports on the design of a sentiment analysis, extracting a vast amount of tweets. Prototyping is used in this development. Results classify customers' perspective via tweets into positive and negative, which is represented in a pie chart and html page. However, the program has planned to develop on a web application system, but due to limitation of Django which can be worked on a Linux server or LAMP, for further this approach need to be done.**

Keywords-component; Twitter, sentiment, opinion mining, social media, natural language processing

I. INTRODUCTION

According to [1], millions of people are using social network sites to express their emotions, opinion and disclose about their daily lives. However, people write anything such as social activities or any comment on products. Through the online communities provide an interactive forum where consumers inform and influence others.Moreover, social media provides an opportunity for business that giving a platform to connect with their customers such as social media to advertise or speak directly to customers for connecting with customer's perspective of products and services.

In contrast, consumers have all the power when it comes to what consumers want to see and how consumers respond. With this, the company's success & failure is publicly shared and end up with word of mouth. However, the social network can change the behavior and decision making 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 customer's think, it would be more beneficial to organize to react on time and come up with a good strategy to compete their competitors.

A. Problem Statement

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

- Sentiment Analysis of Web Based Applications Focus on Single Tweet Only.

With the rapid growth of the World Wide Web, people are using social media such as Twitter which generates big volumes of opinion texts in the form of tweets which is available for the sentiment analysis [3]. This translates to a huge volume of information from a human viewpoint which make it difficult to extract a sentences, read them, analyze tweet by tweet, summarize them and organize them into an understandable format in a timely manner [3].

- Difficulty of Sentiment Analysis with inappropriate English

Informal language refers to the use of colloquialisms and slang in communication, employing the conventions of spoken language [4] such as 'would not' and 'wouldn't'. Not all systems are able to detect sentiment from use of informal language and this could hanker the analysis and decision-making process.

Emoticons, are a pictorial representation of human facial expressions [5], which in the absence of body language 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 happy 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 turn to emoticons to properly express what they cannot put into words [6]. Not being able 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 will be used more frequently on Twitter so as to help to minimize the characters used. This is because Twitter has put a limit on its characters to 140 [7]. For example, 'Tba' refers to be announced**.**

B. Objective

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

II. METHODOLOGY

This project has been divided into 2 phases. First, literature study is conducted, followed by system development. Literature study involves conducting studies on various sentiment analysis techniques and method that currently in used. In phase 2, application requirements and functionalities are defined prior to its development. Also, architecture and interface design of the program and how it will interact are also identified. In developing the Twitter Sentiment Analysis application, several tools are utilized, such as Python Shell 2.7.2 and Notepad.

III. LITERATURE REVIEW

A. Opining Mining

Opinion mining refers to the broad area of natural language processing, text mining, computational linguistics, which involves the computational study of sentiments, opinions and emotions expressed in text [8]. Although, view or attitude based on emotion instead of reason is often colloquially referred to as a sentiment [8]. Hence, lending to an equivalent for opinion mining or sentiment analysis.

[9] stated that opinion mining has many application domains including accounting, law, research, entertainment, education, technology, politics, and marketing. In earlier days many social media have given web users avenue for opening up to express and share their thoughts and opinions [10].

B. Twitter

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

Twitter, with 500 million users and million messages per day, has quickly became a valuable asset for organizations to invigilate their reputation and brands by extracting and analyzing the sentiment of the tweets by the public about their products, services market and even about competitors [12]. [2] highlighted that, from the social media generated opinions with the mammoth growth of the world wide web, super volumes of opinion texts in the form of tweets, reviews, blogs or any discussion groups and forums are available for analysis, thus making the world wide web the fastest, most comprising and easily accessible medium for sentiment analysis.

C. Microblogging with E-commerce

A microblogging platform such as Twitter is alike to a conventional blogging platform just single posts are shorter [13]. Twitter has limited for a small number of words which are designed for the quick transmission of information or exchange of opinion [7]. However, small business or large organizations are initiation to the potential of microblogging as an e-commerce marketing tool [3]. Though, microblogging platform has been developed a few years' time for promoting foreign trade website by using a foreign microblogging 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 microblogging platform has enabled companies do brand image, product important sales channel, improve product sales, talk to consumer for a good interaction and other business activities involved [2,3] [14]. [13] said, in fact, the companies manufacturing such products have started to poll theses microblogs to get a sense of general sentiment for a product. Many times these companies study user reactions and reply to users on microblogs [14].

D. Social Media

[15] defined a social media as a group of Internet-based applications that create on the ideological and technological foundations of Web2.0 which is allowed to build and exchange of user generated contents. In a discussion of Internet World Start, 16] identified that a trend of internet users is increasing and continuing to spend more time with social media by the total time spent on mobile devices and social media in the U.S.across PC increased by 37percent to121billion minutes in 2012, compared to 88 billion minutes in 2011. On the other hand, businesses use social networking sites to find and communicate with clients, business can be demonstrated damage to productivity caused by social networking [17]. As social media can be posted so easily to the public, it can harm private information to spread out in the social world [11].

On the contrary, [18] discussed that the benefits of participating in social media have gone beyond simply social sharing to build organization's reputation and bring in career opportunities and monetary income. In addition, [15], [35] mentioned that the social media is also being used for advertisement by companies for promotions, professionals for searching, recruiting, social learning online and electronic commerce. Electronic commerce or E-commerce refers to the purchase and sale of goods or services online which can via social media, such has Twitter which is convenient due to its 24-hours availability, ease of customer service and global reach [19].

Among the reasons of 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 can be found in the comments or tweet to provide useful indicators for many different purposes [20]. Also, [12] and [36] stated that a sentiment can be categorized into two groups, which is negative and positive words. Sentiment analysis is a natural language processing techniques to quantify an expressed opinion or sentiment within a selection of tweets [8].

Sentiment analysis refers to the general 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 associated with a specific sentiment [21]. The lexicon methods vary according to the context in which they were created and involve calculating orientation for a document from the semantic orientation of texts or phrases in the documents [19]. Besides, [24] also states that a lexicon sentiment is to detect word-carrying opinion in the corpus and then to predict opinion expressed in 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 total polarity score (s) equal 0 -> s=0
iii. Check if token is present in a dictionary, then

If token is positive, s will be positive (+)

If token is negative, s will be negative (-)

iv. Look at the total 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 leaning-based method, is that it has the ability to adapt and create trained models for specific purposes and contexts. In contrast, an availability of labeled data and hence the low applicability of the method of new data which is cause labeling data might be costly or even prohibitive for some tasks [21].

2. Machine-learning-based Approach

Machine learning methods often rely on supervised classification approaches where sentiment detection is framed as a binary which are positive and negative [24]. This approach requires labeled data to train classifiers [21]. This approach, it becomes apparent that aspects of the local context of a word need to be taken into account such as 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 to each tweet post
ii. Collect all the adjective for entire tweet posts
iii. Make a popular word set composed of the top N adjectives
iv. Navigate all of the tweets in the experimental set to create the following:
- 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 is not 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 fixed number of the most regularly happening popular words which assigned an integer value on behalf of the frequency of the word in the Twitter.

F. Techniques of Sentiment Analysis

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

1. Natural Language Processing (NLP)

NLP techniques are based on machine learning and especially statistical learning which uses a general learning algorithm combined with a large sample, a corpus, of data to learn the rules [26]. Sentiment analysis has been handled as a Natural Language Processing denoted NLP, at many levels of granularity. Starting from being a document level classification task [27], it has been handled at the sentence level [28] and more recently at the phrase level [13]. NLP is a field in computer science which involves making computers derive meaning from human language and input as a way of interacting with the real world.

2. Case-Based Reasoning (CBR)

Case-Based Reasoning (CBR) is one of the techniques available to implement sentiment analysis. CBR is known by recalling the past successfully solved problems and use the same solutions to solve the current closely related problems [29]. [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 system can learn by acquiring new knowledge as 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 Network (ANN) or known as neural network is a mathematical technique that interconnects group of artificial neurons. It will process information using the connections approach to computation. ANN is used in finding the relationship between input and output or to find patterns in data[25].

4. Support Vector Machine(SVM)

Support Vector Machine is to detect the sentiments of tweets [23]. [10] together with [37] stated SVM is able to extract and analyze to obtain upto70%-81.3% of accuracy on the test set. [29] collected training data from three different Twitter sentiment detection websites which mainly use some pre-built sentiment lexicons to label each tweet as positive or negative. Using SVM trained from these noisy labeled data, they obtained 81.3% in sentiment classification accuracy.

G. *Application Programming Interface(API)*

Alchemy API performs better than the others in terms of the quality and the quantity of the extracted entities [14]. As time passed the PythonTwitter Application Programming Interface (API) is created by collected tweets [30]. Python can automatically calculated frequency of messages being re-tweeted every 100 seconds, sorted the top 200 messages based on there-tweeting frequency, and stored them in the designated database [12]. As the Python Twitter API only included Twitter messages for the most recent six days, collected the data needed to be stored in a different database [14].

H. *Python*

Python was found by Guido Van Rossum in Natherland, 1989 which has been public in 1991[31]. Python is a programming language that's available and solves a computer problem which is providing a simple way to write out a solution [31]. [32] mentioned that Python can be called as a scripting language. Moreover, [32] and [32] also supported that actually Python is a just description of language because it can be one written and run on many platforms. In addition, [34] mentioned that Python is a language that is great for writing a prototype because Python is less time consuming and working prototype provided, contrast with other programming languages.

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

IIV. RESULT AND DISCUSSION

A. *Twitter Retrieved*

To associate with Twitter API, developer need to agree in terms and conditions of development Twitter platform which has been provided to get an authorization to access a data. The output from this process will be saved in JSON file. The reason is, JSON (JavaScript Object Notation) is a lightweight data-interchange format which is easy for humans to write and read. Moreover, stated that, JSON is simple for machines to generate and 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 others. However, output□s size depends on the time for retrieving tweets from Twitter.

Nevertheless, the output will be categorized into 2 forms, which are encoded and un-encoded. According to security issue for accessing a data, some of the output will be shown in an ID form such as string ID. Sentiment Analysis. The tweets will be assigned the value of each word, together with categorize into positive and negative word, according to lexicon dictionary. The result will be shown in .txt, .csv and html.

B. *Sentiment Analysis*

Tweets from JSON file will be assigned the value of each word by matching with the lexicon dictionary. As a limitation of words in the lexicon dictionary which is not able to assign a value to every single word from tweets. However, as a scientific language of python, which is able to analyze a sense of each tweet into positive or negative for getting a result.

C. *Information Presented*

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

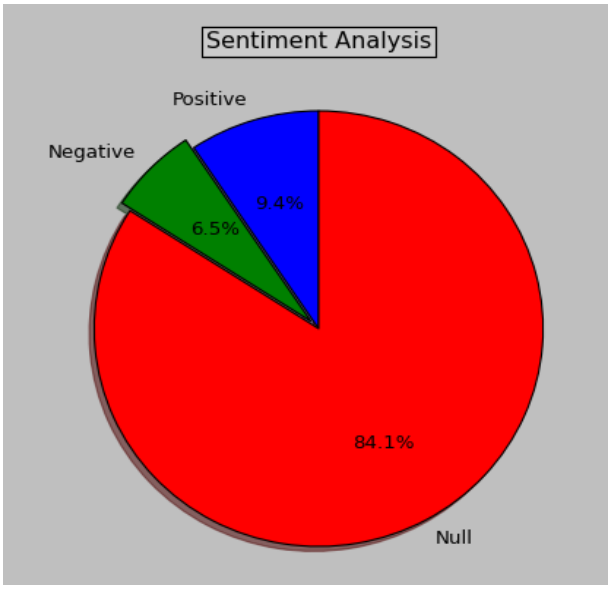

Fig. 1. Pie Chart

As shown in Fig. 1, the pie chart is representing of each percentage positive, negative and null sentiment hash tags in different color.

IIIV CONCLUSION AND RECOMMENDATION

Twitter sentiment analysis is developed to analyze customers perspectives toward the critical to success in the marketplace. The program is using a machine-based learning approach which is more accurate for analyzing a sentiment; together with natural language processing techniques will be used.

As a result, program will be categorized sentiment into positive and negative, which is represented in a pie chart and html page Although, the program has been planned to be developed as a web application, due to limitation of Django which can only work on Linux server or LAMP. Thus, it cannot be realized. Therefore, further enhancement of this element is 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.