Using Social Media Data as Research Data

Suman Silwal
Interdisciplinary Engineering Ph.D. Student
The University of Alabama at Birmingham
ssilwal@uab.edu

Advisor: Dale W. Callahan, Ph.D., P.E.
Director - Information Engineering and Management
The University of Alabama at Birmingham School of Engineering
dcallahan@uab.edu

Abstract

Social Media (SM) is becoming a normal part of everyday life. The information generated from Social Media (SM) data is becoming increasingly utilized as a communication channel for market trend, brand awareness, breaking news, and online social interaction between person to person. SM is also rapidly growing and maturing [1]. Further, SM is becoming a reliable tool for interdisciplinary industries like banks, travel, healthcare, biotech, software, sports etc.

SM data can also be used as a research tool to apply in different areas of Humanities, Art, Science and Engineering. There are unlimited possibilities using Social Networking Site (SNS) to collect, process and evaluate data. This paper reviews the current state of Social Networking Sites and Text-based Language Processes, and how it can be used to generate valuable information.

Key words: *Social Media, Social Network Site, Natural Process Language*

1. Introduction

Human communication via speech and symbols date back more than 30,000 years [2]. With the birth of the internet (which was originally created as a small government and university research tool) in the public domain in 1995, the way humans communicate has changed [3][4]. Even though Social Networking Sites (SNS) started in the 1990's, SNS did not become a mainstream medium of communication until 2000 [5].

Today in 2013, Social Media (SM) is becoming a norm of interaction between businesses and their customers/fans as well as person to person communication. Prior to 2008, web presence was an essential part of a business strategy; now, SM is taking over as one of the additional factors for businesses to succeed [6]. There are many SNS that serve different demographics and interests; Facebook (facebook.com), Twitter (twitter.com), and LinkedIn (linkedIn.com) are taking the lead with more than 2.2 billion registered users combined [7].

Every day, active users on these Social Networking Sites generate millions of posts and updates. In this new era of Social Media, information generated by active SM users can be used as a research tool to identify current trends, brand awareness, marketing campaign success, disease outbreaks, breaking news, and much more. Depending on the privacy rules on each social network, information can be abstracted to generate useful knowledge for everyone to review and understand.

Even though there is a vast amount of interest and enthusiasm about SM, much research and product development are done in the area of marketing and advertising. This SM data give market researchers great opportunities to find people's interests, products they like and use, current trends, etc.

Shared SM information can be used in many different areas: discovering the outbreak of a disease in any corner of the world, finding a solution to a complex problem, providing a source of information during natural disasters, or updating real time on local or global events. There are endless possibilities with regards to how SM information can be used.

2. Social Network Sites Data

Every day, active users on Social Networking Sites (SNS) generate millions of posts. SNS data are growing exponentially and are being used to identify current trends, brand awareness, marketing campaign success, disease outbreaks, breaking news, etc.

More than 200 SNS sites listed on the Wikipedia.org [7]. Each of these sites has its own unique SM presence with its own list of unique users. According to eMarket.com, nearly one in four people are using some kind of Social Networking Site around the world [8]. This is a growing trend. More and more people are using SM to get their current news and update from friends and families around the world.

Table 1. Top ten most popular Social Networking Sites [9]

Social Networking Site (SNS)	Estimated Unique Monthly Visitors
Facebook.com	750,000,000
Twitter.com	250,000,000
LinkedIn.com	110,000,000
Pinterest.com	85,500,000
MySpace.com	70,500,000
Google Plus	65,000,000
DeviantArt.com	25,500,000
LiveJournal.com	20,500,000
Tagged.com	19,500,000
Orkut.com	17,500,000
Total	**1,414,000,000**

Last updated date for above table data was 7/24/2013

Table 1 provides only a fraction of the Social Networking Sites that exist with estimated total of 1,414,000,000 unique monthly visitors. Even though these numbers change from month to month, it is a growing trend that Social Media is becoming an acceptable form of daily communication around the world.

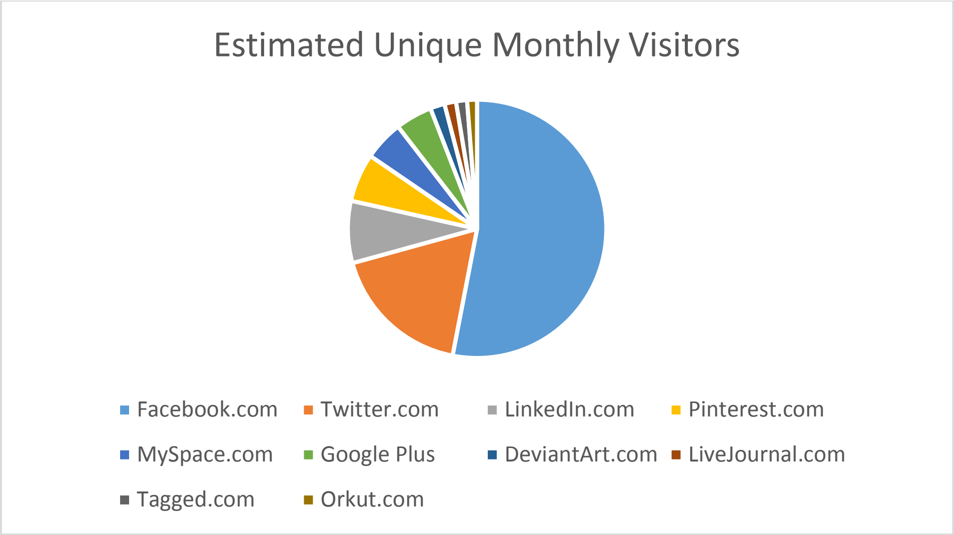

Figure 1. Data from Table 1

Facebook and Twitter are leading SNS. According to statisticbrain.com, there are 70 billion shared posts on Facebook monthly and an average of 190 million tweets daily [10]. These are large amount of monthly data generated by only two Social Networking Sites. Reviewing the top 10 SNS from table 1, we can forecast that more than 200 SNS contributes billions of daily data.

People are expressing their thoughts and sentiments in real time in SM. These SM data can provide a wealth of research materials for business users as well as university researchers. Filtering, validating, and capturing useful information from unstructured SM data are always going to be a challenge.

Due to the rapid change of SM data, real time data analyses are vital in getting valid information [11] to review users' sentiments. Text-based Natural Language Process (NLP) can play an important role in analyzing these SM data. In the next chapter we will discuss the NLP.

3. Natural Language Processing (NLP)

NLP is described as a computer system which processes human language in the context of its meaning [12]. Even with the advancement of computer languages and artificial intelligence, humans and computers do not talk the same language. Computer systems use byte-code.
Table 2 provides a list of NLP toolkits with a description of each and the implementation architect used. Each of these toolkits provides a different option to retrieve and process textual data.

Table 2. Natural Language Processing Toolkits

Name	Description	Implementation Architect Based On	URLs
LingPipe	Processes text using computational linguistics. It automatically classifies Twitter search results into categories.	Java	http://alias-i.com/lingpipe/index.html
Apache OpenNLP	Performs tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and co-reference resolution	Java	http://opennlp.apache.org/
Stanford Parser and Part-of-Speech (POS) Tagger	Reads text in some languages and assigns parts of speech to each word (and other token), such as nouns, verbs, adjectives, etc.	Java	http://nlp.stanford.edu/software/tagger.s html
OpenFst	Keys applications in speech recognition and synthesis, machine translation, optical character recognition, pattern matching, string processing, machine learning, information extraction and retrieval, among others.	C++	http://www.openfst.org/
Natural Language Toolkit (NLTK)	Works in computational linguistics using Python.	Python	http://nltk.org/
Opinion Finder	Processes documents and automatically identifies subjective sentences as well as various aspects of subjectivity within sentences, including agents who are sources of opinions, direct subjective expressions, speech events, and sentiment expressions.	Java	http://mpqa.cs.pitt.edu/opinionfinder/
GATE	Uses for all types of computational tasks involving human language.	Java	http://gate.ac.uk/
NLP Toolsuite	Collections of NLP components.	Java	http://www.julielab.de/Resources/Software/NLP_Tools.html

NLP can play an important role in understanding SM users' sentiments. SM text-based posts can be processed using NLP to get positive, negative, and natural feedback. This feedback can be used to further process these data.

4. Using Social Media data as research data

Starting fall 2013, Nielsen, a leading global information and measurement company that provides market research, started to use Twitter SM data to complement rating systems that exist today [13]. Nielsen purchased SocialGuide.com, whose APIs are focused on the Twitter data on TV viewing. It mainly uses hashtags (#) search and retweets to see how many people are actually talking about a given show in a given period of time.

4.1 Current Social Media Analysis Model

Figure 2 shows how most of the SM analysis are done. In this model, data are filtered and evaluated according to hashtag (#), mention, follow and/or followers information. It provides a lot of information about current trends, popularity of person or subject, breaking news, etc.

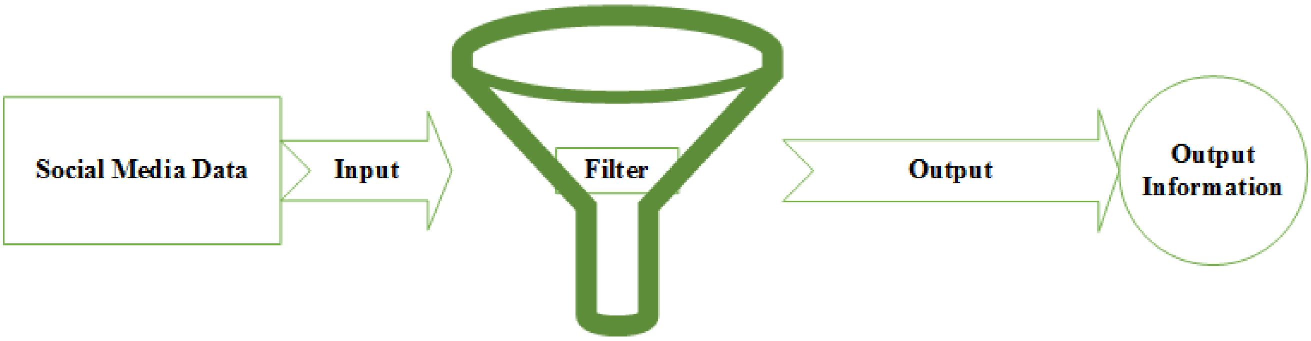

Figure 2. Social Media data process

Even though most Social Media analyses are done in a real time manner, they fail to provide a deeper look into users' sentiment. As researchers, we are missing out on valuable information. To understand the true meaning beyond Hashtag (#) and mentions, we need to further analyze these SM data using other processes.

4.2 Developing Social Media Users' Sentiments Model

Understanding users' sentiment from unstructured Social Media data provides its unique challenges. Some SNS like Twitter only allow 140 characters for a person to express his/her thoughts and sentiments. There are multiple factors involved in outputting useful information to generate a Sentiments Model by using these SM data. Section 3 provides a listing of Natural Language Processing Toolkits. NLP can be used to process users' sentiments.

Figure 3 is showing a recommended input/output Users' Sentiments Model, which can process Social Media data. Once data is filtered, it is sent to the model for further processing. Inside the model, SM data will be processed using NLP and/or some other Text-based processing to understand users' sentiments. Those sentiments will be analyzed, evaluated and processed to get some useful information. Once the information is ready, it will be sent to the output system.

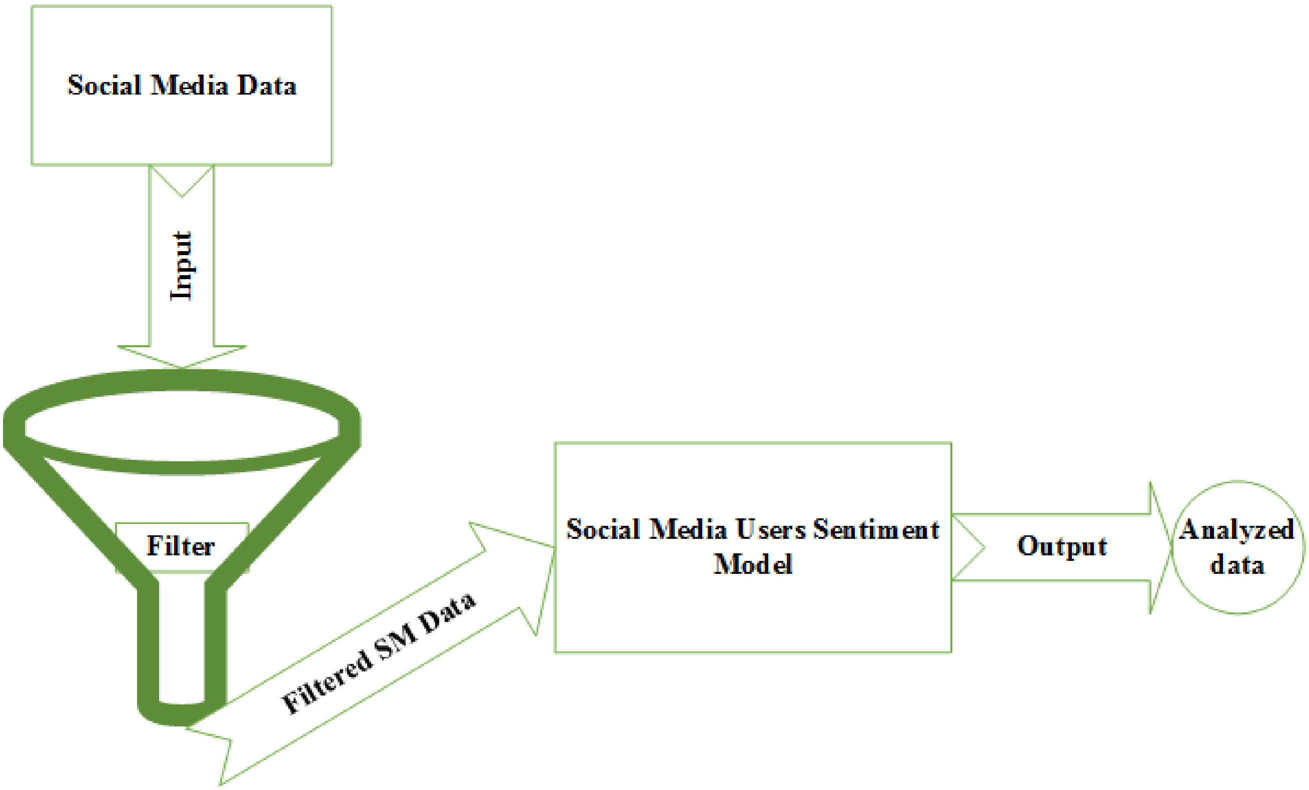

Figure 3. Users' sentiment model

In this model, most of the work is done at the Social Media Users' Sentiment Model stage. Using NLP sentiment analysis is just the first phase of the model development. Even in this initial stage of development of the SM Users Sentiment Model, we see a great potential for a wider variety of uses for interdisciplinary industries.

5. Conclusion

Over the last several years, SNS have been growing rapidly. Businesses have been paying close attention to the growth of the SM boom and the opportunities that this is providing them. This growth is hard to ignore. The active users' participation with and contribution to SNS' data gives researchers like us untapped resources that can be used for finding solutions to complex problems.

Even though understanding a true user sentiment on unstructured data still provides immense challenges, a new way of analyzing SM users' sentiment goes beyond the current state of SM data analysis. It also provides a great opportunity for our research topic.

5. Future works

In future research works, we will extract SM data and develop a Social Media Users' Sentiment Model to process those SM data.

6. References

[1] "How is Social Media Maturing? | International Meetings Review." [Online]. Available: http://www.internationalmeetingsreview.com/technology/how-social-media-maturing-95698. [Accessed: 13-Jul-2013].

[2] *History of communication*. http://en.wikipedia.org/wiki/History_of_communication.

[3] B. M. Leiner, D. D. Clark, R. E. Kahn, L. Kleinrock, D. C. Lynch, J. Postel, L. G. Roberts, and S. Wolff, "A Brief History of the Internet Professor of Computer Science," vol. 39, no. 5, pp. 22–31, 2009.

[4] T. R. Tyler, "Is the Internet Changing Social Life? It Seems the More Things Change, the More They Stay the Same," *J. Soc. Issues*, vol. 58, no. 1, pp. 195–205, Jan. 2002.

[5] S. Edosomwan and S. Prakasan, "The history of social media and its impact on business," *J. Appl. …*, 2011.

[6] C. K. Reid, "Should Business Embrace Social Networking?," *EContent*, 2009. [Online]. Available: http://www.econtentmag.com/Articles/ArticleReader.aspx?ArticleID=54518&PageNum=1.

[7] "List of social networking websites." [Online]. Available: http://en.wikipedia.org/wiki/Social_networking_websites.

[8] "Social Networking Reaches Nearly One in Four Around the World - eMarketer." [Online]. Available: http://www.emarketer.com/Article/Social-Networking-Reaches-Nearly-One-Four-Around-World/1009976. [Accessed: 28-Aug-2013].

[9] "Top 15 Most Popular Social Networking Sites." [Online]. Available: http://www.ebizmba.com/articles/social-networking-websites.

[10] "Social Networking Statistics | Statistic Brain." [Online]. Available: http://www.statisticbrain.com/social-networking-statistics/.

[11] P. Song, A. Shu, and A. Zhou, "A pointillism approach for natural language processing of social media," *arXiv Prepr. arXiv …*, 2012.

[12] J. Rehling, "How Natural Language Processing Helps Uncover Social Media Sentiment." [Online]. Available: http://mashable.com/2011/11/08/natural-language-processing-social-media/. [Accessed: 16-Jul-2013].

[13] "How Nielsen Is Using Twitter for Smarter TV Ratings - The Social Media Monthly." [Online]. Available: http://thesocialmediamonthly.com/how-nielsen-is-using-twitter-for-smarter-tv-ratings/.