2015

Developing A Rating Model Using Social Media Data

Suman Silwal
University of Alabama at Birmingham

DEVELOPING A RATING MODEL USING SOCIAL MEDIA DATA

by

SUMAN SILWAL

DR. DALE W. CALLAHAN, COMMITTEE CHAIR
DR. OLIVIA AFFUSO
DR. ALLEN C. JOHNSTON
DR. ROY P. KOOMULLIL
DR. MURAT M. TANIK

A DISSERTATION

Submitted to the graduate faculty of The University of Alabama at Birmingham,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

BIRMINGHAM, ALABAMA

2015

DEVELOPING A RATING MODEL
USING SOCIAL MEDIA'S DATA

SUMAN SILWAL

PHD IN INTERDISCIPLINARY ENGINEERING

ABSTRACT

Social Media (SM) is becoming a normal part of everyday life for many people around the world. This new form of communication has helped to close social gaps and bring the world closer. The information generated from Social Networking Sites (SNS) is increasingly utilized as a communication channel for market trend, brand awareness, breaking news, person-to-person online social interaction, etc. As more and more people are using SNS, their reach and power are growing rapidly in daily life.

Massive amounts of daily data generated from SNS can be used in many interdisciplinary areas of research such as the Humanities, Art, Science, Engineering, Sports, etc. SM data is readily available through SNS's Application Programming Interface (API). Many SNS provide deeper statistical information to further this research into SM data.

In the recent years, events have been continuously discussed on SM in the form of status updates, posts, discussions and comments by its participants, volunteers, and supporters. SM content generated before, during, and after an event could add valuable insight into the success, popularity, ideas for future improvement of the event, etc. With the fast evolving nature of SM, current events' SM content is ignored, forgotten, and overlooked for new sets of future posts, discussions, and comments.

This dissertation research demonstrates that any publically available SM data can be captured and analyzed to produce a numeric rating for an event such as a marathon. As a result, a rating model was created through combinations of multiple models using SM data to rate an event.

Key words: Social Media (SM), Social Networking Site (SNS), Rating System, Sentiment Analysis, Marathon, Twitter, Hashtag (#), Rating Model, Rating, Tweet

DEDICATION

This dissertation research is dedicated to my mother, Jhuma Silwal, and my late father, Bhaktendar Dhoj Silwal. You have always been a source of great strength, dedication, and inspiration to me.

ACKNOWLEDGMENTS

I would like to sincerely thank the members of my graduate committee: Dr. Dale Callahan, Dr. Olivia Affuso, Dr. Allen Johnston, Dr. Roy Koomullil, and Dr. Murat Tanik. Your guidance and support have been invaluable. I would especially like to thank my committee chair and mentor, Dr. Dale Callahan. Your encouragement, constructive criticism, patience, guidance, and continued mentoring have allowed me to achieve more than I could have envisioned in my educational and professional journey. I would like to thank Jim Merrell and everyone at BlueCross and BlueShield of Alabama for your continued support through my education journey. I would like to thank Dr. David Littlefield, Heather Creel, Sherrye Watson, and the Mechanical Engineering department for making my Ph.D. journey smooth and possible. Thanks to Craig Beard for guiding me in finding research papers through different sources. Thanks to my TeamMRuns and all my running friends for their continued support since the beginning of my Ph.D. journey. Also, thanks to Robert Moore, Ashely Dowson, Matthew Whipple, David Graves, Sunil Silwal, Travis Engram, and others for evaluating and rating marathon events' sample tweets and words. Thanks to Jessica Bonner for proofreading and editing my papers. I also want to thank Dr. Bharat Soni for accepting me into UAB's Interdisciplinary Engineering Ph.D. program. Further, I thank Dr. Sanjay Singh for being a mentor and guide as well as a friend throughout my education journey at UAB. To my mother-in-law and father-in-law, Rosa and Edgar Alvarez, thanks for your support. Thanks to my brother Sunil Silwal, to my sisters Sunita Baskota and Sushma Chhetry for your continued support and encouragement of me to achieve this goal. To my two wonderful daughters, Isabel and Abigail, thank you for being patient, for being my source

of motivation, and for bringing the reality to my journey. Most importantly, I want to thank my wife, Marlene. Your love, patience, trust, and confidence in me made this journey possible.

Lastly, I would like to thank everyone who touched my life directly and indirectly to help me to cross this Ph.D. finish line. Without everyone's support, love, patience, help, confidence, motivation, and much more, I would not have completed this journey that started in the fall of 1989, when I left my home country of Nepal. Now that I'm looking back, I have so much to be thankful for.

TABLE OF CONTENTS

Page

LIST OF TABLES

CAN AN EVENT BE RATED USING HASHTAGS, EMOTICONS, IMAGES, AND URLs? RATING AN EVENT USING SOCIAL MEDIA DATA

LIST OF FIGURES

CAN AN EVENT BE RATED USING HASHTAGS, EMOTICONS, IMAGES, AND URLs? RATING AN EVENT USING SOCIAL MEDIA DATA

CONCLUSION

LIST OF ABBREVIATIONS

SM	Social Media
SNS	Social Networking Sites
NLP	Natural Language Processing
AI	Artificial Intelligence
JSON	JavaScript Object Notation
AJAX	Asynchronous JavaScript and XML
API	Application Programming Interface
RD	Relation Database
App	Application
#	Hashtag
URL	Uniform Resource Locator
P	Positive
N	Negative
NU	Neutral
NA	Not Applicable
B	Both
BK	Backward
F	Forward

INTRODUCTION

Human communication via speech and symbols date back more than 30,000 years[1]. 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 [1][2]. Even though Social Networking Sites (SNS) started in the 1990's, SNS did not become a mainstream medium of communication until 2000 [3].

In 2015, Social Media (SM) is becoming a norm of interaction between businesses and their customers/fans as well as a norm of person-to-person communication. Prior to 2008, the 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 [4]. 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[2].

Every day, active users on these SNS generate millions of posts and updates. In this new era of SM, the information generated by these active SM users can be used as a research tool to identify current trends, brand awareness, marketing campaign success, disease outbreaks, breaking news, etc. Depending on the privacy rules on each SNS, the 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 surrounding SM, much research and product development are conducted in the area of marketing and advertising. SM data give market researchers great opportunities to find people's

[1] http://en.wikipedia.org/wiki/History_of_communication
[2] http://en.wikipedia.org/wiki/Social_networking_websites

interests, products they like and use, current trends, etc. Shared SM information can be used for many different purposes, such as discovering the outbreak of a disease in any corner of the world, finding a solution to a complex problem, providing a source of information, connecting people and resources during natural disasters such as the April 25, 2015 earthquake in Nepal, or updating real time on local or global events. There are endless possibilities with regard to how SM can be used.

A single SM post can consist of words, word abbreviations, numbers, Hashtags, images, mentions, links, symbols, emoticons, etc. Furthermore, each piece of this information can provide insight into understanding the overall sentiment of a SM post. Therefore, all the aspects of a single SM post are candidates for an overall sentiment analysis evaluation of an event. Multiple modeling techniques are being built to capture every single aspect of SM posts to achieve the overall goal of creating a rating system using SM data.

Common Terms

In this dissertation, there are several terms that occur regularly. It is important for readers to understand the meaning of each term:

I. **Social Networking Sites** (SNS) refers to one or more websites where users can create a public profile and interact with other users within the same website[3].

II. **Social Media** (SM) refers to Web 2.0 technology with multiple actors contributing to Social Network Sites (SNS) where people communicate with each other [5][6].

[3] http://www.webopedia.com/TERM/S/social_networking_site.html

III. **Marathon** refers to a 26.2-mile foot racing event, mostly occurring on roads. While long races were a part of ancient Greek competitions, contemporary marathons trace their origin to 1896, when it was one of the original Olympic events[4]. Now, there are many marathon events worldwide. The USA had 570 officially listed marathons as of 2011[5].

IV. **Twitter** is a Social Network Site (SNS) where micro-bloggers are allowed to use up to 140 characters to express their thoughts.

Relation Database (RD)

A Relation Database (RD) is a collection of tables and data using the relation model. Most modern databases are RDs. There are multiple commercial as well as multiple open source RDs. The MySQL[6] database is one of the most widely used open source RDs.

The World Marathon Majors[7]

The World Marathon Majors is a series consisting of six of the largest and most renowned marathons in the world. The cities involved are Tokyo (Japan), Boston (USA), London (UK), Berlin (Germany), Chicago (USA), and New York City (USA). Table 1 lists all 6 marathon majors and their social network presence. In a growing trend, these marathon majors are continuously discussed on their SNS.

Table 1 SM presence on the World Marathon Majors

Marathon Name	Facebook Page Like	Twitter Followers	Total Entry
Boston Marathon [8]	120,346	60,692	27,000
Chicago Marathon[9]	63,591	13, 299	45,000
Berlin Marathon[10]	26,883	4,912	40,000

[4] http://en.wikipedia.org/wiki/Marathon
[5] http://www.statisticbrain.com/marathon-running-statistics/
[6] http://www.mysql.com/
[7] http://worldmarathonmajors.com/
[8] http://www.baa.org/
[9] http://www.chicagomarathon.com/
[10] http://www.bmw-berlin-marathon.com/en/

London Marathon[11]	56,689	42,994	35,700
New York City Marathon[12]	85,121	39,870	47,000
Tokyo Marathon[13]	12,556	24,894	36,000

METHOD

In this section, the method and technology that will be used as a part of this research will be discussed.

Social Network Site (SNS)

At this time, this research is focusing on the Twitter[14] platform to retrieve and process data. Even with 140 characters and the unstructured nature of the data, Twitter is the Social Media of choice for this research because of its easy uses, access to its APIs, the volume of daily data (table 2), its popularity, etc.

Table 2 Twitter statistics [15]

Twitter statistics	Data
Total number of registered users	645,750,000
Number of new Twitter's users signing up every day	135,000
Average number of tweets per day	58 million
Number of Twitter search engine queries every day	2.1 billion
Number of active Twitter users every month	115 million
Number of tweets every second	9,100

(Data posted date: 03/25/2015)

Twitter is built on the idea of an open source project[16]. Through a developer network, Twitter gives low latency access to its data using a specific set of Twitter

[11] http://www.virginmoneylondonmarathon.com/
[12] http://www.ingnycmarathon.org/
[13] http://www.tokyo42195.org/2014/
[14] http://www.twitter.com
[15] http://www.statisticbrain.com/social-networking-statistics
[16] https://dev.twitter.com/opensource

APIs[17]. Twitter provides powerful search access to its data in real time, access that is going to be important for this research.

Here are some of the common terms used on Twitter:

i. Tweets: Twitter user posts.

ii. Follower: A Twitter user following another Twitter user.

iii. Following: A Twitter user following another Twitter user.

iv. Retweet: A Twitter post re-posted by another Twitter user or the original poster.

Application Programming Interface (API)

An application programming interface (API) specifies how some software components should interact with each other[18]. All of the leading SNS such as Facebook, Twitter, and Google+ provide APIs to connect to its network through their developer's network. Each network has its own rules on how a developer can access its network.

a. *Java API*

Java programming language is an open-source computer programming language that is widely used for application development. It has many different APIs to help to connect to databases as well as Social Media.

b. *Spring Java framework API*

The Spring Java framework API provides a lot of different components to build a very powerful application. It glues together with other application frameworks to build an application.

[17] https://dev.twitter.com/docs/streaming-apis
[18] http://en.wikipedia.org/wiki/Application_programming_interface

c. *Twitter's SNS connecting APIs*

Twitter has its own connection and access API (table 3). For this research, Spring Social API was selected to connect to, search, and retrieve Twitter's data.

Table 3 Twitter API libraries

API Library Name	Library Website	API library based
Twitcurl	https://code.google.com/p/twitcurl/	C++
MonkehTweet	http://monkehtweet.riaforge.org/	ColdFusion
LINQ2Twitter	http://linqtotwitter.codeplex.com/	.Net
Tweetsharp	https://github.com/danielcrenna/tweetsharp	.Net
Twitter4J	http://twitter4j.org/en/index.html	Java
Spring Social	http://www.springsource.org/spring-social	Java
STTwitter	https://github.com/nst/STTwitter	Objective-C
FHSTwitterEngine	https://github.com/fhsjaagshs/FHSTwitterEngine	Objective-C
tmhOAuth	https://github.com/themattharris/tmhOAuth	PHP
Tweepy	https://github.com/tweepy/tweepy	Python

d. MySQL database

The MySQL system is a widely used open-source relational database management system (RDBMS), which provides different tools to access and manage the database. It is used as the primary database to warehouse Twitter data as well as many other components needed for this research.

MOTIVATION

Humans have always used word-of-mouth to express sentiment toward products, services, and events. Since the birth of the Internet in the public domain, websites are also used as word-of-mouth tools. In recent years, SM has taken over as social communication channels as well as word-of-mouth tools [7]. People are expressing more and more of their thoughts, ideas, sentiments, and recommendations through SM posts.

The marathon running sport is also a growing sport around the world. In any given marathon race, one can find elite runners to everyday runners competing in it. The World Marathon Majors, listed in Table 2, is a fraction of marathons listed around the world. According to statisticbrain.com, the USA alone has 570 listed marathons with 551,811 finishers. Further, as of 2013, the six World Marathon Majors have more than 230,700 registered runners for these events.

More and more marathon organizers are utilizing SM to communicate up-to-date information with their participants. From the start of a marathon training cycle until the crossing of the finish line and beyond, marathon participants are utilizing SM to talk about their experiences and sentiments toward different aspects of a race. However, as runners are moved from event to event, information posted by these participants are not captured, analyzed, rated, and posted for future use. With the fast pace of SM data generation, previous SM posts are lost in piles of new SM posts.

Even with the growth of marathon running events around the world, the area of marathon running still lacks any type of comprehensive rating system. Regardless, these marathon events are continuously discussed on SM in the form of status updates, posts, and comments by participants, volunteers, and supporters.

This dissertation research is motivated to create a unique way to rate a marathon event using publically available SM data. In this rating model, SM data are imported, analyzed, and evaluated to produce a numeric rating. There is immense value in the information generated by the highly dedicated running community, which is very passionate about what it does. The eventual output of the rating model can be also

valuable to event organizers and sponsors in evaluating marathon/race outcome, popularity, and input for future improvements.

For initial research, only marathons are used as a research topic and Twitter as a SM tool. Future models could expand beyond the initial research topic and tool. This area of rating an event using SM data is still a new field with unlimited research possibilities.

OBJECTIVE

The main objective of this research is to design and develop a rating model using SM data to rate an event. As a part of developing a model, the following steps will be taken:

1. Review the current state of user rating systems, SNS, and Text-based Natural Language Processing.
2. Develop a rating model using Twitter data to generate a numeric rating.
3. Compare and contrast the rating model against manual rating.

BACKGROUND

In the spring of 2012, the earlier research into the importance of SM was started during the University of Alabama at Birmingham's (UAB) EGR 796 - Interdisciplinary Engineering Journal Club class. During that time, SM outlet Twitter was in the earlier stage of becoming a real time news outlet. As a part of background research, several journal publications [8][9] and news articles were reviewed, including an important role the Twitter SNS played during the 2011 Tunisian and Egyptian Arab Spring revolutions [10].

When the current research started, multiple resources were utilized to find journal articles, conference papers, magazine articles, book chapters, video presentations, and many other sources to understand the past and present of this research area. Rating an event using SM data is a relatively new area. Therefore, it lacks extensive publication. However, through the assistance of UAB's Mervyn H. Sterne Library website, multiple online databases, such as the Institute of Electrical and Electronics Engineers (IEEE), Association for Computing Machinery (ACM), etc. as well as Google Scholar, were used for the search and review of relevant publications in this area of research.

During the prototyping phases of rating models, Java programing language and the MySQL relation database were selected due to the researchers' prior knowledge of these tools. Therefore, there was no background research done on these tools.

Figure 1 shows different areas that this research touched on as a part of background reviews. Areas such as SM Data and Natural Language Processing (NLP), SM Data and Events, and SM Data and Sentiment Analysis were reviewed and analyzed.

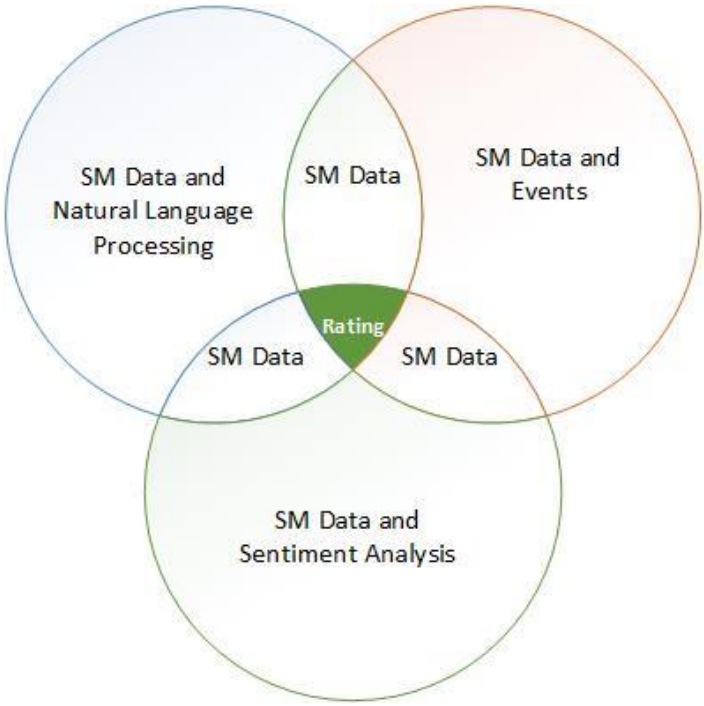


Figure 1: Background research areas

SM data and Natural Language Processing (NLP)

Natural Language Processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages[19]. In recent years, a lot of NLP has been developed to understand textual data. Table 1 lists some of the NLP toolkits.

Table 4. Natural Language Processing Toolkits

Name	Description
LingPipe	Processes text using computational linguistics. It automatically classifies Twitter search results into categories.
Apache OpenNLP	Performs tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and co-reference resolution.
Stanford Parser and	Reads text in some languages and assigns parts of speech to

[19] https://en.wikipedia.org/wiki/Natural_language_processing

Part-of-Speech (POS) Tagger	each word (and other tokens), such as nouns, verbs, adjectives, etc.
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.
Natural Language Toolkit (NLTK)	Works in computational linguistics using Python.
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.
GATE	Uses for all types of computational tasks involving human language.
NLP Toolsuite	Collections of NLP components.
Tweet NLP	Provides a tokenizer, a part-of-speech tagger, hierarchical word clusters, and a dependency parser for tweets, along with annotated data and web-based annotation tools.

As a part of the earlier research, the Stanford NLP Group's works were reviewed. The Natural Language Processing Group at Stanford University is a team of faculty, research scientists, postdocs, programmers and students who work together on algorithms that allow computers to process and understand human languages[20]. The Part-Of-Speech Tagger (POS Tagger) is a NLP the group reviewed. A POS tagger is a piece of software that reads text in some language and assigns parts of speech to each word (and other

[20] http://nlp.stanford.edu/

tokens), such as nouns, verbs, adjectives, etc.[21]. The group developed a part-of-speech tagger that demonstrates the following ideas: (i) explicit use of both preceding and following tag contexts via a dependency network representation; (ii) broad use of lexical features, including jointly conditioning on multiple consecutive words; (iii) effective use of priors in conditional log-linear models; and, (iv) fine-grained modeling of unknown word features[11]. Using these ideas, the group was able to develop the resulting tagger with 97.24% accuracy. The research group provided the tagger under the GNU General Public License, which includes components for command-line invocation, and a Java API[21]. A detailed course video on the NLP can also be found at the coursera.org[22] website.

Since this research was related to the Tweeter and NLP, the Tweet NLP research project[23] was also reviewed in detail. The project developed the Part-of-Speech (POS) tagging for Twitter, which has a tag-set, annotate data, and report tagging with results nearing 90% accuracy [12]. The Tweet NLP POS improved the part of speech through word clustering [13] via the Brown clustering[14] method on a large set of unlabeled tweets. The final results of this research had been publicly released on their website: http://www.ark.cs.cmu.edu/TweetNL. These results include evaluation data, annotation guidelines, open-source tagger, and word clusters. This Tweet NLP gave much needed input and research ideas for this research.

As there is ongoing research and development in the area of NLP to develop and understand the SM data to further advance the NLP, there are also researches to

[21] http://nlp.stanford.edu/software/tagger.shtml
[22] https://www.coursera.org/course/nlp
[23] http://www.ark.cs.cmu.edu/TweetNLP

understand validity of these NLP for sentiment analysis. One piece of the NLP research indicated that "sentiment analysis is only no better than manual analysis of social media data toward the goal of supporting organizational decision-making, but may even prove disadvantageous to such efforts."[15]

After review, analysis, and detailed study of the NLP APIs, the POS portion of NLP was not used in this research due to many unanswered questions on this topic. The Tweet NLP project's clustering ideas were used.

SM Data and Events

Long before SM was part of everyday life, event planning was still going on. With the birth of SM and its uses, these events' chatters could be listed to capture, understand, evaluate, and produce some meaningful information for current and future event planning.

As part of this research, there was much research conducted in the area of event SM data analysis. One of those large events was Super Bowl XLVI. It utilized SM data beyond marketing for hospitality, accommodations, and safety [16]. The research used Facebook, Twitter, and blogs to collect its data. It was interesting research because it fell too closely to this research topic.

There were also many different types of research done in this area of SM data and events, such as the 2010-2011 Australian floods[17], predicting flu trends using Twitter data [18], understanding approval ratings of election candidates[19], predicting national suicide numbers with social media data [20], etc. Even with the unstructured nature of text data and the maximum of 140 characters allowed, the Twitter SNS is still used as one of the main data sources for these research types.

SM outlet Twitter provides great ways to capture data in real time [21][22] and predict its results. Since this research is based on almost real-time importing and processing data, some reviews were done on the research topic, such as "Identifying Relevant Event Content for Real-time Event Detection [23]." This research's approach was to continuously monitor emerging Hashtags and rate them by their similarity to specific pre-defined event Hashtags using TF-IDF vectors [23].

Even with the vast amount of research in this area of SM Data and Events, these ideas were not taken further to actually rate an event using SM data.

SM Data and Sentiment Analysis

Finding the sentiment of a SM post is a big part of this research. Some of this research time was focused on looking at SM data and sentiment analysis. Sentiment analysis is not a new topic, but over the years, many concepts have been developed to understand sentiment of a SM post [24][25][26]. Some of the research done in this area also include sentiment analysis using emotional signals[27], word weight based on context [28], etc. As a part of this background research, the linguistic features for detecting the sentiment of Twitter messages [29] were also reviewed.

Summary of background research

Even after the background research had been conducted, there was still a gap and inadequate information found in this area of an event rating using SM data. Therefore, all the background research ideas were taken further to develop a rating model using SM data.

OVERVIEW

To achieve the research goal of building a rating model to rate an event using SM data, four different journal and/or conference research papers were written. Each paper is built into each other to demonstrate and expand the research idea. Here is the list of the four papers' summaries.

1) **Using Social Media data as research data:** In this journal paper, the current state of SNS, SM data, Natural Language Processing, and current rating systems was reviewed.
2) **Building a Rating Model:** In this conference paper, the foundation and road maps were created for future models.
3) **How is my event rated? Rating an event using Social Media data:** In this journal paper, the first three of nine models to rate an event were discussed in detail.
4) **Can an event be rated using Hashtags, Emoticons, Images, and URLs? Rating an event using Social Media data:** In this journal and conference paper, the remaining six models to complete the rating models were discussed in detail.

Table 2 shows a list of model names, summaries, and page numbers of ten different models that were discussed during this research, including the final model, which is a unified model of the first nine models.

Table 5. Summary of rating models using SM data

Model name	Summary of model	Page number
Word-by-word sentiment model	Each SM post is sliced into multiple words and clustered into either the positive, negative, natural or not applicable sentiment category.	50
Multi-words association sentiment model	A set of words in a SM post is used to make sentiment analysis decisions. Directional indicators such as "forward," "backward," or "both sides" of a word were used to review the sentiment of a set of words.	56
Word weight factors sentiment model	A SM post could have words with different word weight strength values. Each word with numeric value is used in this model.	60
Hashtag (#) sentiment model	Those Hashtags that were not used during the collection of SM posts are used for this sentiment model. The higher the rate of use of a Hashtag for a given dataset trumps a higher ranking.	77
Emoticon sentiment model	Each emoticon is captured and analyzed to make better sense of its sentiment value and then assigned to positive, negative, and neutral categories.	82
Event organizers sentiment model	Event organizers' Twitter handle user ids are used to filter their SM posts, and those	86

	tweets of organizers' are assigned a constant numeric value of 5 to those posts.	
Image sentiment model	To find the sentiment of an image, this model considered finding sentiments expressed by one or more than one person in an image through facial expression, full body expression, etc.	88
Popularity sentiment model	The popularity of SM posts is determined by the number of reposts SM posts received such that the higher the repost values, the higher the sentiment rating of a post.	91
URL sentiment model	A SM post with a URL is used for this sentiment model. Sentiment value associated content with a URL is used to identify the rating of a SM post.	92
Unified rating model	In this final model, an average sum of each SM post's numeric rating values that was generated from the previous 9 models are used to find the overall rating of a SM post.	93

Out of the first nine models, six different prototypes were created that helped to further validate rating models. The following steps were taken during the model building process for each of the working prototype models.

a. During testing, a process was created to generate numeric rating SM data.
b. In the results section, process results were reviewed.
c. Within the validation section, human vs. computer process rating results were compared.
d. In the discussion section, the relevance of the model was discussed.

USING SOCIAL MEDIA DATA AS RESEARCH DATA

by

SUMAN SILWAL AND DALE W. CALLAHAN

International Journal for Innovation Education and Research
Vol-1 No-3 November 2013

Format adapted and errata corrected for dissertation

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 banking, travel, healthcare, biotech, software, sports etc.*

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

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

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 (SNS) generate millions of posts and updates. In this new era of SM, 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 surrounding SM, much research and product development are done in the area of marketing and advertising through investigation of SM data. SM data give market researchers great opportunities to find people's interests, products they like and use, current trends, etc.

Importantly, 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 regard 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. As a result, 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 are listed on 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, as more and more people are using SM to get current news and updates 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 an 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.

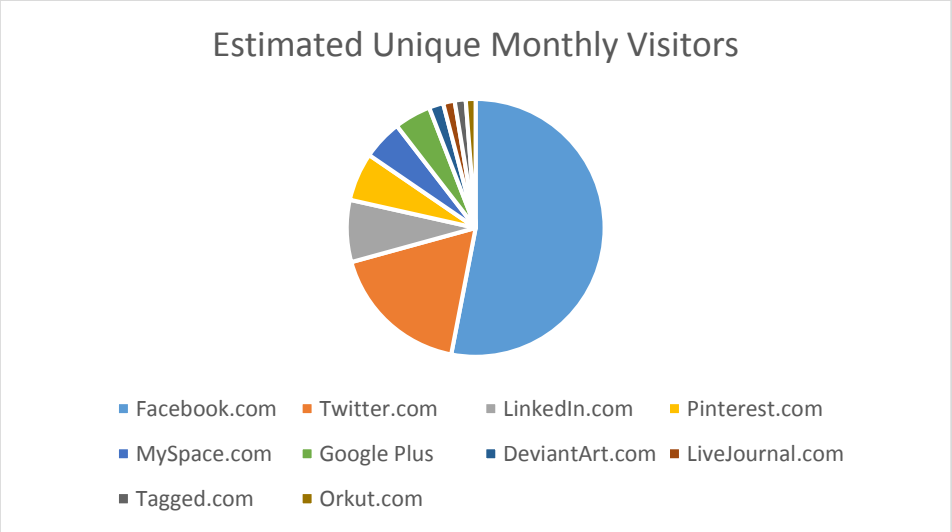


Figure 1: Data from Table 1

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 amounts of monthly data generated by only two Social Networking Sites. Based on review of the top 10 SNS from table 1, it is clear that there are more than 200 SNS contributing 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. However, filtering, validating, and capturing useful information from unstructured SM data is 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 Processing (NLP) can play an important role in analyzing these SM data. In the next chapter, NLP will be discussed.

3. Natural Language Processing (NLP)

NLP is described as a computer system that 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 speak the same language. Computer systems use byte-code.

Table 2 provides a list of NLP toolkits with a description of each as well as 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 tokens), such as nouns, verbs, adjectives, etc.	Java	http://nlp.stanford.edu/software/tagger.shtml
OpenFst	Keys applications in speech recognition and synthesis, machine	C++	http://www.openfst.org/

	translation, optical character recognition, pattern matching, string processing, machine learning, information extraction and retrieval, among others.		
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 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 hashtag (#) searches 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 analyses are done. In this model, data are filtered and evaluated according to hashtag (#), mention, and following and/or followers information. It provides a lot of information about current trends, popularity of a person or subject, breaking news, etc.

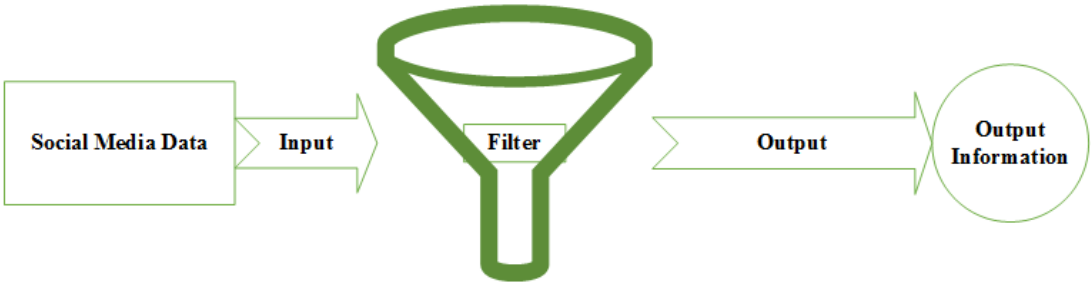


Figure 2: Social Media data process

Even though most Social Media analyses are done in a real time manner, they fail to provide a deeper look into users' sentiments. Consequently, researchers are missing out on valuable information. To understand the true meaning beyond Hashtags (#) and

mentions, these SM data need to be analyzed further by using other models and processes.

4.2 Developing Social Media Users' Sentiments Model

Understanding users' sentiments 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. Because of such limitations, 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 obtain some useful information. Once the information is ready, it will be sent to the output system.

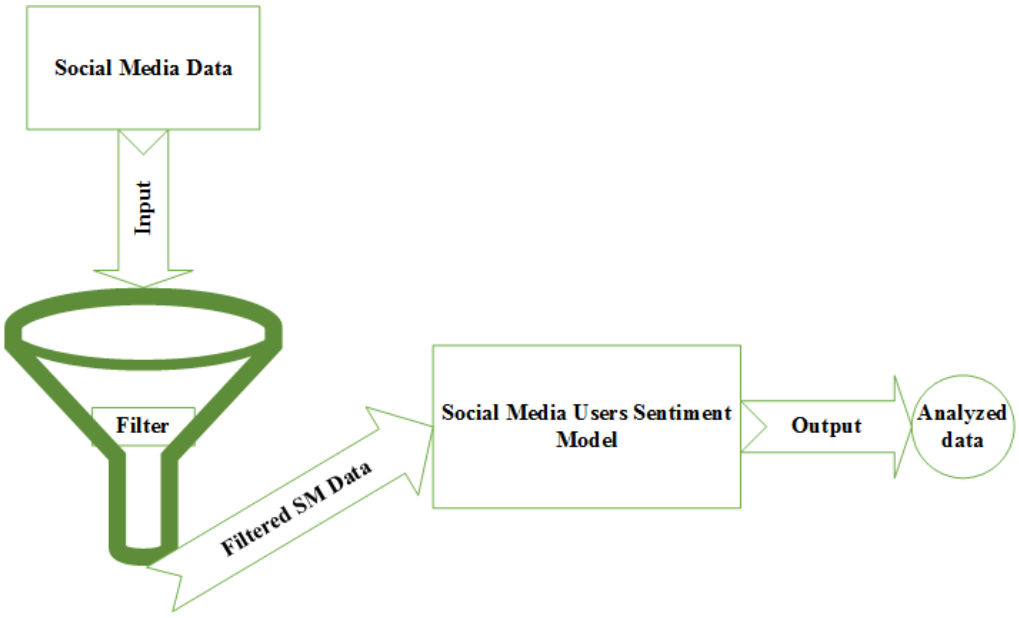


Figure 3: Users' Sentiments Model

In this model, most of the work is done at the Social Media Users' Sentiments Model stage. Using NLP sentiment analysis is just the first phase of the model's development. Even in this initial stage of development of the Users' Sentiments Model, there is 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 it is providing them. This growth is hard to ignore. The active users' participation with and contribution to SNS' data gives researchers untapped resources that can be used for finding solutions to complex problems.

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

6. Future works

In future research works, SM data will be extracted to develop a rating model process to rate an event.

.

7. 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/.

BUILDING A RATING MODEL

by

SUMAN SILWAL and DALE W. CALLAHAN

IEEE SOUTHEASTCON 2014

Format adapted and errata corrected for dissertation

Abstract— Social Media (SM) data are growing, and SM is becoming an acceptable part of daily life for billions of people around the world. Extracting information from Social Networking Sites (SNS) can provide great challenges as well as opportunities. One opportunity is that using SM data beyond day-to-day communication can provide additional values. Specifically, there is much research and many products dedicated to taking SNS beyond communication channels.

This research goes beyond specific tools inherent to the SM, such as Hashtag mentions and Like counts. Instead, it will use text-based modeling, data mining techniques, Natural Language Processing, machine language, etc., to understand SM content to produce numeric ratings. The final contribution of this research is building a rating model for an event using SM data. At this point, this research is laying out a road map.

Keywords— Social Networking Sites; Social Media; Rating Model; Rating; Natural Language Processing

Introduction

Trading goods, providing services and organizing events have been rated in some fashion by word of mouth, print format, or, in recent years, via websites like Amazon.com, TripAdvisor.com and MarathonGuide.com. Each of these websites provides great options for a person to rate and review a product, a service or an event. In order to rate, a user is required to register. Once registered, the user can log into the rating system, add comments, and give a rating. Figure 1 has an example from the TripAdvisor.com rating and reviewing process.

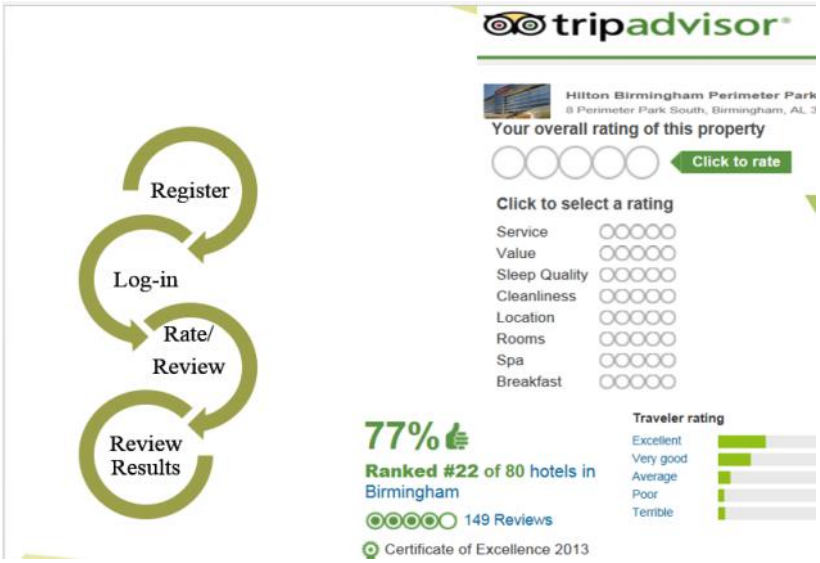


Figure 1: TripAdvisor.com rating and reviewing process

Since year 2000, one of the leading marathon survey websites, MarathonGuide.com, has collected little more than 1,800 surveys for the World Marathon Majors[24] like the Boston Marathon, the New York City Marathon, and the Chicago Marathon. This number of surveys is less than 0.01% of the total participants for these events over the past 13 years. Consequently, the survey data do not provide a large pool of data for these races. In comparison, SNS like Twitter.com and Facebook.com have an influx of comments, feedback and experience posted for the same marathons almost in real time.

The previous journal paper [1] briefly discussed possibilities of building an input/output model where SM data can be retrieved, filtered, and analyzed. to produce numeric results [1]. In this paper, this idea will be taken a little further and create a road

[24] http://worldmarathonmajors.com/

map for a rating model for an event using SM text data. This research is still in the preliminary stage of developing a full version of a rating model for an event.

For initial research, Twitter.com will be used as a Social Media data feed, and the Chicago Marathon will be used as the event to build a rating model. Twitter.com provides a great option for a person to express his/her thoughts and emotions in an open platform [2] within a limit of 140 characters.

Preview of Rating Model

Social Media Data - Manual Rating

Table 1 lists 5 random tweets, during and after the 2013 Chicago Marathon, with a Hashtag (#) mention of #ChicagoMarathon from Twitter.com. Just looking at each tweet, a person can identify related or unrelated data for the marathon rating system. Also, a random 1-5 numeric rating was assigned for related Chicago Marathon tweets, as shown in column 4 in table 1. If numeric ratings are averaged for all related tweets from table 1, the manual rating model will produce a 3.5 average rating for the 4 related Chicago Marathon tweets.

TABLE 1 2013 CHICAGO MARATHON TWITTER.COM POST DATA

	#ChicagoMarathon Hashtag mentioned data with a 1-5 rating		
	Tweets	***Related/ Not Related***	***Rating***
1	4:17 in my marathon debut. Fell off my goal, but this was so incredible. Hungry for more! Sub-4 then eyes on Boston. **#ChicagoMarathon**	Related	4
2	**#chicagomarathon** fails. No spectators aka friends and family	Related	1

	are allowed to watch their runners cross the finish line for fear of terrorist		
3	No PR but I finished the **#chicagomarathon**. Thanks for all the support. I am officially retired!	Related	4
4	Many thanks to the volunteers and spectators of the #**chicagomarathon**! You made this day possible! You all #ownchicago	Related	5
5	Does anyone know if the #**chicagomarathon** is over yet?	Not related	NA

Rating Model – Bigger Picture

The rating model in figure 2 previews a complete goal of this research to produce an output numeric rating for inputting tweets. As in the previous example of the manual rating model presented in table 1, the objective of the rating model is to produce similar rating results automatically without any manual intervention.

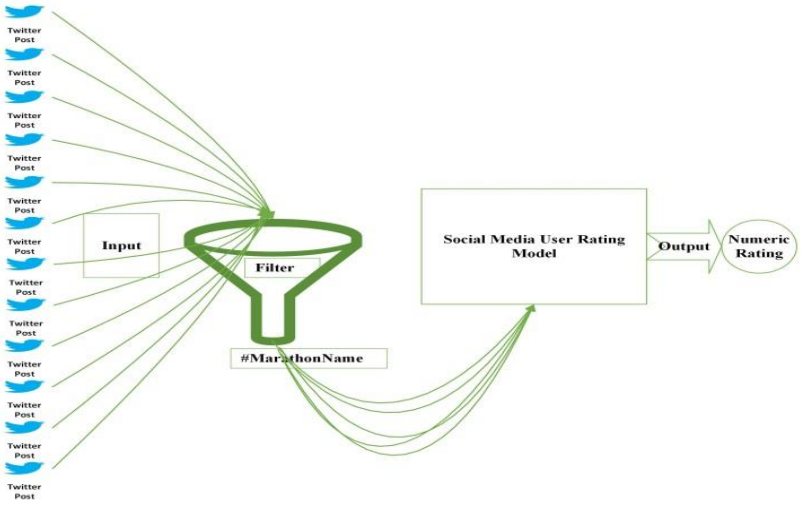


Figure 2: Rating model – a complete view

There are four steps to complete a numeric rating model using Twitter.com: retrieving data, filtering valid data, processing data, and producing a numeric rating for each tweet. In future chapters, each of these components will be discussed in a little more detail.

Retrieving SM Data

The first step in building a rating model using Twitter is to sign up for Twitter.com's developer network, which provides authority to access Twitter from an application. Twitter.com's security key and token are used with an Application Programming Interface (API) like Spring Social Twitter[25] to connect and access Twitter.com's APIs and its data.

Filtering Valid SM Data

Twitter.com provides a Hashtag (#) filtering/searching tool, which helps to retrieve only valid data to process. The Spring Social Twitter API has a built-in search algorithm to return the first 50 matching tweets per API call, which can be stored into a database. Figure 3 shows an inside look of the SM data filtering process.

Database schema for the rating system can look like the following: Twitter Id - Integer, Source – Char (20), Filter Criteria - Char (250), Type - Char (10), Tweets - Char (200), Rating - Numeric.

[25] http://projects.spring.io/spring-social-twitter/

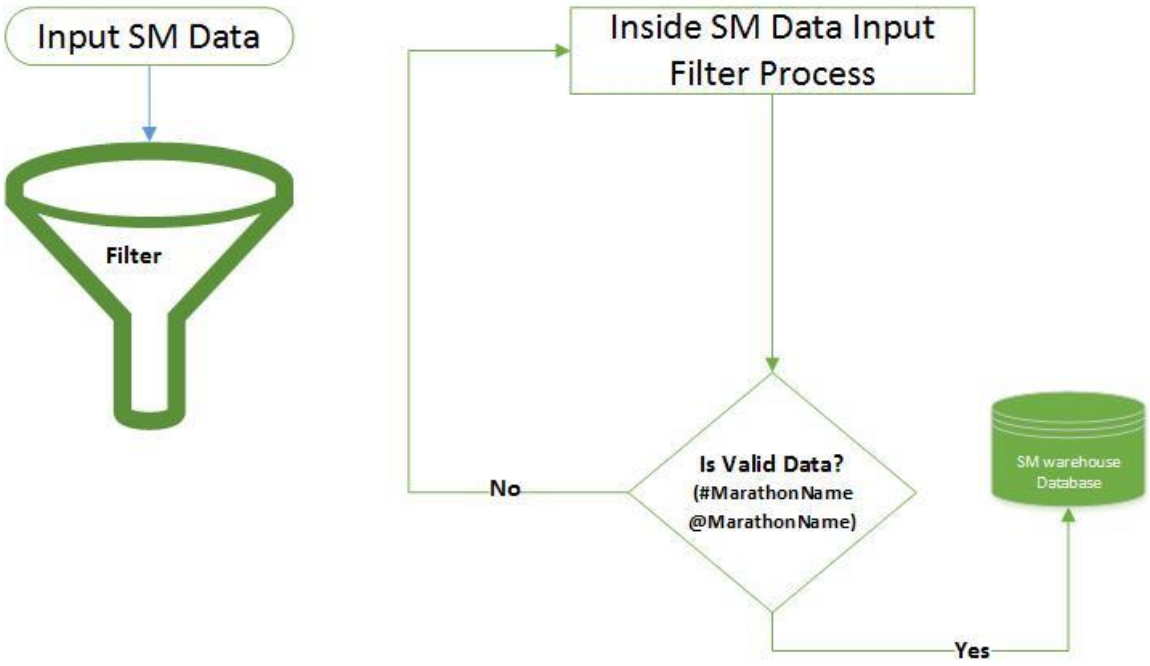


Figure 3: Inside SM data input filtering process

SM Data Users' Rating Model Processing

Processing SM data to understand emotions, facts, and related and unrelated information to build a users' rating model is the core of this research. A tweet can contain text, images, links, emoticons and/or combinations of all types. Each of these different types of tweets can be utilized differently to a get a comprehensive users' rating model for an event, product, or service. At this time, this research is only looking at text-based tweets.

Figure 4 provides a road map for this research. This rating model can be considered a "system of systems"[3] where each individual component can contain its own system with processes, measures, and matrixes.

There are many NLPs, APIs, and tools being developed for linguistic analysis of SM data [4]. For the first step of understanding Twitter data, this research is looking at

Twitter's NLP[26] for a tokenizer and a part-of-speech tagger[5]. There are also many other NLPs, which can help to develop the rating model[1]. At this time, this research is still working on evaluating suitable APIs or tools to help to achieve its goal of rating an event using SM data. An abundance of research where Twitter data were used was reviewed [6][7][8].

Initially, a specific SNS or several tools and APIs to develop a users' rating model may be used. The final model should be platform independent.

SM Text Data Rating Process

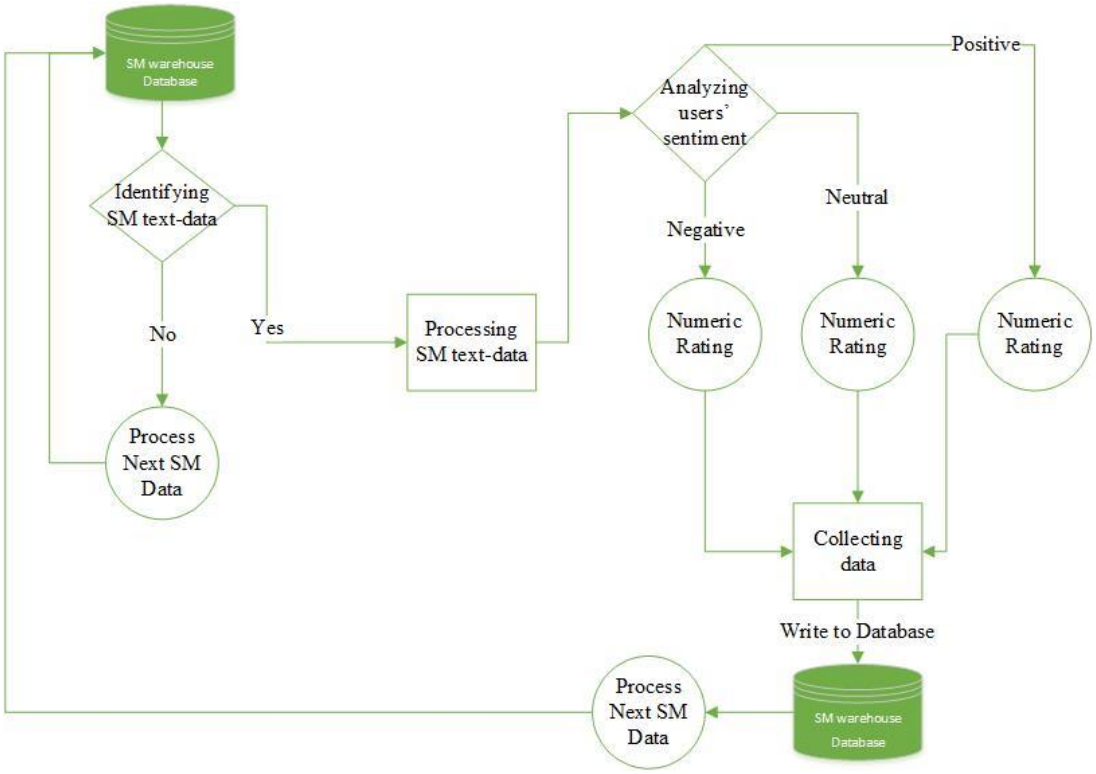


Figure 4: SM text data rating model

In the next few bullet points, each step will be briefly broken down (figure 4).

[26] http://www.ark.cs.cmu.edu/TweetNLP/

1. *Identifying SM text-data*

Unrated SM data from the warehouse database tables are passed to the next process where each of these tweet rows is filtered to identify its types. If a tweet does not have any text data other than Hashtag text, it will be ignored. Then text SM data will be passed on to the next process.

2. *Processing SM text-data*

This process will look at each text data row to understand emotions, facts, and related and unrelated information for the #ChicagoMarathon Hashtag data. At the end of this process, this research is expected to reflect a very good idea of what kind of data are getting reviewed as well as the meaning for these types of SM data. Most of the time conducting this research will be spent finding the best possible APIs, tools, and/or algorithms to understand these SM data.

3. *Analyzing users' sentiment*

In this process, SM text data will be taken through an NLP sentiment analysis process to give a negative, positive or neutral understanding of the row. Each SM data row will be tagged with a numeric rating according to rules that will be defined.

4. *Collecting data*

In this process, each row will be updated with numeric rating numbers produced by the previous process. Once this update is done, the process will move to the next row of SM data.

Conclusion

The final results from processing Twitter data rows should be similar to the rating data in table 1. This research is still in the early stages of design, development and implementation of the final product. The final contribution of this research is to create a rating model for an event using SM data.

As this research expands on the idea of developing a rating model building process, multiple APIs, databases, and tools will be used to complete the rating model. A complete model will be a framework of many frameworks or "system of systems"[3], where the final model should be a platform and SNS independent.

Building an automatic rating model is an exciting opportunity to provide an almost real time rating for an event like the Chicago Marathon. Once it is fully developed, it can help to rate products and services without direct input into a rating system like Amazon.com, TripAdvisor.com and MarathonGuide.com.

Future Works

At this point of research, there are still a lot of unanswered questions. In coming months, different programming languages, APIs, algorithms, databases and processes will be evaluated and used to build a rating model from SM data for an event.

References

[1] S. Silwal, “Using Social Media Data as Research Data,” vol. 1, pp. 49–55, 2013.

[2] “Legal Loop: NY Judge Rules Public Tweets are Public | The Daily Record | nydailyrecord.com.” [Online]. Available: http://nydailyrecord.com/blog/2012/07/06/legal-loop-ny-judge-rules-public-tweets-are-public/. [Accessed: 12-Jul-2013].

[3] A. Odusd and T. Sse, *Systems Engineering Guide for Systems of Systems*, no. August. 2008.

[4] K. Gimpel, N. Schneider, B. O’Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, “Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments,” *Hum. Lang. Technol.*, vol. 2, no. 2, pp. 42–47, 2011.

[5] O. Owoputi, B. O’Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith, “Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters,” in *Proceedings of NAACL-HLT 2013*, 2013.

[6] C. Wojtalewicz, “Social Media Use for Large Event Management,” no. 1, pp. 24–29, 2012.

[7] H. Achrekar and A. Gandhe, “Predicting Flu Trends Using Twitter Data,” ... *WKSHPS), 2011 IEEE …*, pp. 702–707, 2011.

[8] K. M. Larson and R. T. Watson, “The Impact Of Natural Language Processing-Based Textual Analysis Of Social Media Interactions On Decision Making,” 2013.

HOW IS MY EVENT RATED?
RATING AN EVENT USING SOCIAL MEDIA DATA

by

SUMAN SILWAL and DALE W. CALLAHAN

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY SCIENCES AND ENGINEERING

VOL. 6, NO. 3, MARCH 2015

Format adapted and errata corrected for dissertation

How is my event rated?
Rating an event using Social Media data

Suman Silwal, Dale W. Callahan, Ph.D., P.E.

Abstract — In recent years, events have been continuously discussed on Social Media in the form of status updates, posts, and comments by participants, volunteers, and supporters. Social Media content generated before, during, and after an event could offer valuable insight into the success and popularity of an event. It can also generate ideas for future improvement of the event.

With the fast evolving nature of Social Media, current events' Social Media content is ignored, forgotten, and overlooked for new sets of future posts, discussions, and comments.

This research demonstrates that any publically available Social Media data can be captured and analyzed to produce some meaningful information. As a result, a rating model was created through combinations of multiple models using SM data to rate an event.

Key Words: *Social Media(SM), Social Networking Site (SNS), Rating Model, Sentiment Analysis, Marathon, Twitter, Hashtag (#), Marathon Rating, Event rating*

I. INTRODUCTION

Social Media (SM) data are "unstructured, informal, and fast-evolving" [1] in nature. In recent years, more and more people have been sharing their thoughts, feelings, sentiments, etc. on SM about events, products and services [2]. As the growth of SM

uses is happening, so is the interest on research and development to utilize these SM data [2][3][4].

In recent years, many different groups developed Natural Language Processing (NLP) tools such as the Tweet NLP (http://www.ark.cs.cmu.edu/TweetNLP/) and the Stanford NLP (http://nlp.stanford.edu/) to understand sentiments of a SM post. The Tweet NLP uses tokenizing, clustering, and part-of-speech tagging approaches for Twitter data [5]. Even though it takes some effort to understand and obtain meaningful results using SM data, there are still a lot of unanswered questions regarding the findings of SM sentiments using existing sentiment tools and NLP [6][7]. Social Networking Sites (SNS) have a lot of opinion spam [8] and fake opinions [9]. Due to the unstructured nature of SM data, finding quality user-generated content [10] from SM posts is always a challenge. Even with a data set that is filtered and domain specific, understanding and producing meaningful information by processing an individual SM post through a computer program provides added challenges.

A single SM post can consist of words, abbreviations, numbers, hashtags, images, mentions, links, special symbols, emoticons, etc.; furthermore, this information can provide insight into understanding the overall sentiment of a SM post.

In previous research [2][11], a foundation and road maps were built where many ways to understand positive, negative and neutral sentiments of an event's SM posts were discussed. In this research, multiple modeling techniques are used to capture different aspects of SM posts to achieve the overall goal of creating a rating system using SM data. A "systems of systems" [12] is built using a model of models in an interdisciplinary manner to capture and analyze SM data to produce some meaningful information. The

final outcome of this research is to build a numeric rating system of an event using SM data as well as to compare and validate those output data.

Here are some of the core systems, events, and frameworks that are used to build the user rating models and processes:

- ***Twitter.com*** (Twitter) is used as the main SNS for this research. It provides its developer network limited access to its publically available data through its Application Programming Interface (API) [13].
- ***Marathon*** events are 26.2-mile foot races that are considered the event topic of this research.
- ***MySQL database system*** is a widely used open-source relational database management system (RDBMS) that provides different tools to access and manage the database.
- ***Java programming language*** is an open-source computer programming language, which is widely used for application development.
- ***Spring*** java framework provides a lot of different components to build a very powerful application, including Spring Social API. It is used to glue these applications together.

In the future sections of this research paper, the rating building process will be broken into 3 different parts:

1) ***Data importing*** consists of importing and inserting SM data into a local database.
2) ***Sentiment dictionary*** building consists of creating processes to generate sentiment dictionaries.

3) ***Sentiment Modeling*** defines different possible modeling techniques needed to build a rating system.

   a. During the testing section, a process to generate numeric rating SM data will be built.

   b. In the results section, results generated by the testing process will be reviewed.

   c. Within the validation section, the human rating from results will be compared with the computer-generated rating.

   d. In the discussion section, the relevance of this research model will be discussed.

II. SM RESEARCH DATA COLLECTION PROCESS

Collecting valid sets of data plays an important role in the success of any research, including this research. Any open forum on the Internet can contain noise and misleading information [14]. In this research, it is important to find, filter, and collect only necessary data to avoid overloading of unnecessary and excessive data.

Once a developer's access account is set up with the Twitter SNS (https://dev.twitter.com/), a request is sent to set up a consumer key, consumer secret, access token, and access token secret for authentication and authorization to its API.

During the data import process, once a valid handshake is made through Twitter's authentication APIs, the data import process gets access to Twitter's dataset. By default, the Spring Social (http://projects.spring.io/spring-social/) search API for Twitter can retrieve up to 50 of the most recent matching tweets per call. Also, Twitter allows only 180 requests/queries per 15 minutes to its API per run.

This research is mainly based on yearly marathon events, which are heavily discussed close to the actual race day. To prevent accessing irrelevant Twitter data, a search look up table using search criteria (table 1) with active status was created. The Twitter data collection process (figure 1) runs almost in real time; as a result, this process imports data into the SM warehouse database table.

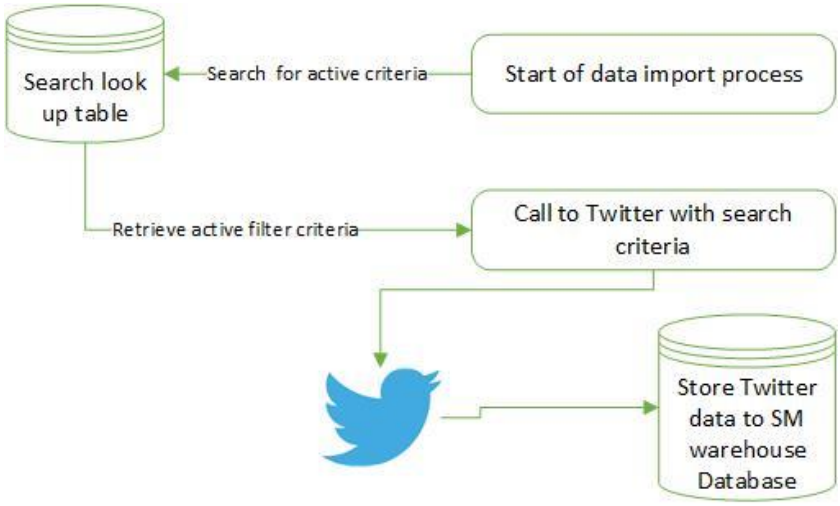


Figure 1: Twitter data import

Once the search look up criteria are handed to the import process, most of the heavy lifting to retrieve proper Twitter data is done within the Twitter search API.

Table 1: Some of the Search Criteria

Search Lookup Criteria
#mercedesmarathon
@Run_Mercedes
#BostonMarathon2014
#ChicagoMarathon

At this time of this research, Tweet Id, Tweets Text, Generated from user, Tweet created date, and Retweet count information are captured from each Twitter per API call.

Imported tweets are stored in a local warehouse database table for future uses (table 2). As more marathons are brought into the rating mix, this list of data is bound to grow.

Table 2: Some of the marathons' data collection counts

Event Names	Total Count
Boston Marathon	168357
Country Music Marathon	2997
Flying Pig	1884
Marine Corps Marathon	3349
OK City Marathon	1526
Richmond Marathon	305
St. Jude Marathon	1257

III. DICTIONARY BUILDING

Building a valid dictionary is a very important part of this research. A dictionary gives an advantage in creating a structure around unstructured SM data. Each word in the dictionary table can have multiple attributes, such as sentiment, trending count, weight, etc. to help understand more about each word.

Furthermore, each word in this dictionary can also be clustered into a positive (P), negative (N), neutral (NU) or not applicable (NA) sentiment category.

A. Initial sentiment dictionary building process

Initially, predefined sentiment words from different websites were imported into the sentiment dictionary table (figure 2). This provided a good set of data to start with predefined values.

Since this research is based on specific Twitter data and marathon running events, these initial sentiments were not sufficient. Therefore, additional words were added to the dictionary using the sentiment dictionary building process.

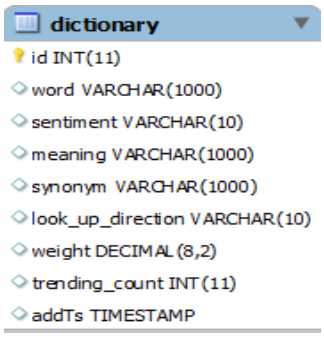


Figure 2: Dictionary table

B. SM data sentiment dictionary building process

In this sentiment dictionary building process (figure 3), at first, a call is made to the Twitter data warehouse table. Then, each of the resulting tweet posts are split into multiple word rows and stored into a temporary dictionary table. For this process, each word with a special character, symbol, link, numeric value, emoticon, etc. is ignored. A valid and unique word from a tweet post is inserted into a dictionary database table for future use.

At the time of writing this paper, each word on this dictionary is manually clustered into one of the default sentiment categories (table 3).

Table 3: Default sentiment indicators

Sentiment Indicators	Descriptions
P	Positive
N	Negative
NU	Neutral
NA	Not Applicable

Even though it is a laborious process to create a word-based dictionary, this process gives control over how each word is perceived and evaluated without knowing the full context of a sentence. In this approach, each sentiment is defined purely on a word level. Table 4 lists some pros and cons of creating a domain-specific lexicon.

Table 4: Pros and cons of creating a sentiment dictionary

Pros	Cons
Quickly build dictionary words	Have to look for words
Domain-specific word	Getting unnecessary words into database table
Ability to expand attributes to understand a word	Manually enter into dictionary
Clustering words to different categories	Misleading sentiment by just looking at one word
Ability to create structure around words	
Reusability	
Grouping SM shorthanded words to real words	
Search ability	
Reusability	

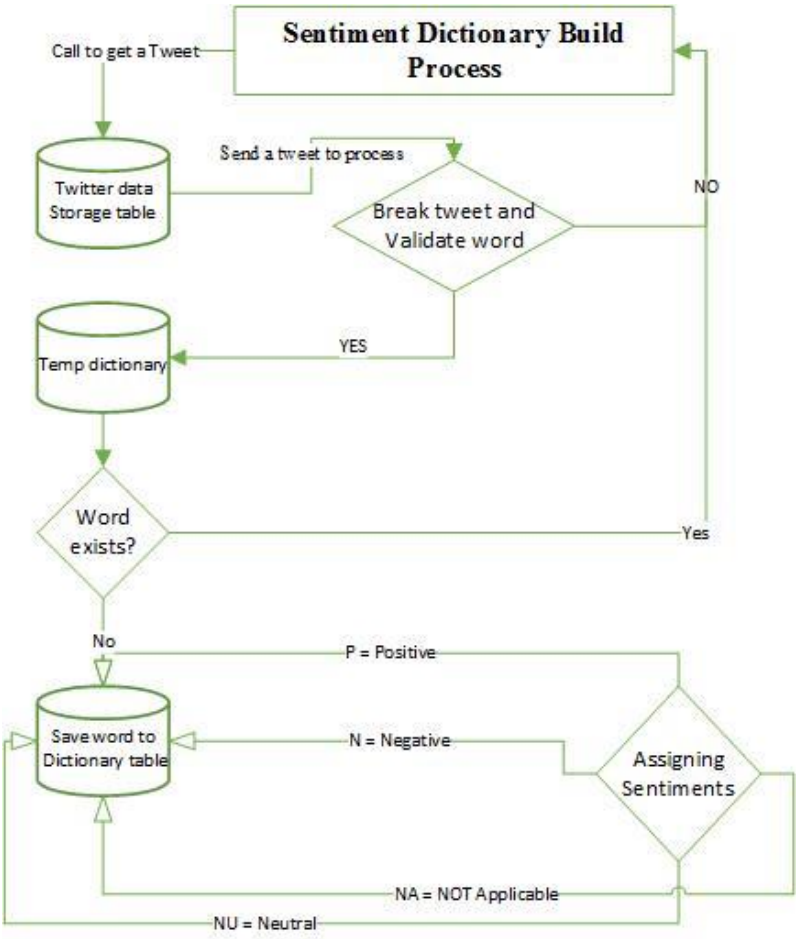


Figure 3: Dictionary data building process

IV. BUILDING A RATING MODEL USING SM DATA

Building a marathon event rating model using SM data in an interdisciplinary manner is core to this research. Thus, publically available SM data are captured and analyzed to produce some meaningful information. An ultimate outcome of this research is to build a numeric rating model through a combination of multiple sentiment analysis models using SM data. Importantly, a SM API such as Twitter gives access to different types of datasets, including geo location, add timestamp, tweet id, text, etc. This research is generally interested in text data of a SM post.

In recent years, there has been a lot of interest in the study of SM data to find social interactions, emotion [15], election approval rating [4], etc. At this time of research, this area of processing SM data to create a numeric rating system is still a new field of interest.

A rating model is built on the simple idea of an input-process-output model (figure 4), where input data is retrieved from a SNS data source. Those SNS data are processed to produce some meaningful information.

Figure 4: Input-process-out model

This section will present multiple rating models to reflect a possible sentiment of a SM post. Eventually, each of these models is put together to create a unified rating model that contains a model of models. Figure 5 shows an example of a tweet post that consists of many different aspects of a SM post such as text, images, URL, mentions, etc.

Figure 5: MarathonRuns' tweet before Houston Marathon

A. Word-by-word sentiment model

The word-by-word sentiment model is the first of many sentiment models that will be built as a part of this research. For this rating model, each word of a SM post is qualified to be reviewed for sentiment analysis. A single SM post can consist of many different types of words, abbreviations, numbers, hashtags, images, mentions, links, symbols, emoticons, etc. For this model, punctuation, reference to images, URLs, numbers, emotions, etc. were ignored.

In this model, each SM post is sliced into multiple words. Then, each of these words is clustered into either the positive, negative, natural or not applicable sentiment category. Having these words clustered into 4 different sentiment categories give a little sense of structure around the unstructured nature of a SM post.

The design of this model is dependent on the accuracy of each word's sentiment in obtaining the overall sentiment of an entire SM post. Eventually, each SM word's sentiment rating will produce an overall rating for an event associated with that SM post. Figure 6 shows what a word-by-word rating looks like in a bigger picture.

Table 5: Default sentiment indicators numeric value

Sentiment Indicators	Descriptions	Numeric Rating
P	Positive	5
N	Negative	1
NU	Neutral	2.5
NA	Not Applicable	0

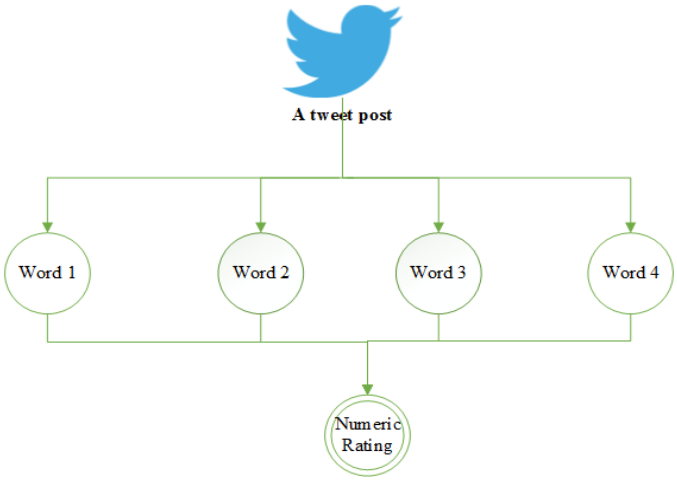


Figure 6: Word-by-word rating – a bigger picture

Table 6 shows how a SM post from figure 1 is broken into different words and sentiment categories.

Table 6: Word-by-word sentiment

Word	Sentiment
Good	Positive
Luck	Positive
To	Not Applicable
Racing	Neutral
#HouMarathon	Neutral
@HoustonMarathon	Not Applicable
!	Positive
Gr8t	Positive
Weather	Neutral
www.weather.com/weather/weekend/l/	Not Applicable
Go	Positive
Get	Neutral
Finisher	Positive
Medal	Positive
:)	Not Applicable

Each of these sentiment indicators from column 2 of table 5 is associated with a numeric value. Table 5 shows a list of sentiment indicators and the assigned default numeric value for this model. For this model, those values defined on table 5 are considered as default sentiment indicator values for all of the current and future models. Since numeric rating is based on a 1-5 rating system, this is a standard constant value for each sentiment indicator parameter. Also, it gives a structure and consistent look at SM data.

For a single SM post, the sum of sentiment indicators is multiplied by each numeric rating value associated with it. The sum of these values is then divided by the sum of the total sentiment. The following formula (1) provides a numeric rating for a SM post:

Numeric rating using word-by-word sentiment

$$= \frac{\sum(\boldsymbol{P})\times\mathbf{5}+\sum(\boldsymbol{N})\times\mathbf{1}+\sum(\boldsymbol{NU})\times\mathbf{2.5}}{\sum\boldsymbol{P}+\sum\boldsymbol{N}+\sum\boldsymbol{NU}} \qquad (1)$$

Based on this formula, the numeric rating for figure 5's SM post using the word-by-word rating model is 4.09. This result is very close to a numeric rating compared to a manual rating.

Due to the unstructured nature of SM data, the word-by-word sentiment model provides a great benefit in understanding SM posts through breaking each word into small units of its own. It can provide some insight into users' sentiments.

1) Testing

To test the word-by-word rating model, a testing process model was built. Initially, this process (figure 7) retrieves a single tweet post from the warehouse table and splits it into multiple words. Each of the valid words is sought in the sentiment dictionary table to find an associated positive, negative, neutral, and not applicable sentiment category. Figure 7 shows the core logic for collecting different sentiments of each word.

A tally is kept for each tweet word's sentiment category assignment count and numeric value associated with them. At the end of processing each tweet post, the formula (1) defined by the word-by-word rating model to find the numeric rating of a tweet was used. These values are stored in the Twitter warehouse database table field for future calculations.

1) Results

Table 7 shows the final results of the word-by-word rating model after processing three different events' rating results. Based on an initial observation of these results, each of these events is getting positive ratings.

Table 7: Word-by-word rating model results

Event Name	Word-by-word rating
Boston Marathon	3.6161
Richmond Marathon	3.7403
Twin Cities Marathon	3.5991

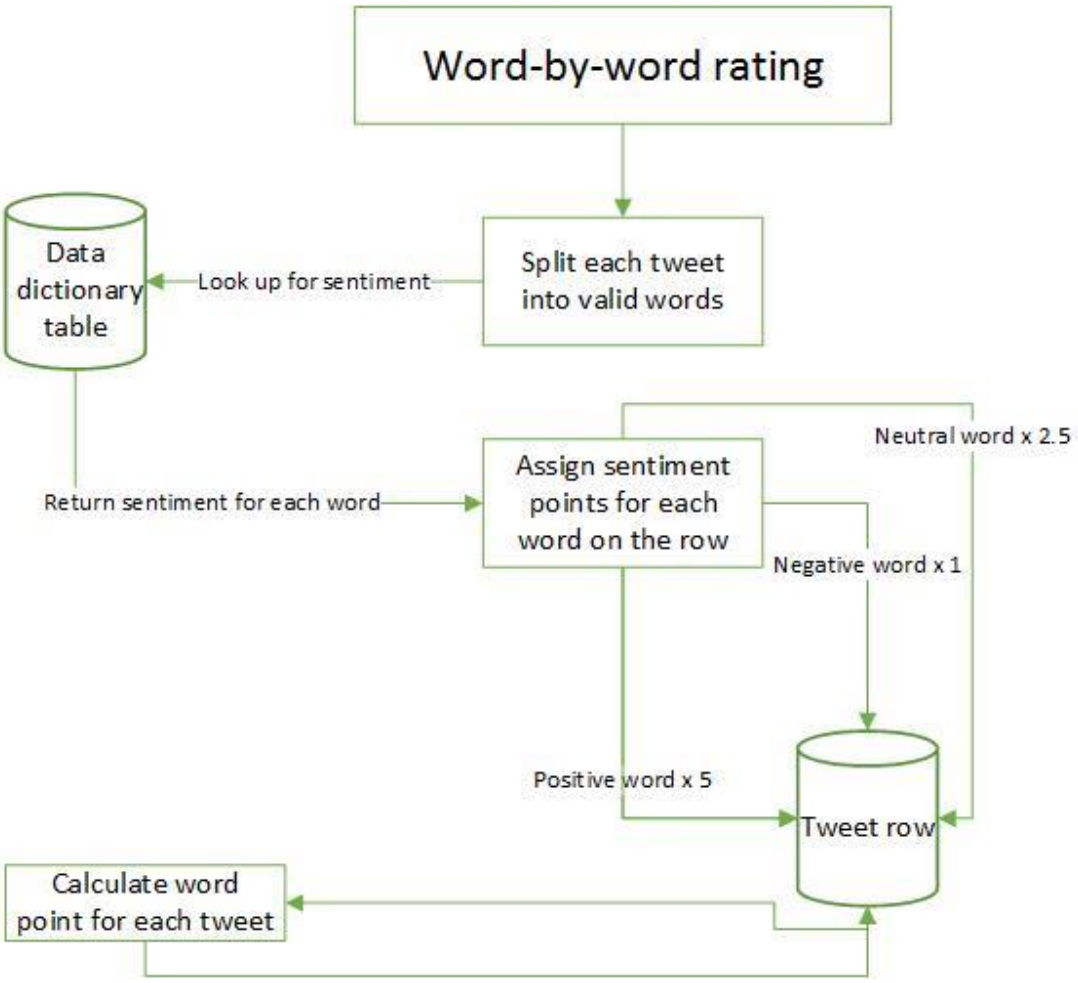


Figure 7: Word-by-word sentiment rating process

2) Validation

Validating these results of the rating system is an important part of this model. Since this model is trying to create a rating model using a computer process, there will always be a misunderstanding between human speech and a computer process's translation of such speech.

By comparing the overall results of the human rating and the word-by-word rating, the following results (table 8) were acquired. By keeping the human rating as the standard rating, the result is less than 4% difference (2) between the word-by-word rating and the human rating.

Difference % =

$$\left(\frac{Absolute\ Value\ of\ (Word\ by\ word\ rating - Human\ rating)}{Word\ by\ word\ rating + Human\ rating}\right) \text{ X } 100 \qquad (2)$$

Table 8: Comparing results

Event Name	Word-by-word rating	Human rating	Difference %
Boston Marathon	3.8258	3.5549	3.67%
Twin Cities Marathon	3.5991	3.5870	0.17%
Richmond Marathon	3.7403	3.7637	0.31%

3) Discussion

Due to the unstructured nature of SM data, the word-by-word sentiment model provides the benefit of understanding SM posts through breaking each word into small units of its own. It also provides some insight into users' sentiments, but it does not provide all the answers. As cumulative data comparisons, the biggest difference between the human rating and the process model rating is 3.67%, which is within the confidence level of 10%. It is great news for the rating model.

The review of line by line results from the Boston Marathon shows that more than 72% of tweets with more than a 10% difference in values were found. Some tweets were rated higher by the human rating, while other tweets were rated higher by the rating process (table 9). This may be due to the process only looking at one word at a time to make sense of the whole sentence, while the human rating is looking at the whole context of a tweet.

Table 9: Example tweets

Tweet	Word-by-word rating	Human rating
You don't get it, do you? It's not about winning--it's about participating. #BostonMarathon @JoeyDips	2.40	4.6
I could not imagine running 5 min miles for 26 miles #BostonMarathon	2.70	4.2

Ultimately, the word-by-word model is still a valid rating model as an initial model, but it is not sufficient enough to look at one word at a time to build an overall model of a SM post.

B. Multi-words association sentiment modeling

Within the multi-words association sentiment model, a set of words is looked up to make sentiment analysis decisions. Unlike the word-by-word rating model, in this model, not every word is qualified for bi-direction look up. For those qualified words, a directional indicator is used on each word so that it can be viewed from a specific direction: forward, backward, or both sides of a word (figure 8).

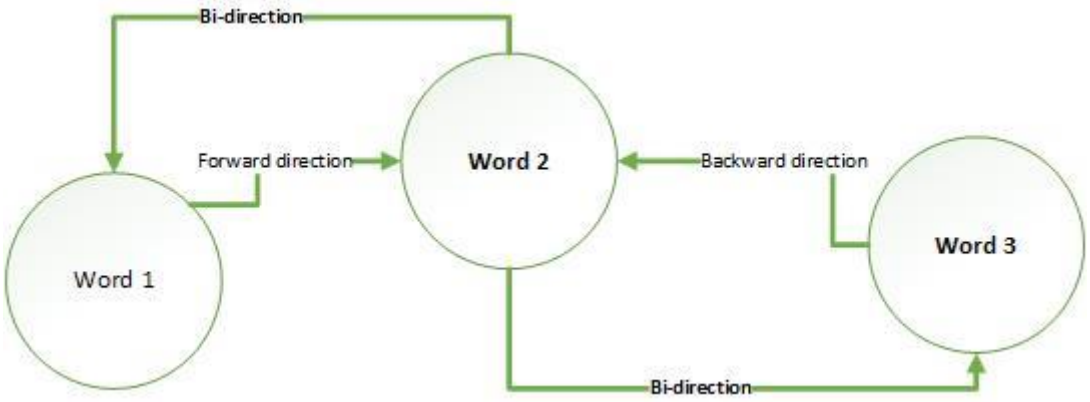


Figure 8: Multi-words association model

A multi-word association matrix (table 10) is created to identify sentiment for multi-direction look up words. In this matrix, a negative word generates negative sentiment for an associated word regardless of whether it has a positive or neutral sentiment; also, a positive word with neutral sentiment will generate a positive sentiment.

Table 10: Multi-word sentiment matrix

	Positive (P)	Negative (N)	Neutral (NU)
Positive (P)	P	N	P
Negative (N)	N	N	N
Neutral (NU)	P	N	NU

Table 11 shows how each of these associated sentiment words generate positive or negative sentiments using multi-word association model.

Table 11: Multi-word association sentiment examples

Sentence	First lookup Word	Associated Word	Sentiment
Good Luck today!	Good	Luck	Positive
Good grief run fast.	Good	Grief	Negative
Love my Finisher Medal.	Finisher	Medal	Positive
Gr8t weather	Gr8t	Weather	Positive
I do not like this race	Not	Like	Negative

Similar to the word-by-word sentiment model, a default numeric value of 1-5 is assigned to each word and its associated word(s) with similar calculations. For those words that do not have an associated sentiment, it follows the word-by-word sentiment model to assign sentiment values. For one sample SM post (figure 5), the process results received around a 4.16 rating using the multi-word sentiment model due to two positive words' association "Finisher" and "Gr8t" with other words.

1) Testing

To test the model, a multi-word association process is built to look up more than one word to obtain a more in-depth sentiment rating of a tweet post. To properly tag this new rating model, one more attribute to the dictionary table to indicate looking forward (F), backward (BK) or at both (B) directions of a word (table 12) were added manually.

Table 12: Word with sentiment and multi-direction look up

Word	Sentiment value	Look up direction
quick	P	F
lit	P	B
can't	N	F
looking	P	F
having	P	F
!	P	BK
cross	P	F

In this process, a complete sentence like "I do not like this run" has a bi-direction word "not," which looks at both sides of the word. Even if "do" and "like" are two positive words with sentiment value of 5, in this process the word "not" looks at both directions, which creates negative outcomes for words "do" and "like."

This process works very similar to the word-by-word rating process that was described in the previous section but with the added difference of look up in the multi-words association sentiment matrix (table 10). Every word in the dictionary table does not have bi-direction look up indicator. For those words without bi-direction indicator, they are treated as a word-by-word rating process. In general, this process works similar to the word-by-word rating model with multi-direction look up attributes.

2) Results

At a glance, word-by-word rating and multi-word rating results (table 13) had improved numeric rating values.

Table 13: Multi-word rating

Tweets	Word-by-word Rating	Multi-word rating
Congrats to @dianamchard for finishing the #BostonMarathon !	4.38	5.00
Amazing that an American man won the #BostonMarathon. #StorybookEnding	4.00	5.00
.@TylerPennel just won the @tcmarathon. His. First. Marathon. That's wild. #tcmarathon	3.33	5.00
Meb Wins Boston: Amazing things happen. Never stop believing. #BostonMarathon http://t.co/eFCnF4KQCC via @Flotrack	4.29	4.64

After comparing cumulative results from word-by-word rating and multi-word rating process results (table 14), very little difference was found in overall rating results.

Table 14: Comparing results from word-by-word and multi-word process results

Event Name	Word-by-word	Multi-word
Boston Marathon	3.6161	3.6083
Twin Cities Marathon	3.5991	3.5919

3) Validation

To validate results from the multi-word association process, the sum of the results retrieved from the sample data were taken and compared against the human rating (table 15). The Boston Marathon rating is still higher than the human rating, while Twin Cities Marathon's rating shows a consistent look.

Table 15: Validation after multi-word association process run

Event name	Multi-word rating	Human rating
Boston Marathon	3.8242	3.5549
Twin Cities Marathon	3.5991	3.5870

To further validate Boston Marathon's rating, another sample was taken for the marathon data, which were rated by the multi-word rating process and were not part of previous sample collections. Those sample data were sent for further validation. After taking the average from these new data sets, results were again compared (table 16). At this point, the overall Boston Marathon rating dropped by .6185 from the prior overall rating; also, there was a 0.2861 difference between sample data used by the multi-word process and human ratings.

Table 16: Validation after multi-word re-process run

Event name	Multi-word rating	Human rating
Boston Marathon	3.2057	3.4918

4) Discussion

A look at the sample set of data side by side (figure 9) shows that a lot of data from the human rating results are very closely rated to computer processing results. Even

though there was a big drop in numbers for overall rating using sample data, these were expected results due to multi-words rating.

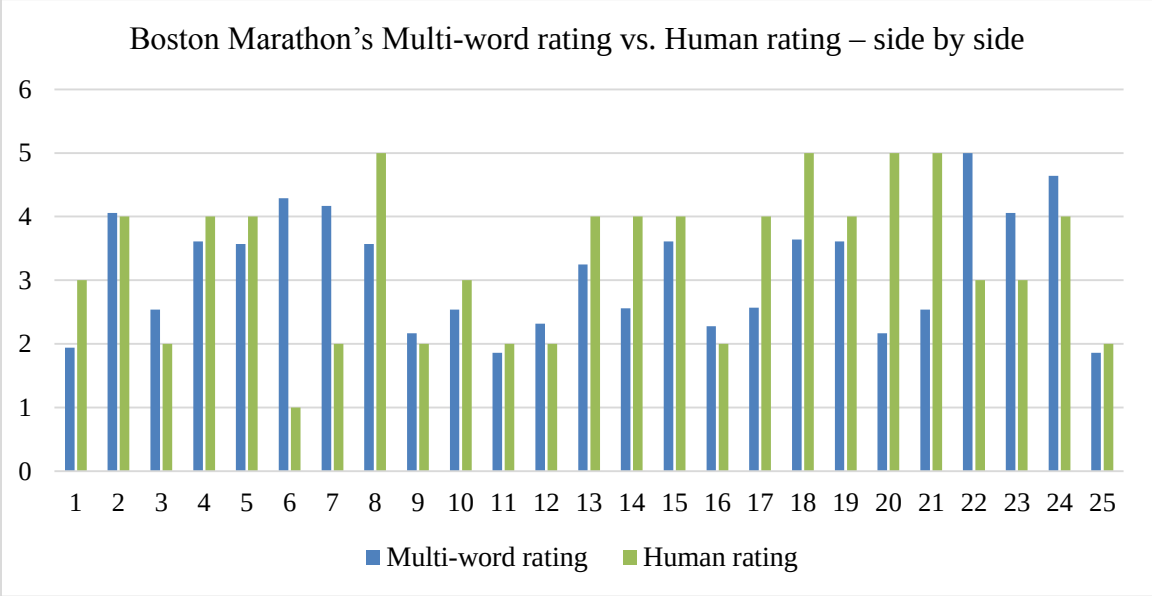


Figure 9: Sample of multi-word rating vs. human rating

The multi-word association process is a great addition to the overall look up of the rating process. At this time, a limited amount of words for this multi-direction look up is used. As more words will be added to find multi-directions, better results can be expected in the future.

Even though the multi-word association process offers a lot more insight into a SM post and its sentiment, this research is looking further into analyzing a SM post. In the next section, the word weight factors sentiment model will be discussed.

C. Word weight factors sentiment model

In the word weight factors sentiment model, any word may or may not have the same static numeric value even though it may be a synonym and have the same default sentiment category value (table 5). Thus, in this model, each word is manually assigned a word weight numeric sentiment value; for example, "*good*" and "*great*" both are positive

sentiment words, but the word "*great*" can be given a higher word weight than the word "*good*."

Table 17: Word weight chart

Word	Sentiment	Word weight
excited	P	5.00
good	P	4.50
great	P	5.00
please	P	3.00
join	P	3.50
us	NU	2.50
praying	P	4.00
injured	N	1.00

A SM post could have words with different word weight strength values. Therefore, each of the numeric values associated with a word from a single SM post is summed together and divided by the sum of words to generate a rating. The following formula (3) shows how the word weight sentiment is calculated:

Numeric rating using word weight factor

$$= \frac{\sum(\boldsymbol{word} \times \boldsymbol{weight})}{\sum \boldsymbol{words}} \quad (3)$$

Even though assigning weight to every word may be a difficult task, this model adds a different dimension to rating models overall. With this model, there is no dependence with static weight sentiment value. Table 18 gives a snapshot of how a word weight range could look, where a positive word could have any numeric value from 3.5 to 5, while a negative word could have any numeric value from 1 to less than 2.5.

Table 18: Word weight range

Sentiment	Range
Positive (P)	3.5 - 5
Negative (N)	1 – less than 2.5
Neutral (NU)	2.5 to less than 3.5

1) Test

In this rating process, each word of a tweet post is viewed against its predefined word weight. Every word that gets processed has a word weight associated with it (table 17). Some words have greater strength than others. If predefined words are not found, a default sentiment is used to find a word weight.

2) Results

Since each word could have potentially different weights, it is possible to acquire different results compared to the other rating models that have been used so far. In figure 10, 12 sample results were compared after running the word weight sentiment process. These sample results show that some of the tweets' ratings improved, while most of these tweet rating results from the word weight model went down in numeric value.

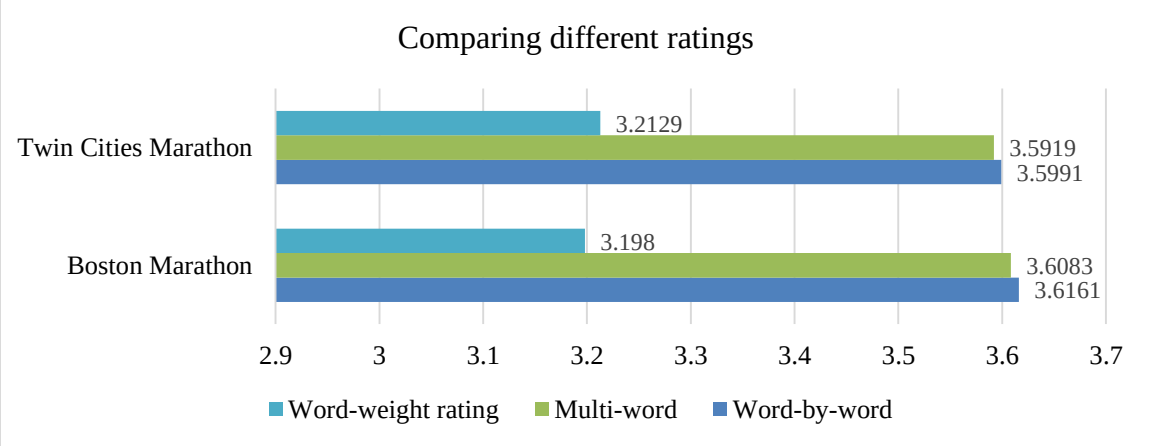


Figure 10: Comparing different rating results

Finally, table 19 shows cumulative word weight ratings. Due to a decrease in numeric rating on each tweet, there is also an overall downward rating for each event. Since the rating looks at the rating of the individual word weight, it is important to have an accurate word weight rating for each word. To improve the results for word rating, further modifying was made to the word weigh-in value of each word.

Table 19: After word weight

Marathon Name	Word-by-word	Multi-word	Word-weight rating
Boston Marathon	3.6161	3.6083	3.1980
Twin Cities Marathon	3.5991	3.5919	3.2129

3) Validation

Comparison of results from the multi-word rating (table 20) reveals that the Boston Marathon rating is still higher for the multi-word rating than the human rating, but the Twin Cities Marathon numeric rating has gone down. The reason for a higher Boston Marathon rating may be due to the process not having enough sample data associated with the word weight rating.

Table 20: Validation after word weight process run

Event name	Word-weight rating	Human rating	Difference %
Boston Marathon	3.1980	3.5549	5.2%
Twin Cities Marathon	3.2129	3.5870	5.5%

4) Discussion

After looking at the detailed results from human and word weight ratings, the absolute difference is still less than the 10% error range. Overall, the actual results match well with expected results, but there is still a lot of discrepancy for line-by-line ratings.

D. Unified rating model

The unified rating model is the final and core model for this research, which consists of a model of models. Thus far, three different ways of looking at a single SM post were discussed. Even though each model has different ways to look at SM data, each model is built with a vision to create a single unified (figure 11) rating model. The previous three models work together to contribute to build an aggregate rating of an event.

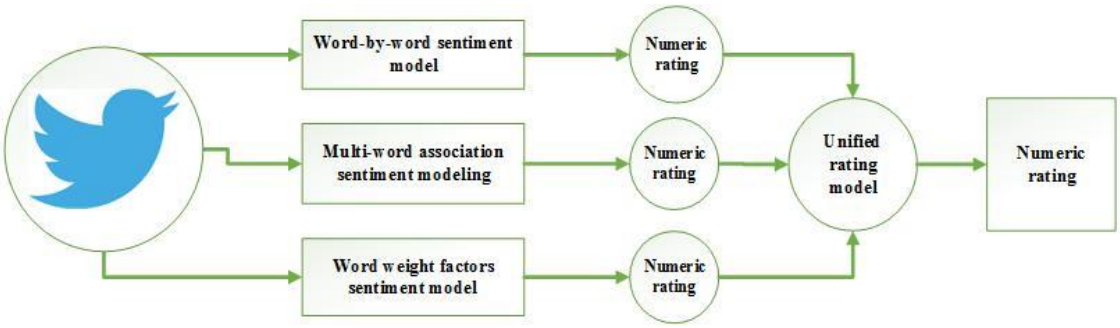


Figure 11: Unified rating model

Since most of the hard work is done by the previous models, in this model an average sum of each SM post's numeric value was taken. This cumulative numeric value is generated using the following formula (4):

Rating model=

$$\frac{\sum(\boldsymbol{Word\ by\ word\ rating + multi\ word\ association\ rating + word\ weight\ factors\ rating})}{\sum(\boldsymbol{Number\ of\ sentiment\ model\ used})} \quad (4)$$

1) Test

This is the core and final processing model where different modeling data were brought together to create a unified model for an event using Twitter data. As of this research paper's composition, this model consists of 3 different processing models to review a SM post where each SM post was sliced and diced into 3 different dimensions to find its rating. As described in the previous sections, each of these models have their own way of looking at a SM post. Since most of the work was done by the previous processes, in this process the average sums of the word-by-word rating, the multi-word rating, and the word weight rating results were taken to get a final rating of each tweet post. Finally, the average sum of all SM posts from each event creates a final rating for an event.

2) Results

In these results (table 21), the average sum of 3 different events were taken to produce a final rating result for each SM post.

Table 21: Unified rating model's results

Event Name	Rating Models used	rating
Boston Marathon	3	3.4742
Twin Cities Marathon	3	3.4680

3) Validation

The validation process still produced a cumulative rating for both the Boston and Twin Cities Marathon that is less than a 2% (table 22) difference between the human and process ratings. This is much less than the 10% margin. Thus far, these modeling technique looks promising.

Table 22: Comparing unified model results vs. human rating

Event name	Process rating	Human rating	Difference in %
Boston Marathon	3.4742	3.5550	1.15%
Twin Cities Marathon	3.4680	3.5870	1.69%

4) Discussion

Further review of the Boston Marathon's line-by-line items with the overall rating reveals that more than 47% of data is above the standard 10% threshold between the human rating and the computer process rating. Even though it is a drop from the previous rating, still these are very high percent values that do not match.

These rating models are getting better as computer processes improve and more word sentiment categories are added as well as develop new ways to review SM data.

V. CONCLUSION

Creation of three models to break SM posts into small units of words, multi-words, and word weight to understand sentiment of a SM post shows that the rating

model's result is getting closer to providing an understanding of the sentiment of a SM post by using computer-generated processes. In this approach, SM data are looked at beyond current trend and social experience.

Furthermore, this research is built in a truly interdisciplinary manner to connect the multidiscipline of big data computation processing, social networking, sports, event, linguistics, etc. As these models and processes mature, this idea of event rating can be used for any event in an almost real time manner.

Due to many known and unknown variables, there will always be misunderstandings regarding true human sentiments vs. computer-analyzed sentiments of a SM post. In such cases, an individual SM post rating may vary between the human rating and the process rating, while the overall rating results are within standard threshold. These differences are due to the ways of looking at a SM post using multiple modeling techniques and dimensions.

This research is successfully able to offer a model of models to capture and analyze SM data to produce meaningful information. Initially, this research is able to achieve its goal of producing a numeric rating utilizing SM data by rating two different events (table 20).

Despite its successes, there is still a lot of work remaining to complete the model of models. This rating model concept is not invalid, but it needs further improvements.

VI. FUTURE WORKS

In the future, there needs to be ways to bring in other part of a SM post, such as hashtags, emoticons, images, etc., to complete the users' sentiments model. Also to further validate these processes, more marathon events will be evaluated.

Since word sentiment category and word weight are important parts of the overall rating models and processes, there needs to be ways to automate weighting words, categorizing sentiment, creating rating matrixes, and finding more predefined words with associated sentiments.

VII. APPENDIX

A. Appendix I

For the validation process, sample sets of multiple data from 10 different individuals were sent with combined experience of more than 50 years of social media and more than 50 years of running. Even though these are not the same people who posted these SM posts, having humans to review helped to validate the computer process results. Each individual was asked to read each sample Twitter post and rate them 1-5 according to the sentiments toward a marathon.

REFERENCES

[1] X. Hu, J. Tang, H. Gao, and H. Liu, "Unsupervised Sentiment Analysis with Emotional Signals," pp. 607–617.

[2] S. Silwal, "Using Social Media Data as Research Data," vol. 1, pp. 49–55, 2013.

[3] C. Wojtalewicz, "Social Media Use for Large Event Management," no. 1, pp. 24–29, 2012.

[4] D. Contractor and T. Faruquie, "Understanding Election Candidate Approval Ratings Using Social Media Data," *... 22nd Int. Conf. ...*, pp. 189–190, 2013.

[5] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, and N. Schneider, "Part-of-Speech Tagging for Twitter: Word Clusters and Other Advances," 2012.

[6] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter Sentiment Analysis : The Good the Bad and the OMG !," *Artif. Intell.*, pp. 538–541, 2011.

[7] T. Baldwin, “Social Media : Friend or Foe of Natural Language Processing ?,” pp. 58–59, 2012.

[8] B. Liu, “Sentiment Analysis and Subjectivity,” *Handb. Nat. Lang. Process.*, pp. 1–38, 2010.

[9] B. Liu, “Sentiment Analysis and Opinion Mining,” *Synth. Lect. Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, May 2012.

[10] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, “Finding High-Quality Content In Social Media,” *Proc. Int. Conf. Web search web data Min. - WSDM ’08*, p. 183, 2008.

[11] S. Silwal and D. W. Callahan, “Building a Social Media Rating Model,” IEEE *Southeastcon 2014*, pp. 1–3, Mar. 2014.

[12] A. Odusd and T. Sse, *Systems Engineering Guide for Systems of Systems*, no. August. 2008.

[13] J. Weng, E.-P. Lim, Q. He, and C. W.-K. Leung, “What Do People Want in Microblogs? Measuring Interestingness of Hashtags in Twitter,” *2010 IEEE Int. Conf. Data Min.*, pp. 1121–1126, Dec. 2010.

[14] T. Ahlqvist, A. Bäck, M. Halonen, and S. Heinonen, “Social Media Roadmaps,” *Helsinki Ed. Prima Oy*, 2008.

[15] F. Kivran-swaine and M. Naaman, “Network Properties and Social Sharing of Emotions in Social Awareness Streams,” *Media*, no. July, pp. 379–382, 2009.

CAN AN EVENT BE RATED USING HASHTAGS, EMOTICONS, IMAGES, AND URLs?
RATING AN EVENT USING SOCIAL MEDIA DATA

by

SUMAN SILWAL and DALE W. CALLAHAN

International Handbook of Academic Research and Teaching, Volume 40

Summer 2015 Nashville, TN

Reprinted, with permission, from the Proceedings of the Intellectbase International Consortium Conference

Format adapted and errata corrected for dissertation

CAN AN EVENT BE RATED USING HASHTAGS, EMOTICONS, IMAGES, AND URLs?
RATING AN EVENT USING SOCIAL MEDIA DATA

Suman Silwal
University Of Alabama at Birmingham
Birmingham, AL
ssilwal@uab.edu

Dr. Dale Callahan
University Of Alabama at Birmingham
Birmingham, AL
dcallahan@uab.edu

Conference: *Nashville, TN*
Paper Category: *Full Paper*
Track: *Science and Technology*

ABSTRACT

Hashtags, emoticons, images, and URLs are parts of everyday Social Media (SM) posts. These different types of SM data can provide insights into users' sentiments. Further, each piece of SM data can add values to a SM post. A Social Networking Site(s) (SNS) such as Twitter.com allows only 140 characters per post. Despite the limited space for expression, SNS contributors are finding alternative ways of expressing their thoughts and sentiments without writing a full sentence by using these datatypes.

The Twitter SNS revolutionized how Hashtags (#) are used in SM. Hashtags are becoming an indication of brand, a symbol of hope, a quick text index technique, and a search tool. Hashtags such as *#BostonStrong* are widely used and have become the symbol of hope and unity after the 2013 Boston Marathon bombing. As more and more

people are using Hashtags to express their thoughts and feelings in creative ways, Hashtags' popularity and uses are here to stay.

It is also important to note that today's emoticons have gone from simple happy faces to complicated icons. Regularly, a SM contributor uses emoticons to express emotions and sentiments beyond words. Thus, to understand someone's SM post, emoticons must be examined.

In recent years, most SNS allow posts of images. Beyond words, images can also be used to express SM posts' sentiments. Images from an event such as a marathon are posted in a real time manner. These images can be evaluated to find user sentiment.

In this research, each event-related SM post is processed and evaluated using rating models to generate a numeric rating of the event. Further, Hashtags, emoticons, images, URLs, etc. are utilized to create a unified SM model of models to generate the rating of a SM post.

***Keywords*:** *Social Media(SM), Social Networking Sites (SNS), Rating System, Sentiment Analysis, Rating, Twitter, Hashtag (#), Marathon Rating, Event rating, Emoticons, URLs, Hashtags rating, Emoticons rating*

INTRODUCTION

In previous research, word-based rating models were discussed to rate an event as well as ways to import data from the Twitter SNS (Silwal & Callahan 2015). Imported data and results generated from previous models will continue to be used for this research to compare and generate unified models.

In this research, different aspects of SM posts will be reviewed. Also, three different rating models to further understand a SM post will be developed.

In recent years, many research studies have been conducted in the area of using Hashtags (Wang et al. 2014) (Mohammad 2012)(Weng et al. 2010), emoticons (Hu et al. n.d.), and images (Images & Analysis 2013) to find user sentiment. As the popularity of SM use is growing, so are the ways to find the sentiments of a SM post.

This research is not only looking at sentiments of SM posts but also providing numeric ratings for those posts. Due to many reasons, it is always difficult to understand 100% user sentiment of a SM post. At this time, sentiment validation within a 10% margin of error is acceptable.

In this research, Hashtags, emoticons, URLs, the popularity of a SM post, images, and event organizers' posts will be reviewed and discussed as a part of creating a rating model to rate an event. This approach will provide a deeper look into SM posts. Expanding the rating beyond text data will help to increase understanding regarding the sentiments of a SM post to rate an event.

Here are some of the core systems, events, and frameworks that are used to build a user rating models and processes:

- **Twitter.com** (Twitter) is the Social Network Site (SNS) of choice that provides limited access to its publically available data on its developer network through its Application Programming Interface (API) (Weng et al. 2010).
- **Marathon** events are 26.2-mile foot races that are considered the research event topic.
- **Java programming language** is an open-source computer programming language widely used for application development.

- **Spring java framework** provides a lot of different components to build a very powerful application including Spring Social API. It is used to glue the research application processes together.

The upcoming sections of this paper will be broken down into 4 different parts to build a rating model:

1) **Dictionary** building process consists of creating processes to generate Hashtag and emoticons dictionaries.
2) **Sentiment Modeling** section defines different possible modeling techniques needed to build a rating system.
    a. During the testing section, a process to generate numeric rating SM data will be created.
    b. During the results section, results generated by testing process will be reviewed.
    c. During the validation section, human rating vs. computer rating results will be compared.
    d. During the discussion section, the relevance of this research model will be discussed.
3) **Future SM modeling** ideas will also be discussed.
4) **Conclusion and Future Works.**

DICTIONARY BUILDING

The previous research publication (Silwal & Callahan 2015) described how to build a dictionary to store text data and provide sentiment, word weight, and multi

direction of those words (Silwal & Callahan 2015). To further expand this research to evaluate a SM post and acquire user ratings, two distinct dictionaries were created. Creating a dictionary from unstructured SM data can provide some kind of structure.

Similar to prior dictionary building processes, each word or symbol in the dictionary can also be clustered into a positive (P), negative (N), neutral (NU) or not applicable (NA) sentiment category.

Hashtag dictionary

Since the research goal is to rate a yearly event, a Hashtag trend can be different event to event as well as year to year. For the Hashtag dictionary building process, the year a Hashtag is used to identify its trend is saved into the Hashtag dictionary table. For this model, Hashtag #BostonMarathon, is ignored because it was used during the Twitter data collection process. This exclusion prevents uneven ratios of trend information. The Hashtag dictionary building process (figure 1) has three phases.

In the first phase, a tweet post is split into many different words. To build a Hashtag dictionary, only Hashtag words are retrieved and stored into a temporary Hashtag table. Once all data is stored into the temporary Hashtag table, a further process was run on those Hashtag data to find trend count information, event code, and the year it was used. This information is stored within the Hashtags dictionary table (figure 2).

In the second phase, each Hashtag is looked up against the sentiment dictionary, which was built as a part of the previous research (Silwal & Callahan 2015). Since these Hashtags are still text data, they may already have been captured into the sentiment dictionary table. If Hashtags are found, an update is made to the Hashtag dictionary with the sentiment from the sentiment dictionary.

In the third phase, a manual process is completed to assign the positive (P), negative (N), neutral (NU) or not applicable (NA) sentiment category to each Hashtag that did not have a sentiment assigned to them.

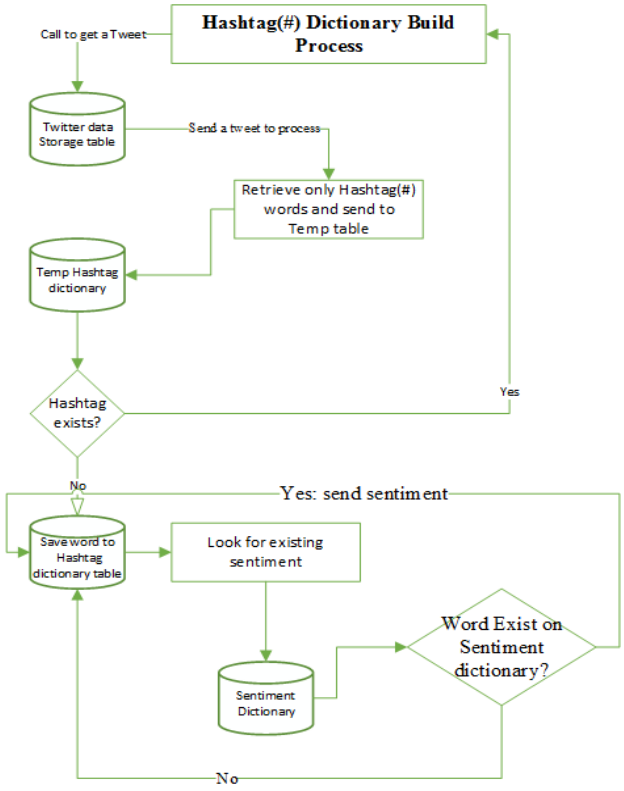


Figure 1: Hashtag dictionary building process

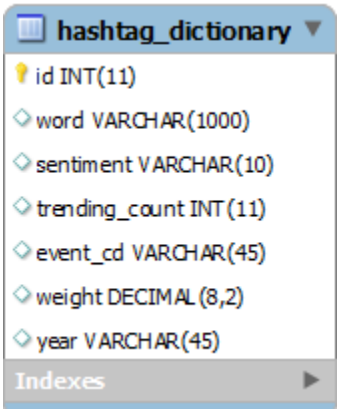


Figure 2: Shows all different attributes listed in a Hashtag table

Emoticon sentiment dictionary

Emoticons provide different ways to express users' sentiments. They are, consequently, a common feature on microblogging sites (Kouloumpis et al. 2011). Over the years, the popularity of emoticons have been growing rapidly. Simple emoticons from the early computer age have grown into many new and complex emoticons. A SNS such as Twitter has limited characters per post. Thus, emoticons also provide a quick and easy-to-use way to further express thoughts and sentiments of a SM post.

In this research, emoticons are stored into the emoticon dictionary for current and future research uses. At this time, these emoticons are manually searched and retrieved through different sources such as prior-used knowledge, websites, and SM data posts. These emoticons are categorized into positive, negative or neutral sentiments, which have a default numeric rating value (table1).

Table 1: Default sentiment indicators numeric value

Sentiment Indicators	Descriptions	Numeric Rating
P	Positive	5
N	Negative	1
NU	Neutral	2.5

Table 2 lists some of the emoticons and their sentiment in the emoticon sentiment dictionary.

Table 2: Emoticons data

Emoticon	Sentiment
:-)	P
:)	P
:D	P
♥	P
:-|	NU
:|	NU
:-(	N
=(	N

BUILDING A RATING MODEL USING SM DATA

In this section of research, different user rating models used to rate SM posts will be discussed. With the added models to rate SM data, the primary goal of this research is getting closer to reality.

Hashtag (#) sentiment model

A Hashtag symbol "#"is used, before a word or phrase in SNS such as Twitter.com, Facebook.com and Instagram.com for categorizing, trending, and searching (Weng et al. 2010). It can be also used to express sentiment and emotions (Mohammad 2012). The popularity of using Hashtags is growing. People are finding clever ways to express their thoughts and emotions by using Hashtags. Importantly, Hashtags are also used for searching, filtering, and retrieving data from Twitter.com.

For the Hashtag sentiment model, each SM post is cataloged into positive, negative, neutral, and not applicable sentiments. Since this research is based on rating a specific event, each Hashtag used during collecting data is not a candidate for this model. For the Example 1 SM post, #BostonMarathon was used for searching and filtering to collect Boston Marathon SM data, so it would not be a candidate for this rating model because it could create an uneven trend.

Example 1 SM post: "Good luck to all running the #BostonMarathon, sending our love from #KC! #BostonStrong"

In this model, a higher rate of use of a Hashtag for a given dataset trumps a higher ranking. A highly ranked Hashtag could be a negative or positive sentiment. Therefore, this model combines sentiment indicators such as positive, negative, and neutral with a Hashtag trend for a given year and event to find the sentiment rating of a SM post. With

the growing use of Hashtags to express sentiments, thoughts, and ideas, this rating model adds value to the overall rating of a SM post (table 3).

Table 3: Hashtag words

Hashtag Words	Sentiment Value	Hashtag Trend	Event Name	Year
#BostonStrong	P	581	Boston Marathon	2014
#Boston	P	88	Boston Marathon	2014
#GoNatalieGo	P	44	Boston Marathon	2014
#WeRunTogether	P	35	Boston Marathon	2014
#Eid	NA	18	Twin Cities Marathon	2014
#WhatWillIMissWhenWWATEnds	P	18	Twin Cities Marathon	2014
#BostonMarathonBombing	N	5	Boston Marathon	2014

Hashtag trend counts are used to determine different rating values so that the higher the trend of a Hashtag for a given year and event, the higher the rating for that particular Hashtag. Table 4 presents the frequency vs. rating numeric values matrix.

Table 4: Frequency vs. rating numeric values matrix

Frequency of use	Rating numeric values
500 or greater	1.5
100-499	1
50-99	.75
25-50	.5
5-25	.25
1-5	.1

For this model, sentiment numeric values are adjusted from prior default values so that the Hashtag frequency matrix from table 4 and table 5 could be used.

Table 5: Hashtag sentiment value

Sentiment	Numeric Values
Positive	3.5
Negative	1.5
Neutral	2.5

The weight of a Hashtag can be determined by using the following formula (1) where each Hashtag's sentiment numeric value from table 3 is combined with the frequency of use matrix from table 4.

Formula for Hashtag weight =

Hashtag sentiment numeric value + Frequency of use (1)

Based on the above formula (1), *#BostonStrong* will receive a numeric rating of 5, while *#MarathonMonday* will receive only a 3.75 rating.

Finally, a SM post is rated using the following formula (2) for a Hashtag's rating system in a post.

SM post numeric rating using Hashtag

$$= \frac{\sum(\boldsymbol{Hashtag\ weight})}{\sum \boldsymbol{Hashtags}} \quad (2)$$

Test

To test the Hashtag sentiment rating, a sentiment rating process was built. In this process, only a Hashtag of each tweet was examined. For this process, non-Hashtag words and symbols were ignored. Hashtags used during the data retrieval process for data searching and filtering were ignored as well.

Through a trending count, some Hashtags are used more frequently than others for an event. For this process, a frequency of use rating idea to determine the overall sentiment value of a tweet post was implemented.

During this Hashtag sentiment rating process, initially the Hashtag dictionary table used to find each Hashtag's sentiment and trend count. Using the frequency matrix (table 4) allows Hashtag tweets to be processed and receive numeric results (table 6) from SM data.

Table 6 Initial run of Hashtag rating process

Event Name	Word-by-word	Multi-word	Word-weight rating	Hashtag Rating
Boston Marathon	3.6161	3.6083	3.1980	3.6443
Twin Cities Marathon	3.5991	3.5919	3.2129	2.7711

Results

By comparing the results, the Boston Marathon rating was consistent with the previous rating results, while the Twin Cities Marathon rating dropped almost a point from previous ratings. Further review reveals that this drop was due to an uneven volume of tweet data for these two test marathons. Thus, the frequency of use matrix was also changed to account for volumes of data processed for Hashtag ratings. Therefore, one more frequency matrix (table 7) was added to account for lower volume datasets.

Table 7: Updated frequency of use matrix

Frequency of use	Rating numeric values
50 or greater	1.5
25-49	1
10-24	.75
5-10	.5

2-4	.25
1	.1

After the reprocessing of SM data using the new frequency of use matrix, table 8 shows improved results from the previous rating process.

Table 8: Hashtag rating model process results

Event Name	Word-by-word	Multi-word	Word-weight rating	Hashtag Rating
Boston Marathon	3.6161	3.6083	3.1980	3.99
Twin Cities Marathon	3.5991	3.5919	3.2129	3.0529

Validation

Similar to prior research, validating results from these rating systems is an important part of this model. Since the rating model uses a computer process, there will always be misunderstandings between human speech and computer processes' translation of such speech. During the validation process, the computer-generated rating and the human rating (Appendix I) for only Hashtag data were compared. Table 9 shows results for the Boston Marathon. The rating results are higher than expected.

Table 9: Validation after Hashtag rating process run

Event name	Hashtag Rating	Human rating
Boston Marathon	3.99	3.55

Discussion

As more and more people are using Hashtags to communicate, Hashtags are becoming a common tool for searching and indexing. Hashtags will always be an important part of any SM research. Review of two different results from the rating results (table 9) show that having a correct volume and frequency of use is an important part of this model calculation.

There is still a big difference between human ratings and process ratings. After further discussion with the human rater, it was discovered that the human rater was not only focusing on Hashtags like this model but also whole tweet posts to make sense of it. For the purpose of research and building a Hashtag rating model, focusing just on Hashtags is a valid way to look at SM data.

Also due to the volume of SM data from event to event, the frequency of use cannot be a fixed value. To further improve this model, it needs to continuously evaluate the volume vs. frequency of use ratios.

Emoticon sentiment model

Beyond words and Hashtags, emoticons are widely used to express positive, negative, and neutral sentiments (Maynard et al. 2012) within a SM post. In recent years, a lot of research has been conducted in the area of emoticon sentiment understanding (Hu et al. n.d.).

In recent years, a lot of new emoticons were developed (table 10) and used to better express thoughts and sentiments of a SM post (example 2). They can be a very powerful tool in determining the sentiment of a SM post.

Table 10: Emoticons

Emoticon	Sentiment
:-) :) :D :o) :] :3 :c	Positive
>:[:-(:(:-[:[:	Negative
:-| :| >:\ >:/ :-/ :-. :/	Neutral

Example 2 SM post: :) Awesome job you are doing xoxo #runJoeyrun!! #bostonmarathon2014

In this model, each valid emoticon analysis adds value to the overall sentiment modeling process. In this emoticon sentiment model, each emoticon can be captured and analyzed to make better sense of its sentiment value. Similar to other sentiment modeling processes, each emoticon is also categorized into positive, negative, and neutral categories.

The Emoticon and Sentiment Matrix (table11) gives a better understanding of how a positive sentiment of a SM post with a negative emoticon can bring the overall sentiment rating value of a post down, while a negative post with a positive emoticon can bring the overall sentiment rating value of a post up.

Table 11: Emoticon and sentiment matrix

	Positive sentiment	Negative sentiment	Neutral sentiment
Positive emoticon	Up	Up	Up
Negative emoticon	Down	Down	Down
Neutral emoticon	No change	No change	No change

Example 3 SM post: Screaming & dancing for 3 hrs almost non stop is EXHAUSTING but fun. Woman said I was the best cheerleader she's ever seen. :) #HouMarathon

The Example 3 SM post can be rated 3.5-3.6 using one of these modeling techniques – without an emoticon. Including a positive emoticon on the SM post can allow post's numeric rating to go up.

Finally, a SM post with an emoticon is rated using the following formula (3).

SM post numeric rating with emoticons =

$$\frac{\sum(\boldsymbol{Emoticon's\ numaric\ data\ in\ a\ SM\ post})}{\sum \boldsymbol{number\ of\ emoticons\ in\ a\ post}} \quad (3)$$

Testing

Similar to the Hashtag sentiment rating, this model only looks at emoticons of SM posts. In this process, each tweet that has one or more emoticon is viewed against an emoticons dictionary. The dictionary has emoticons cataloged with positive, negative, and neutral category sentiment values. Similar to the previous models, the default numeric value approach of assigning 5 for positive sentiment, 2.5 for neutral sentiment, and 1 for negative sentiment was used.

Results

Table 12 shows a few sample tweets' results after processing of the emoticon rating. Due to the negative nature of the first tweet sample, it was not evaluated correctly by previous rating models. However, the emoticon rating model process was able to pick up the negative sentiment of this post.

Table 12: After processing emoticon rating

Tweet	Word-by-word rating	Multi-word rating	Word weigh rating	Hashtags rating	Emoticons rating
Watching #Bostonmarathon on SS 207. Camera work is crap, directing sucks, too much sideshows. Making it hard to enjoy the race :-(	3.60	3.60	3.09		1.00
Listening to the tv while I paint, I just want to run the #BostonMarathon but my leg surgeon would kill me! :-/	3.17	3.17	2.75		2.50
Ducks in Boston Commons are running too :) #BostonMarathon #BostonStrong http://t.co/7Hm9wBiaVv	3.33	3.33	3.06	5.00	5.0
Good luck to everyone running the #BostonMarathon today. May you	4.17	4.17	3.23	5.00	5.0

all be fleet of foot. :-) #BostonStrong http://t.co/qpYMiqyKzn

Table 13 shows the results of overall event ratings after processing of emoticon sentiments. Even though the ratio of emoticons used during this rating process is very small compared to the overall volume of data for sample events, it is important to include the emoticon rating for the overall rating of an event.

Table 13: After emoticon rating

Event Name	Word-by-word Rating	Multi-word Rating	Word-weight rating	Hashtag Rating	Emoticon Rating
Boston Marathon	3.6161	3.6083	3.1980	3.7815	4.9074
Twin Cities Marathon	3.5991	3.5919	3.2129	3.0529	4.9582

Validation

To validate the emoticon rating, 100 tweets were rated by human evaluators manually. Comparing the results from the human rating vs. the process rating (table 14) show that the human rating is much lower than the process rating. It is similar to the Hashtag rating, where human raters are not only looking at emoticons but also the context of a tweet. In contrast, this process only focused on emoticons.

Table 14: Process vs. human rating after emoticon run

Event name	Emoticon Rating	Human rating
Boston Marathon	4.53	4.16

Discussion

Even though the overall emoticon rating is higher than expected, this model brings additional dimensions to overall SM modeling. An insufficient amount of emoticons in the sample dataset as well as in the emoticon dictionary makes this model unreliable at this time. However, the emoticon rating should always be a part of the overall rating process. The future work of this model should include improvement of collecting and processing more emoticons.

Event organizers sentiment model

In this age of SM, every event organizer has his or her own SNS account (Silwal & Callahan 2013) to post up-to-date event information, updates, announcements, etc. Event organizers frequently update their SNS during, before, and after an event. Table 15 shows some of the event names with their Twitter handle user ids.

In this model, event organizers' Twitter handle user ids are reviewed to filter their SM posts. Once a SM post of an event organizer is found, a constant numeric value of 5 is assigned to that SM post. In this model, the assumption is made that a SM post from any event organizer is mostly positive.

Table 15: Event name and its twitter handle id

Event Name	Event Twitter handle id
7 Bridges Marathon	7BMarathon
Boston Marathon	Bostonmarathon
Chicago Marathon	ChiMarathon
Mercedes Marathon	Run_Mercedes
OK City Marathon	OKCMarathon
Richmond Marathon	RichmondMarathon
St. Jude Marathon	StJude
Twin Cities Marathon	Tcmarathon

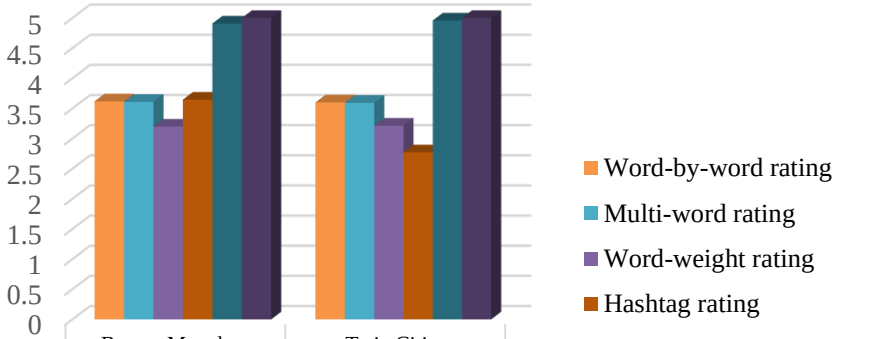
In most cases, the ratios of SM posts from event organizers are much lower than that of event goers. At this time, it is safe to assign a higher rating for a SM post generated by an event organizer.

Test

In this event organizers' sentiment rating process, all tweets generated from an event organizers' tweet handle are given a positive rating with a numeric value of 5. Even though tweet posts may acquire different ratings from other processes, this rating process assumes that event organizers' tweet posts are mostly positive and each tweet receives a standard numeric rating of 5.

Results

Figure 3 shows the results after processing of the organizers' tweets. These tweets were rated differently by other processes, but when this process was run, the rating of each tweet with organizers' user ids were assigned a numeric value of 5.



	Boston Marathon	Twin Cities Marathon
Word-by-word rating	3.6161	3.5991
Multi-word rating	3.6083	3.5919
Word-weight rating	3.198	3.2129
Hashtag rating	3.6443	2.7711
Emoticon Rating	4.9074	4.9582
Event organizers ' rating	5	5

Figure 3: After event organizers rating

Validation

Similar to previous models, this model is also validated by using the process vs. the human rating results. As in prior cases, the human rater rated the SM post as a whole, while this process gives a standard 5.

Table 16: Process vs. human rating after process run

Event name	Event organizers	Human rating
Boston Marathon	5	4.16
Twin City Marathon	5	4.07

Discussion

Looking at the results in table 16, there is a big difference between how this process rated vs. how the rating is done by a human rater. After comparing these results, an observation was made that all SM posts by event organizers cannot have a constant value.

After reviewing all event organizers' tweets and finding a big difference in results, a conclusion was made that this model cannot be used as standard rating model. However, it can be used as an optional rating model.

FUTURE SM MODELING IDEAS

In the next few sections of this chapter, some of the models will be discussed as a part of the future expansion of the current rating models. The following models add a lot of value for understanding the overall rating of a SM post.

Image sentiment model

With the revolution of smart phones in recent years, the phone has become a minicomputer with many applications (apps), including camera and SM apps to upload photos on SNS. Most SNS allow photo upload capabilities to its sites via phone

applications (app). Due to accessibility and ease of use, an event can generate numerous photos with or without any textual information. Also, a SM post can have one or more images associated with it. With the use of advanced photo recognition technology, the sentiment of a SM post including a photo can be identified (Yuan et al. 2013).

In figure 4, a runner is very excited to be at the finish line of a race. The sentiment expressed by the image can be rated 4.5-5.

Figure 4: A runner at the finishing line

This model needs to account for facial expressions, body expressions and postures, eye expressions, etc. to understand the sentiments expressed by a person. Similar to a dictionay for text, Hashtags, and emoticons, there shall be ways to collect and store common expressions, ways that can help this model to recognize the sentiment of an image quickly.

Further in this model, every aspect of an image, including background, should also be considered to grasp the overall sentiment value of a whole image (figure5). Also, a SM post image can have more than one person. To find the sentiment of an image, sentiments expressed by more than one person need to be considered as well.

Figure 5: Image of a runner looking at the finish line of the Boston Marathon for the first time

Further in this model, finding the actual sentiment of an image needs to be a research topic of its own. Once the sentiment of an image is recognized, a process can put it into one of the standard sentiment categories positive, negative, and neutral with numeric values of 5, 1, and 2.5 respectively. Also, a SM post can have one or more images associated with it. If a post has multiple images associated with it, it should be considered as part of the overall sentiment building process.

The following formula (4) provides the numeric rating of a SM post with images.

Numeric rating using Image sentiment model =

$$\frac{\sum(\textbf{Posive Images})\times 5+\sum(\textbf{Negative Image})\times 1+\sum(\textbf{Neutral Image})\times 2.5}{\sum \textbf{Positve Image}+\sum \textbf{Negative images}+\sum \textbf{Neutral Image}} \quad (4)$$

Overall, event participants uploading more and more photos as SM posts to express their feelings before, during, and after an event can provide another dimension to understanding the overall sentiment of a SM post.

Popularity sentiment model

Many SNS allow its users to repost someone else's SM post as retweets, likes, shares, etc. In this model, the popularity of a SM post is reviewed and analyzed. In the universe of SNS, certain SM posts are receiving more attention as well as are being shared and discussed more than others. In this model, the popularity of SM posts using different popularity indicators aspects is examined. In the following example (figure 6), a SM post by Mercedes Marathon had 11 posts and 14 favorites. The example tweet was a popular post among the Mercedes Marathon's audience.

Figure 6: Tweet post by Mercedes Marathon event

For this popular sentiment model, at first, the sentiment of a post is identified by using the word-by-word sentiment model that was discussed in a previous section. Once the word sentiment is identified, further sentiment reviews need to be done using the popularity sentiment model, in which the more popular SM posts can acquire higher sentiment values.

Table 17: Popularity model outcomes

	Repost count	Sentiment rating
Positive repost	7	More positive

Negative repost	2	More negative
Neutral repost	10	positive

Table 17 shows how sentiment outcomes should look after a SM post with multiple reposts has been processed. Table 18 lists different numeric values that need to be added to a SM post rating. The higher the popularity of a SM post, the higher the numeric rating it can receive.

Table 18: Popularity sentiment model numeric value matrix

Repost counts	Rating numeric values
100 or greater	1.5
50-99	1.0
25-50	.75
5-25	.5
1-5	.25

Finally, the following formula (5) provides the numeric rating of a SM post using the popularity model matrix. As noted in a previous section, since the rating system is 1-5, it will not go higher than 5 and lower than 1.

Numeric rating using popularity sentiment model =

$$\sum \textit{(Word sentiment numeric value + popularity of a post matrix)} \quad (5)$$

This model should be able to provide better numeric ratings for those SM posts that may have been ignored by previous models.

URL sentiment model

A SM post can also have an external link to different websites, other social networking sites, etc. with or without textual data. An external link can provide added insight into the sentiment of a SM post. These external links are a part of SM posts for a

reason; thus, the importance of those external URLs in a SM post cannot be ignored. Therefore, this model is including URLs as a part of the numeric rating sentiment process.

To find the sentiment of a SM post with a link, a two-step approach needs to be taken. First, access will be needed to retrieve the content of the external URL. For the next step, this model needs to understand sentiments of this content.

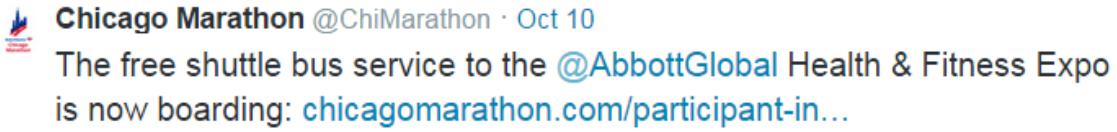


Figure 7: Chicago Marathon's Twitter post with link

This model assumes that the other parts (figure 7) of a SM post have already been evaluated by one of the previous models. Hence, it is only interested in looking at the sentiment of information provided by the URL of the SM post. For now, these external links' sentiments are categorized into positive, negative, and neutral values with 5, 1, and 2.5 numeric values, respectively.

UNIFIED RATING MODEL

The unified rating model is the final and core model for this research, which consists of a model of models. Thus far, 9 different ways to look at a single SM post were discussed. Even though each model has different ways to look at the SM data, each model is built with a vision to create a single unified rating model using SM data.

Figure 8 shows a complete flow of the rating model where the previous 9 models work together to contribute to a unified model.

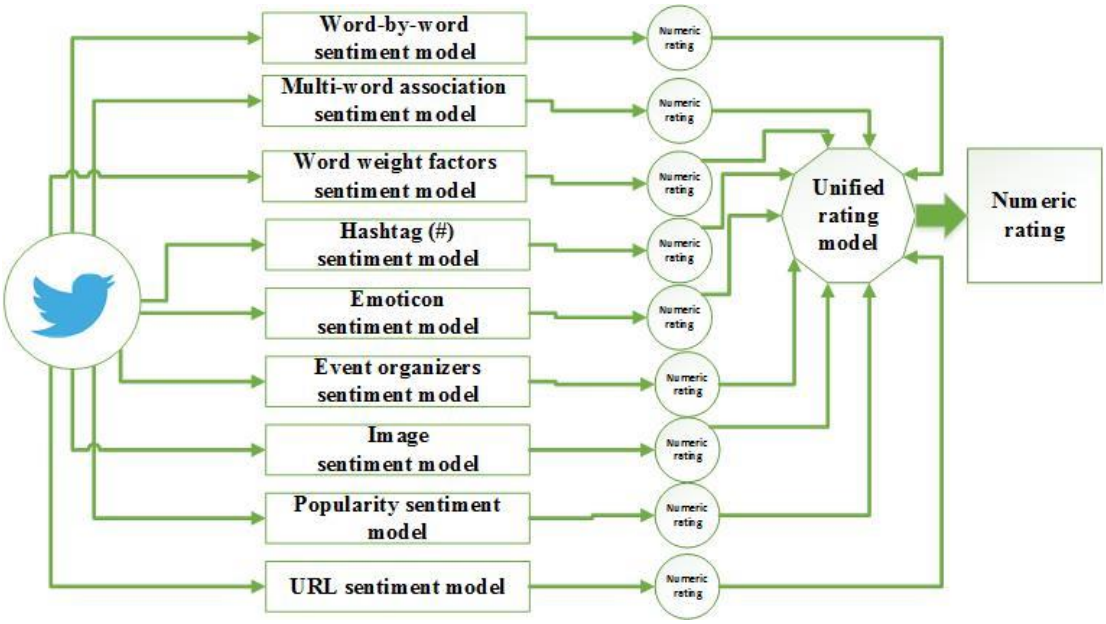


Figure 8: Unified rating model

Since most of the hard work is done by the previous models, in this model, an average sum of each SM post's numeric values were taken. Those values was generated from the previous 9 models using formula 6. The rating generated from this cumulative rating model can create a close to accurate rating of an event.

Rating model=

$$\frac{\sum \begin{matrix}(Word\ by\ word\ rating + multi-word\ association\ rating + word\ weight\ factors\ rating + \\ Hashtag\ rating + Popularity\ rating + Image\ rating + Event\ organizers\ rating + \\ URL\ rating)\end{matrix}}{\sum(Number\ of\ sentiment\ model\ used)} \quad (6)$$

Test

The unified rating model is the core and final processing model, where different modeling data are brought together to create a unified model of an event using Twitter data. As of this research paper's composition, this model consists of 6 different processing models (out of the total 9 models discussed in this research) to review a SM post. As described in previous sections, each of these model has its own way of look at a SM post. Since most of the work was done by previous processes, in this process the

average sums of word-by-word rating, multi-word rating, word weight rating, Hashtags rating, emoticon rating, and event organizers' rating results are taken to obtain the final rating of each tweet post.

Results

Table 19 shows the final results from 6 different rating models. Further review shows that the result outputs from each rating processes could not be equally divided into 1/6 ratios. Figure 9 shows how dividing all ratings can create an uneven rating for SM posts because the volume for different types of rating models are not equal.

Table 19: Unified rating model results

Event Name	Rating Models used	Rating
Boston Marathon	6	3.9832
Twin Cities Marathon	6	3.8907

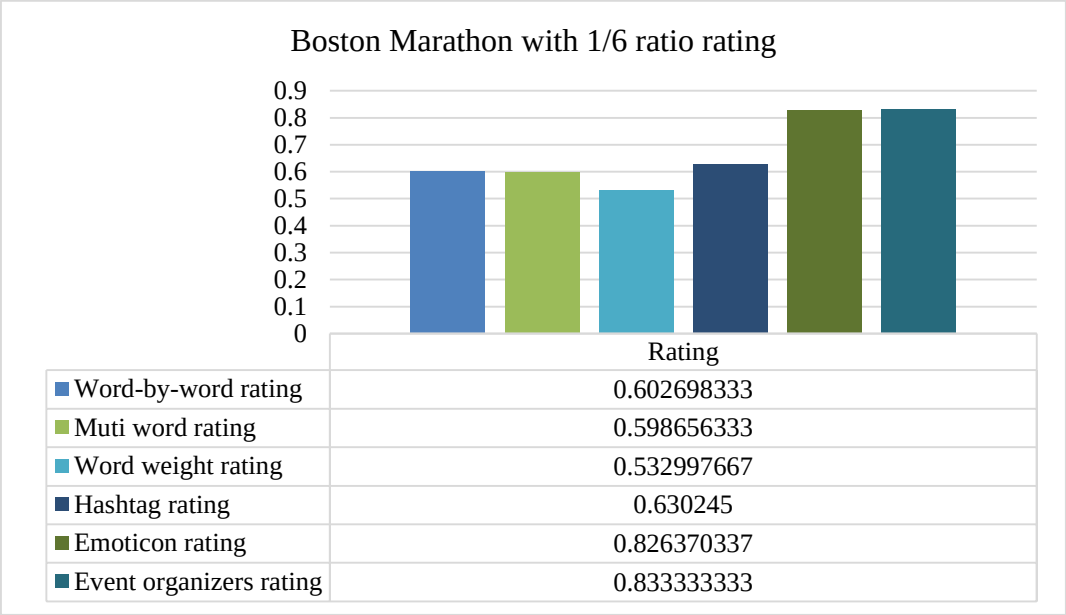


Figure 9 Unified rating with 1/6 ratio rating

To resolve this issue of obtaining uneven ratings, a matrix was created where each rating model has different ratios toward the overall rating. Table 20 shows rating models and calculation ratios. Each of these percentages is multiplied to corresponding average sum ratings that were generated by the previous processes.

Table 20: Rating calculation matrix

Rating Model	% of whole rating
Word-by-word rating	25%
Multi-word rating	25%
Word weight rating	20%
Hashtag rating	20%
Emoticon rating	8%
Event organizers' rating	2%

The updated formula (7) has a better representation of distributed values vs. ratios of reviews done for each rating model.

Rating model=

$$\sum (\textit{word by word rating}*0.25+\textit{multiword rating}*0.25+\textit{word weight rating}*0.20 + \textit{Hashtag rating}*0.20+\textit{Emoticon rating}*0.08+\textit{Event organizers rating}*0.02) \quad (7)$$

Table 21 shows results after rating model parameters were readjusted; these parameters are lower than the one from table 19. Readjusting the value of SM posts has a better outcome.

Table 21: Rating model output after readjusting parameters

Event Name	Rating Models used	Rating
Boston Marathon	6	3.7598
Twin Cities Marathon	6	3.5510

	Boston Marathon	Twin Cities Marathon
Word-by-word	0.9040474	0.89979955
Multi-words	0.90209663	0.89798441
Word weight	0.63959727	0.64258352
Hashtag (#)	0.81744989	0.70690667
Emoticon	0.396657762	0.4
Event organizers'	0.1	0.1

Figure 10: Unified rating results with % of ratios

Based on the information in figures 9 and 10 compared side by side, the outcomes of these two processes are much different. Figure 10 clearly shows how data gets readjusted as results using percentage ratios vs. dividing everything by even ratios of 1/6. In the future, as more ways to review a SM post will be added, the rating calculation matrix from table 17 will be modified. For now, these percentage matrixes are assigned according to their volume counts and results that represent those data. These percent assignments are done manually. As a part of future works, computer algorithmically should generate these percentages ratios according to volume of data reviewed.

Validation

Similar to previous results, the human and overall processes rating are compared to validate the rating model. Table 22 shows how these data are compared against each

other. The final results look very close for the Twin Cities Marathon, while there is still about a 0.20 difference within the results for the Boston Marathon.

Table 22: Rating process models rating vs. human rating

Event name	Rating using processing models	Rating using human
Boston Marathon	3.7598	3.5549
Twin Cities Marathon	3.5510	3.5870

Discussion

Having an overall process that ties all different SM models to one unified model gives a lot of value. Looking at SM posts through different parts and understanding them as a whole makes the process much better than only focusing on only one dimension of a SM post.

The lesson learned from previous models' results as well as this model is that the volume of data plays an important role in this research. As the ratios of data processed through models, the word-based rating model processed more tweet posts than that of event organizers' tweets.

Finally, having a unified processing model provides a better understanding of a SM post than just having only one or two models rate an event. Results for the two sample events are within 10 % of the margin of error.

CONCLUSION

Currently, this research built 9 SM models, which consist of 6 working processes and 3 future ideas. Thus far, this research was focused on breaking a SM post into small units of words, Hashtags, symbols, emoticons, URLs, images, etc. to understand the sentiment of a SM post. This research took a slightly different approach than other

research took (Corley et al. 2010). Based on a deeper look into results generated from processes as well as unified models, finding the rating of an event using SM data is getting closer to being a reality. With a little more work on identifying the gaps, there may be a true rating of an event using SM data that can help event organizers enhance their current rating system.

As discussed in previous research, there will always be some difference between human rating and process model rating (Silwal & Callahan 2015). Also, the meaning of human sentiments expressed in SM posts may be totally different when processing these SM posts through the rating model, but breaking a SM post into many different parts truly helps to look at SM posts in small parts to generate a whole rating.

For now, as long as the results are within a margin of error of 10 percent, they should be acceptable. As this research moves forward, those gaps of error to find an accurate sentiment of a SM post need to be closed. A look at the final results (table 19) from the two sample marathons reveals that the results still fall within the current margin of error.

Finally, all the models are created using the idea of frameworks of frameworks in an interdisciplinary manner. From this research, the conclusion can made that an event can be rated using unified rating models including Hashtags, emoticons, images, and URLs.

FUTURE WORKS

As this field of rating is still new, there are a lot more new discoveries and innovations needed to further analyze SM posts. The future rating models need to be considered as new innovations to improve the current rating models.

During this process of implementation and validation, a discovery was made in the area of an effect of data volume vs. the rating outcomes. In the future, volume ratios to generate ratings need to be automatically figured such that higher volumes of data receive the bigger ratios of rating for overall processes. Also, closing the gap on error needs to be further explored.

At this time, most of dictionary tables have limited information; they need to be expanded to include more words, Hashtags, and emoticons. Also, all dictionary tables may need to merge into a unified table with different datatypes for better and quicker searches.

As part of the current rating model processes, 3 future rating models discussed in previous sections were not included. In the future, 3 remaining models need to be discussed further to complete the research.

Finally, there is a lot more research yet to conduct in this field of rating models using SM data. This research is just a start of an on-going research project in this area of rating an event using SM data. Also, other SNS such as Facebook need to be added to enhance the rating model.

REFERENCES

Corley, C.D. et al., 2010. Text and Structural Data Mining of Influenza Mentions in Web and Social Media. *International Journal of Environmental Research and Public Health*, 7(2), pp.596–615. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2872292&tool=pmcentrez&rendertype=abstract.

Hu, X. et al., Unsupervised Sentiment Analysis with Emotional Signals. , pp.607–617.

Images, W. & Analysis, I.S., 2013. Sentribute : Image Sentiment Analysis from a Mid-level Perspective Sentiment : Why from a Mid-level Perspective Sentribute : The Framework Facial Sentiment Detection using Eigenfaces Facial Sentiment Detection using Eigenfaces. , pp.8–11.

Kouloumpis, E., Wilson, T. & Moore, J., 2011. Twitter Sentiment Analysis : The Good the Bad and the OMG ! *Artificial Intelligence*, pp.538–541. Available at: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewPDFInterstitial/2857/3251.

Maynard, D., Bontcheva, K. & Rout, D., 2012. Challenges in Developing Opinion Mining Tools for Social Media. *Proceedings of@ NLP can u tag# …*. Available at: http://gate.ac.uk/sale/lrec2012/ugc-workshop/opinion-mining-extended.pdf [Accessed July 6, 2013].

Mohammad, S.M., 2012. #Emotional Tweets. , pp.246–255.

Silwal, S. & Callahan, D.W., 2014. Building a Social Media Rating Model. *Ieee Southeastcon 2014*, pp.1–3. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6950748.

Silwal, S. & Callahan, D.W., 2015. How is My Event Rated ? Rating an Event Using Social Media Data. , 6(3), pp.7–16.

Silwal, S. & Callahan, D.W., 2013. Using Social Media Data as Research Data. , 1, pp.49–55.

Wang, X., Tokarchuk, L. & Poslad, S., 2014. Identifying Relevant Event Content for Real-time Event Detection. , (Asonam), pp.395–398.

Weng, J. et al., 2010. What Do People Want in Microblogs? Measuring Interestingness of Hashtags in Twitter. *2010 IEEE International Conference on Data Mining*, pp.1121–1126. Available at:

http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5694095 [Accessed May 26, 2013].

Yuan, J. et al., 2013. Sentribute: Image Sentiment Analysis From a Mid-level Pperspective. *Workshop on Issues of Sentiment*. Available at: http://dl.acm.org/citation.cfm?id=2502079.

CONCLUSION

Discussion

Interdisciplinary Research

This research is interdisciplinary in nature. As of the writing of this dissertation research, the following disciplines were identified as a part of this research (figure 1): Social Science, Data Science, Marketing, Mathematics, Linguistics, Health and Wellness, Sports, Computer Science, and Engineering. Directly or indirectly, this research touched on multiple disciplines to import, build, process, and generate a numeric rating of an event in an interdisciplinary manner. As the research expands to different areas to evaluate SM data, this research topic will impact and touch more disciplines.

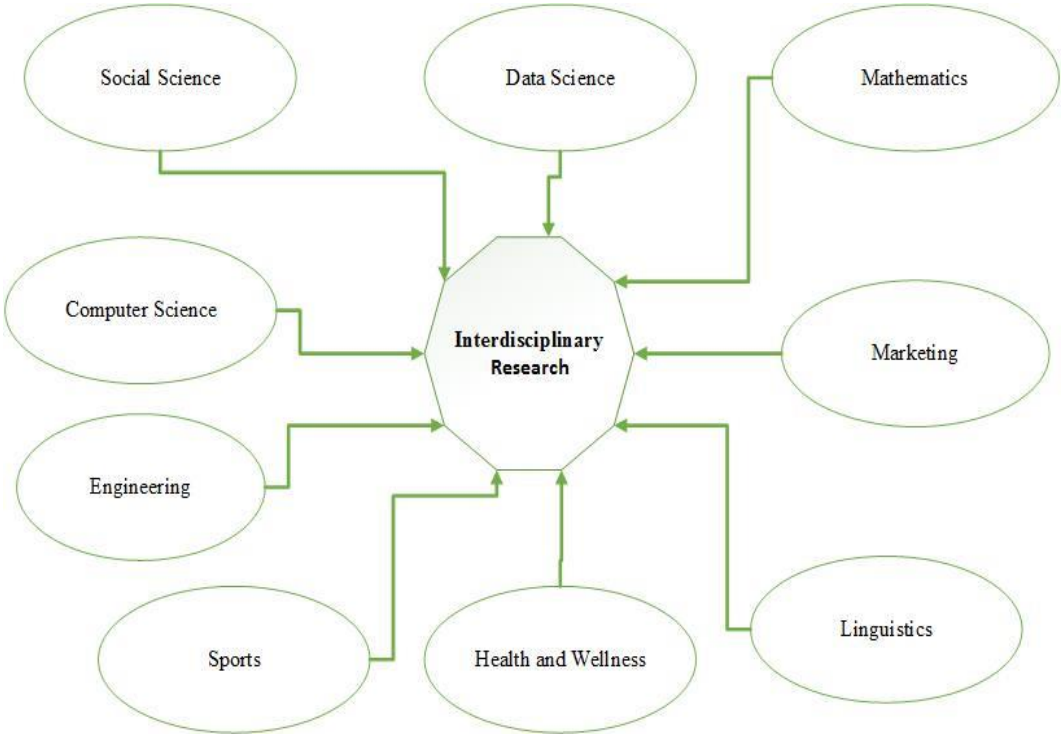


Figure 1: Interdisciplinary Research

Rating framework

This research is based on an interdisciplinary nature to rate an event, which includes many "systems of systems" [8][9]. In this research, models and processes include many disciplines, frameworks, and databases. Without support from each system and its framework, its initial goal of rating an event using SM data may not have been achieved. Figure 2 shows how different systems and frameworks are utilized to build a rating framework. As is the nature of any framework, this rating framework can also be reused for future research topics and ideas.

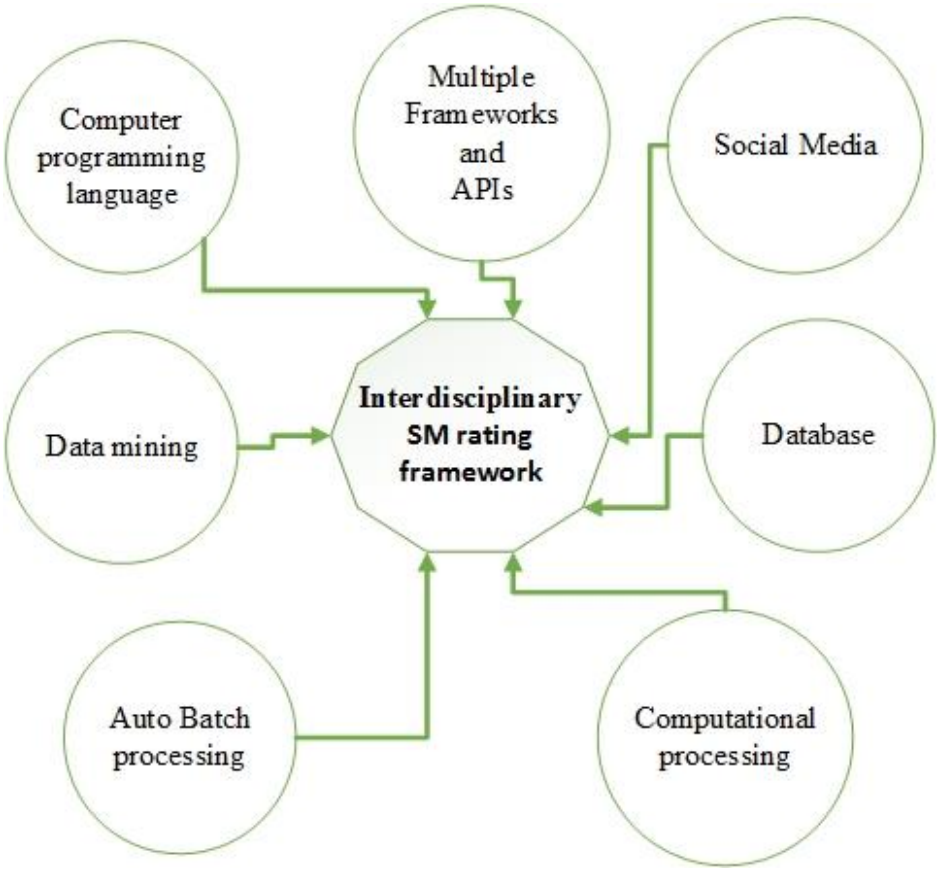


Figure 2: Rating framework of frameworks

As a part of this framework, an event's SM data are collected in real time using an auto batch process. Those collected data are stored in a database table for data mining as well as computational research and processes. For this research, computer programming

language, APIs and multiple frameworks were used to retrieve, store, and process SM data to genereate a numeric rating.

To understand the domain-specific research of running and rating a marathon, multiple dictionary tables such as sentiment, Hashtag, and emoticon dictionaries were built. It helped to look up and reuse common words, terms, word weights, codes, sentiment values, etc. Adding the dictionary as a part of this research provided a great benefit and value as well as gave a structure to the mostly unstructured SM data. In the future, these dictionaries' data can be improved as well as reused.

Data Collection

As long as SNS are the source of data collections, the increase of data for this research will always be possible. As the growth of usability is happening around all SNS, so is the growth of data input. As discussed in previous research, finding quality data will always be a challenge [10]. Over time, however, as these import processes mature, collecting and analyzing quality data will improve as well.

As of now, Twitter's SM data are the only data imported and utilized to build the rating model. In the future, other SNS data sources such as Facebook will be imported to further evaluate a SM post. In such a case, a different importing framework will be created to import new datasets.

Rating Model

At the beginning of this research, only text-based SM models were processed. After the initial run, a realization was made that this research needs to look beyond the simple text-based process to obtain a valid rating of SM data. Additional SM models focusing on Hashtags, emoticons, images, URLs, etc. were added to further analyze SM

posts. That is the main reason a single SM post was processed through 9 different models. This method of a slice-and-dice approach helps to fill the gaps for an otherwise missed opportunity to review, analyze, and rate a SM post.

Even though a SM post can processed through multiple models, eventually a unified model is created to generate a single numeric rating for each SM post. Overall, most of the heavy lifting is done at the individual modeling level, while a unified model is used for combining different models to generate a single rating of an event.

Figure 3 shows a complete flow of the rating model process. In this process, initially SM data such as Twitter data are imported and stored in the warehouse table. Then, the row-by-row SM data from database warehouse tables are processed into 9 different models and will be eventually calculated using a unified model. Figure 3 shows how a SM post is imported, filtered, and processed through multiple SM models as well as the final unifined rating model to generate a final numeric rating. The numeric rating generated from each process is eventually stored into a database table field for future processes and reportings.

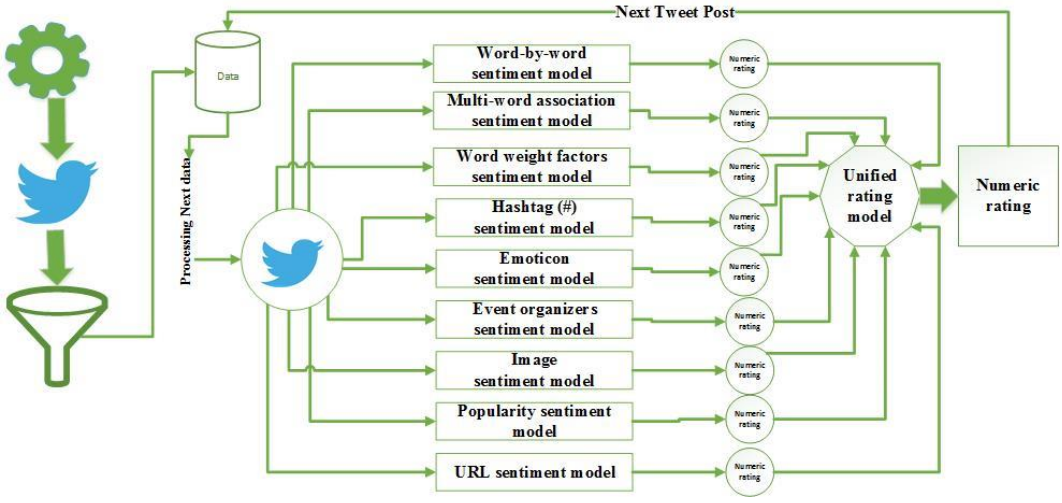


Figure 3: Complete rating model flow

As the growth of SM uses is happening, so are the innovations. It is inevitable that more models will be added into the collection of current rating models in the future to further understand a SM post (figure 3). Even though this rating model is built using only Twitter data, in the future other SM such as Facebook needs to be examined to obtain a compelete view of SM data. These initial model concepts should equally apply to other SM data as well.

Validation

The proper test and validation is an important part of this research. Nine models were discussed in detail, and out of that 6 different working processes were created to further validate current research models. Each of the 6 models generates its own numeric rating. Those ratings contribute its own ways to generate an overall rating of an event.

	Boston Marathon	Twin Cities Marathon	Richmond Marathon
Process rating result	3.7598	3.551	3.7154
Human rating result	3.5549	3.587	3.7761

Figure 4: Process vs. human rating results

As part of the validation process, a comparison was made between process-generated and overall human-assigned average numeric ratings. Even though these validation data were not significant, the overall results were still within 95% accuracy

between rating methods (figure 4). In the future, as computer processes and validations improve, so are these results bound to improve as well.

A deeper look into these results shows that there are still differences between the rating methods' line-by-line results. One of the main reasons for these differences may be due to the fact that the person who posted on SM is not the same person who helped to rate the event's SM post. In the future, there needs to be a better ways to validate SM posts by including different dimensions of validation processes.

Meeting research objectives

A review of all the research reveals that all 3 objectives set during the proposal phase of this research were met. Table 1 shows 3 research objectives vs. 4 research journal and conference paper outcomes. Each of these papers is built in each other to create the final rating model using SM data.

Table 1: Research objective vs. research outcomes

Research objective	Research paper covering objectives
1. Review of the current state of user rating systems, Social Networking Sites, and Text-based Language Processes	Using Social Media Data as Research Data (2013)
2. Develop a Rating Model using Twitter data to generate a numeric rating system	Building a rating model (2014)
3. Compare and contrast Rating Model against manual rating	How is my event rated? Rating an event using Social Media data (2015)
	Can an event be rated using Hashtags, Emoticons, Images, and URLs? Rating an event using Social Media data (2015)

Conclusion

As a part of this research, a lot of lessons were learned in the aspects of importing, processing, and building multiple models to understand and rate an event using unstructured SM data. As this effort to rate a marathon event using SM data moves forward, there is still a lot more research and discovery needed in this area of rating an event using SM data.

Even after years of research, this research has only touched a very small part of the ever-growing universe of SM. As every day billions of SM users connect on their SNS of choice to interact, get current news, event inform, post photos, create movements, etc., this field of SM is rapidly growing and changing as well.

This dissertation research has successfully demonstrated that unstructured SM data can be imported and processed through multiple models to generate meaningful information in an interdisciplinary manner. Even though these prototype results' for the row-by-row tweet were not matched 100% between human rating and process rating, in time, the overall look of these models and processes can improve as new algorithms to calculate and analyze SM data are implemented.

Even though these rating models were able to create numeric ratings for multiple events using SM data, this dissertation research recommends that these concepts discussed in this research should complement other traditional rating methods.

Finally, a statement can be made that with proper methods of importing, modeling, processing, testing, and validating, an event can be rated using SM data.

Future Works

As discussed in previous sections and chapters, this research touched only on SNS as well as a small part of this big universe of SM research. There is more work

remaining for research and product development to fully understand rating of an event using SM data. As the world of SM is changing at a rapid pace, so are the new APIs for these SNS giving more power to developers and researchers to process and understand SM data. In the future, those new features can be used to rate an event using SM data further and better.

The following goals have been set to further enhance and improve the current dissertation research models:

- Look into building prototypes and validate the 3 unfinished rating models.
- Add and improve current dictionary tables' data by importing exciting sentiment assigned words, word weight, and word direction.
- Find better algorithms to distribute overall rating results among different models for the final unified model.
- Improve ways to validate process results data.
- Build a website to disseminate current results to runners, event organizers, and event sponsors.
- Create algorithms to assign different words weight.

Finally, with few modifications and additions, these models and frameworks created as a part of this dissertation research can be taken beyond the current scope of this research. To futher this research, a web application tool will be built so that authorized users can import, process, rate, and review their domain-specific SM data for their own research using these rating models' ideas and tools that were discussed so far during this dissertation research. This will further help to validate and improve this rating model.

GENERAL LIST OF REFERENCES

[1] 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.

[2] 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.

[3] S. Edosomwan and S. Prakasan, "The history of Social Media and its Impact on Business," *J. Appl. ...*, 2011.

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

[5] W. van Osch and C. K. Coursaris, "Organizational Social Media: A Comprehensive Framework and Research Agenda," *2013 46th Hawaii Int. Conf. Syst. Sci.*, pp. 700–707, Jan. 2013.

[6] A. M. Kaplan and M. Haenlein, "Users of the World, Unite! The Challenges and Opportunities of Social Media," *Bus. Horiz.*, vol. 53, no. 1, pp. 59–68, Jan. 2010.

[7] Z. Zhang, X. Li, and Y. Chen, "Deciphering Word-of-Mouth in Social Media : Text-Based," *ACM Trans. Manag. Inf. Syst.*, vol. 3, no. 1, pp. 1–23, 2012.

[8] N. B. Ellison and D. M. Boyd, "Social Network Sites: Definition, History, and Scholarship."

[9] a. Marwick and D. Boyd, "To See and Be Seen: Celebrity Practice on Twitter," *Converg. Int. J. Res. into New Media Technol.*, vol. 17, no. 2, pp. 139–158, 2011.

[10] G. Lotan, E. Graeff, M. Ananny, D. Gaffney, I. Pearce, and D. Boyd, "The Arab Spring| The Revolutions Were Tweeted: Information Flows during the 2011 Tunisian and Egyptian Revolutions," *Int. J. Commun.*, vol. 5, p. 31, 2011.

[11] K. Toutanova, D. Klein, and C. D. Manning, "Feature-rich part-of-speech tagging with a Cyclic Dependency Network," *Proc. 2003 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Vol. 1 (NAACL '03),* pp. 252–259, 2003.

[12] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, "Part-of-speech Tagging for Twitter: Annotation, Features, and Experiments," *Hum. Lang. Technol.*, vol. 2, no. 2, pp. 42–47, 2011.

[13] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith, "Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters," in *Proceedings of NAACL-HLT 2013*, 2013.

[14] Y. Heights, P. F. Brown, P. V. DeSouza, R. L. Mercer, V. J. Della Pietra, and J. C. Lai, "Class-Based n-gram Models of Natural Language," *Comput. Linguist.*, no. 1950, 1992.

[15] K. M. Larson and R. T. Watson, "The Impact Of Natural Language Processing-Based Textual Analysis Of Social Media Interactions On Decision Making," 2013.

[16] C. Wojtalewicz, "Social Media Use for Large Event Management," no. 1, pp. 24–29, 2012.

[17] F. Cheong and C. Cheong, "Social Media Data Mining: A Social Network Analysis Of Tweets During The 2010-2011 Australian Floods.," *PACIS*, 2011.

[18] H. Achrekar and A. Gandhe, "Predicting Flu Trends Using Twitter Data," ... *WKSHPS), 2011 IEEE ...*, pp. 702–707, 2011.

[19] D. Contractor and T. Faruquie, "Understanding Election Candidate Approval Ratings Using Social Media Data," *... 22nd Int. Conf. ...*, pp. 189–190, 2013.

[20] H.-H. Won, W. Myung, G.-Y. Song, W.-H. Lee, J.-W. Kim, B. J. Carroll, and D. K. Kim, "Predicting National Suicide Numbers with Social Media Data.," *PLoS One*, vol. 8, no. 4, p. e61809, Jan. 2013.

[21] J. Weng, E.-P. Lim, Q. He, and C. W.-K. Leung, "What Do People Want in Microblogs? Measuring Interestingness of Hashtags in Twitter," *2010 IEEE Int. Conf. Data Min.*, pp. 1121–1126, Dec. 2010.

[22] M. Russell, *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. 2011.

[23] X. Wang, L. Tokarchuk, and S. Poslad, "Identifying Relevant Event Content for Real-time Event Detection," no. Asonam, pp. 395–398, 2014.

[24] S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. a. Reis, and J. Reynar, "Building a Sntiment Summarizer for Local Service Reviews," *WWW Work. NLP Inf. Explos. Era*, 2008.

[25] B. Liu, "Sentiment Analysis and Subjectivity," *Handb. Nat. Lang. Process.*, pp. 1–38, 2010.

[26] “Opinion Mining, Sentiment Analysis, Opinion Extraction.” [Online]. Available: http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html. [Accessed: 13-Jan-2015].

[27] X. Hu, J. Tang, H. Gao, and H. Liu, “Unsupervised Sentiment Analysis with Emotional Signals,” pp. 607–617.

[28] Y. Q. Xu, Y. H. Zhu, W. H. Wang, and L. C. Gao, “A Dynamic Adjustment Algorithm Research of Sentiment Word Weight Based on Context,” *ICCRD2011 - 2011 3rd Int. Conf. Comput. Res. Dev.*, vol. 3, pp. 19–22, 2011.

[29] E. Kouloumpis, T. Wilson, and J. Moore, “Twitter Sentiment Analysis : The Good the Bad and the OMG !,” *Artif. Intell.*, pp. 538–541, 2011.

[30] S. Silwal and D. W. Callahan, “How is My Event Rated ? Rating an Event Using Social Media Data,” vol. 6, no. 3, pp. 7–16, 2015.

[31] A. Odusd and T. Sse, *Systems Engineering Guide for Systems of Systems*, no. August. 2008.

[32] S. Silwal and D. W. Callahan, “Using Social Media Data as Research Data,” vol. 1, pp. 49–55, 2013.

APPENDIX A

Spring Application XML Configuration File

The Spring application XML configuration file is the main file needed to configure spring with other components as well as database access to processing application.

```
<?xml version="1.0" encoding="UTF-8"?>
<beans xmlns="http://www.springframework.org/schema/beans"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xmlns:aop="http://www.springframework.org/schema/aop"
          xmlns:tx="http://www.springframework.org/schema/tx"
           xmlns:jdbc="http://www.springframework.org/schema/jdbc"
            xmlns:context="http://www.springframework.org/schema/context"
             xsi:schemaLocation="http://www.springframework.org/schema/context
http://www.springframework.org/schema/context/spring-context-3.0.xsd
http://www.springframework.org/schema/beans
http://www.springframework.org/schema/beans/spring-beans-3.0.xsd
http://www.springframework.org/schema/jdbc
http://www.springframework.org/schema/jdbc/spring-jdbc-3.0.xsd
http://www.springframework.org/schema/tx
http://www.springframework.org/schema/tx/spring-tx-3.0.xsd
http://www.springframework.org/schema/aop
http://www.springframework.org/schema/aop/spring-aop-3.0.xsd">
<bean
class="org.springframework.beans.factory.config.PropertyPlaceholderConfigurer">
        <property name="location">
                <value>database.properties</value>
        </property>
</bean>

<bean id="dataSource"
      class="org.springframework.jdbc.datasource.DriverManagerDataSource">
        <property name="driverClassName" value="com.mysql.jdbc.Driver" />
        <property name="url" value="jdbc:mysql://localhost:3306/smdata" />
        <property name="username" value="userid " />
        <property name="password" value="password" />
</bean>


<tx:annotation-driven transaction-manager="txManager"/>
          <!-- a PlatformTransactionManager is still required -->
          <bean id="txManager"
class="org.springframework.jdbc.datasource.DataSourceTransactionManager">
            <!-- (this dependency is defined somewhere else) -->
            <property name="dataSource" ref="dataSource"/>
          </bean>
```

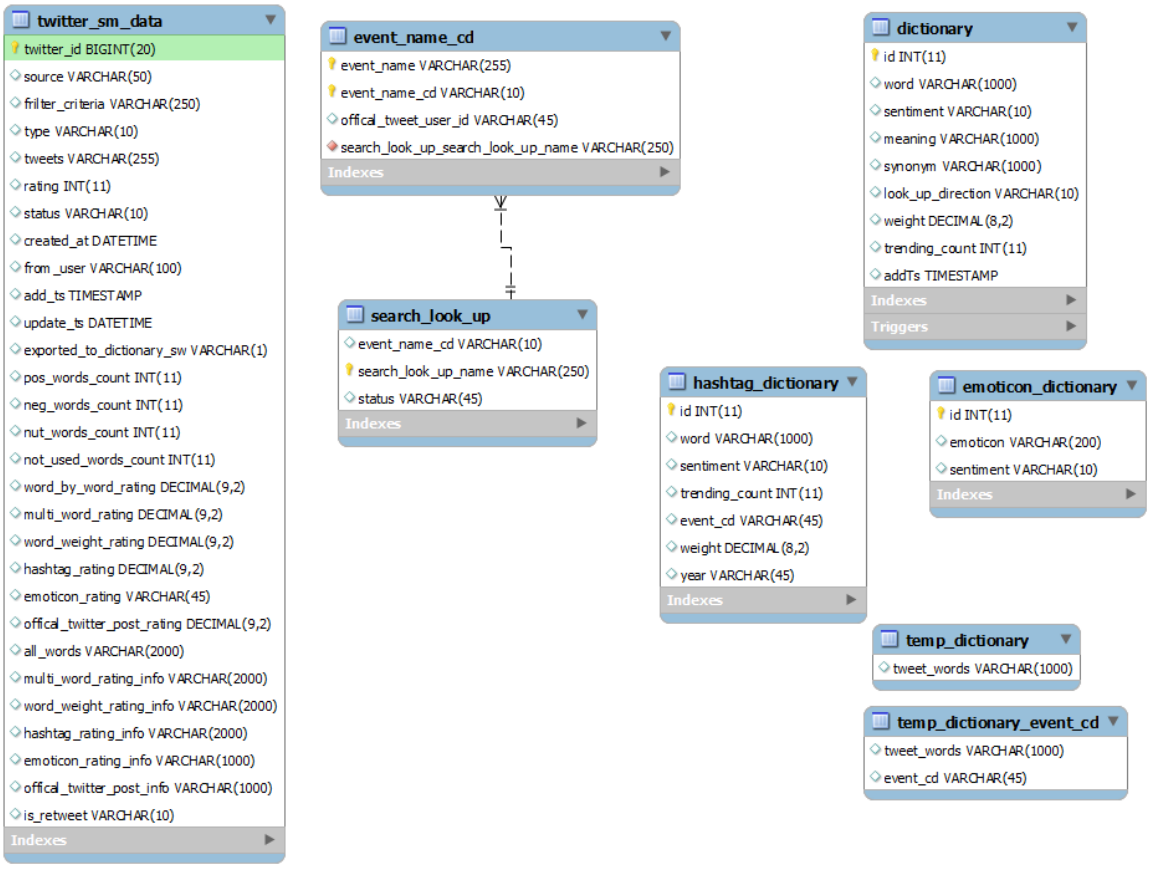
```
<bean id="sqlSessionFactory" class="org.mybatis.spring.SqlSessionFactoryBean">
  <property name="dataSource" ref="dataSource" />
</bean>
<bean class="org.mybatis.spring.mapper.MapperScannerConfigurer">
  <property name="basePackage" value="com.silwal.mapper" />
  <property name="sqlSessionFactory" ref="sqlSessionFactory" />
</bean>

 <context:annotation-config/>
 <context:component-scan base-package="com.silwal.service"/>


</beans>
```

Appendix B

Research Database Schema


twitter_sm_data
twitter_id BIGINT(20)
source VARCHAR(50)
frilter_criteria VARCHAR(250)
type VARCHAR(10)
tweets VARCHAR(255)
rating INT(11)
status VARCHAR(10)
created_at DATETIME
from _user VARCHAR(100)
add_ts TIMESTAMP
update_ts DATETIME
exported_to_dictionary_sw VARCHAR(1)
pos_words_count INT(11)
neg_words_count INT(11)
nut_words_count INT(11)
not_used_words_count INT(11)
word_by_word_rating DECIMAL(9,2)
multi_word_rating DECIMAL(9,2)
word_weight_rating DECIMAL(9,2)
hashtag_rating DECIMAL(9,2)
emoticon_rating VARCHAR(45)
offical_twitter_post_rating DECIMAL(9,2)
all_words VARCHAR(2000)
multi_word_rating_info VARCHAR(2000)
word_weight_rating_info VARCHAR(2000)
hashtag_rating_info VARCHAR(2000)
emoticon_rating_info VARCHAR(1000)
offical_twitter_post_info VARCHAR(1000)
is_retweet VARCHAR(10)
Indexes
event_name_cd
event_name VARCHAR(255)
event_name_cd VARCHAR(10)
offical_tweet_user_id VARCHAR(45)
search_look_up_search_look_up_name VARCHAR(250)
Indexes
search_look_up
event_name_cd VARCHAR(10)
search_look_up_name VARCHAR(250)
status VARCHAR(45)
Indexes
dictionary
id INT(11)
word VARCHAR(1000)
sentiment VARCHAR(10)
meaning VARCHAR(1000)
synonym VARCHAR(1000)
look_up_direction VARCHAR(10)
weight DECIMAL(8,2)
trending_count INT(11)
addTs TIMESTAMP
Indexes
Triggers
hashtag_dictionary
id INT(11)
word VARCHAR(1000)
sentiment VARCHAR(10)
trending_count INT(11)
event_cd VARCHAR(45)
weight DECIMAL(8,2)
year VARCHAR(45)
Indexes
emoticon_dictionary
id INT(11)
emoticon VARCHAR(200)
sentiment VARCHAR(10)
Indexes
temp_dictionary
tweet_words VARCHAR(1000)
temp_dictionary_event_cd
tweet_words VARCHAR(1000)
event_cd VARCHAR(45)

Appendix C

Twitter's JSON Metadata File

```
[
 {
  "coordinates": null,
  "truncated": false,
  "created_at": "Thu Oct 14 22:20:15 +0000 2010",
  "favorited": false,
  "entities": {
    "urls": [
    ],
    "hashtags": [
    ],
    "user_mentions": [
      {
        "name": "Matt Harris",
        "id": 777925,
        "id_str": "777925",
        "indices": [
          0,
          14
        ],
        "screen_name": "themattharris"
      }
    ]
  },
  "text": "@themattharris hey how are things?",
  "annotations": null,
  "contributors": [
    {
      "id": 819797,
      "id_str": "819797",
      "screen_name": "episod"
```

```
32.       }
33.     ],
34.     "id": 12738165059,
35.     "id_str": "12738165059",
36.     "retweet_count": 0,
37.     "geo": null,
38.     "retweeted": false,
39.     "in_reply_to_user_id": 777925,
40.     "in_reply_to_user_id_str": "777925",
41.     "in_reply_to_screen_name": "themattharris",
42.     "user": {
43.       "id": 6253282,
44.       "id_str": "6253282"
45.     },
46.     "source": "web",
47.     "place": null,
48.     "in_reply_to_status_id": 12738040524,
49.     "in_reply_to_status_id_str": "12738040524"
50.   }
51. ]
```

Source: https://dev.twitter.com/docs/twitter-ids-json-and-snowflake