Gergana Dimitrova. Correlating Twitter Sentiment with Basketball Game Events and Characteristics. A Master's Paper for the M.S. in I.S degree. April, 2015. 75 pages. Advisor: Stephanie Haas.

This study performed sentiment analysis on Tweets created during 30 basketball games. 14,440 Tweets were collected using the Twitter Streaming API, parsed by a script and analyzed using the program SentiStrength. There was significant strong positive correlation between game sentiment of fan Tweets and the outcome of the game. Significant correlation was also found between fourth quarter sentiment of fan Tweets and the outcome of the game. Findings suggest that fans participate in live-Tweeting game events as a way to interact with the game and other fans. This study proposes that improvements can be made to sentiment analysis through domain-specific approaches such as improved domain dictionaries and lexicons.

Headings:

Sentiment Analysis.

Microblogging -- Twitter.

Sports -- Basketball.

CORRELATING TWITTER SENTIMENT WITH BASKETBALL GAME EVENTS AND CHARACTERISTICS

by
Gergana Dimitrova

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

April 2015

Approved by

Stephanie Haas

Table of Contents

Table of Figures

Table of Tables

1. Introduction

Information retrieval (IR), a broad field in information science, is the study and practice of retrieving relevant collection objects based on a query. A query is a formal expression of an information need. A system takes the query and matches it against a corpus to return relevant documents using some chosen model or models. Many modern systems of IR use natural language for querying as opposed to an artificial language like Interslavic, which was constructed to allow international communication between Slavic nations[1]. Thus, text processing is a big component of IR.

Sentiment analysis is a field in natural language processing that intersects with computational linguistics, which is used to expand aspects of information retrieval. Research in this field has grown rapidly in the last fifteen years and continues to do so. Sentiment is subjectivity or emotion that is expressed. Sentiment analysis, opinion mining or analysis, and subjectivity analysis all refer to the same process. The purpose of sentiment analysis is to computationally extract sentiment from a corpus, document or text. Unlike information retrieval, opinion mining aims to analyze not just what a document contains but also the sentiment expressed in it.

Sentiment analysis is often used in product reviews. Hu and Liu (2004) use multiple sentiment analysis techniques to gather information from reviews and summarize the sentiment expressed in a large number of the product reviews. They extract sentiment for features of a product and summarize how many positive and negative reviews a feature received. For example, a digital camera can have several features such as size and picture quality so for each feature, Hu and Liu provide a count of positive reviews and negative reviews that a feature such as size has received (p. 168). This type of summary could help non-expert shoppers wading through a large number of reviews, as well as product manufacturers to see which products are faring well and where products can be improved (p. 176).

Tumasjan et al. (2010) apply sentiment analysis to opinions on social media regarding politics. Specifically, they looked into 100,000 Tweets that mentioned a political party during the German federal elections. They found that sentiment expressed on Twitter closely reflected the political positions of the parties mentioned in the sentiment. The authors conclude that Twitter can be used as a valid real-time indicator of political sentiment that might reflect the offline political landscape (p. 184). Hu and Liu, and Tumasjan et al. show that use of sentiment analysis can be very helpful over a large number of domains for real-world applications.

Use of sentiment analysis can be applied to many areas including politics, advertising and business. Because this field is relatively new, there are many challenges to be researched and addressed. For example, in order to mine opinions, researchers must specify what classifies as an opinion and how to label it. Additional challenges also include mining for context. While some content has certain words that indicate the opinion of the user (“great,” “terrible,” “delicious,” etc), other content must be put in context to understand the opinion. For example, “go read the book” can mean one thing for a book review and another for a movie review. There is an increasing number of opinion mining approaches for different types of content. The number of application domains, such as social media and health-related information, in which opinion mining is being used is growing rapidly.

For approaches that focus on social media, Twitter is a popular resource. An enormous number of people use Twitter to share messages about any imaginable topic. There are currently 288 million active users[2] sharing opinions on anything from politics to appliances to sports. Given this diversity and the volume of data, researchers collect anywhere from 8,000 Tweets (Agarwal, et al., 2011) to 34,000,000 Tweets (Thelwall, Buckley & Paltoglou, 2011) for analysis.

An additional benefit of Twitter is that data can be collected as people write it. Live-Tweeting is when people discuss an event (e.g. TV show, conference, hostage situations, etc.) as the event is unfolding. Using this data, organizations involved with the event, product producers and advertisers, and researchers can see when opinions are strongest, when they change and possibly identify sub-events that affect sentiment (e.g. a particular speaker at a conference, or scene from a show). The information can be used to improve services, events or even an organization's image by analyzing where sentiment is present. For example, sentiment analysis on movies or TV shows can show what the audience finds exciting or awful, which can help with future project development.

The purpose of this study is to determine if there is a correlation between the sentiments expressed in Tweets about the event and the event's characteristics. I investigated if there is a correlation between sentiment expressed during a basketball game in game-related Tweets and the outcome of that game. If a team is winning, fan participation might indicate this through positive sentiment for the team throughout the game. Fans might express their positive opinions of an exciting play or for a particular player who is having an impact on the game overall. On the other hand, sentiment turning negative during the last quarter of the game can indicate a loss for the team. If the team is missing a lot of baskets or turning the ball over a lot, fans might indicate their frustration with negative sentiment. Analyzing sentiment for basketball games on Twitter can be useful in a few ways. The analysis can help with market research to gauge fan participation, which is very important for a team to stay relevant to fans; for example, it can help with targeted advertisement depending on how a team is perceived to be performing based on sentiment. Information about the sentiment associated with games might suggest relationships between sentiment and distribution of games (e.g. being nationally televised), opponent popularity or regular and post-season games.

Additionally, this topic can be of interest to psychology and sociology. The reactions and behavior of sports fans during games has been studied in the context of these fields. Using manual sentiment analysis, Iliycheva (2005) studied Bulgarian sports fans writing on online forums for football (soccer), weightlifting and basketball in order to show how nationalism and sports sentiment are linked. She found that “What becomes an emotional centre for the fans is the image of ‘us.’ Most of the postings, especially the ones that refer exclusively to the sports events and people, are written in first person plural. The fans completely identify with the athletes and accept themselves as equal and active part in the entire process of winning or losing…” (p. 261). By looking into the behavior of fans, broader trends can be seen. Sports are an important part of American life and rivalries between states or even cities can begin with sports. Therefore my research question is:

1. Is there a correlation between the average sentiment strength of fan Tweets and the outcome of a game?

More specific questions focus on where sentiment might be influenced during a game.

1a. Is sentiment stronger or weaker during close games?
1b. Is sentiment stronger or weaker during blowout games?
1c. Do sentiment strength and polarity stay the same over all quarters?
1d. Do events affect sentiment strength?

---

[1] http://www.interslavic.org/
[2] https://about.twitter.com/company.

2. Literature Review

Two leading researchers in the field of sentiment analysis are Bo Pang and Lillian Lee. Pang currently works at Google, Inc. where his research areas are Natural Language Processing (NLP) and social media (*bo.pang*). Lee is currently at Cornell where she researches NLP and social interactions (*Lillian Lee: Research Summary*). Their monograph detailing sentiment analysis, “Opinion Mining and Sentiment Analysis,” (2008) is a significant article cited throughout the field and covers background on opinion mining, examples of applications, and an overview of challenges and approaches. The challenges to sentiment analysis that they discuss are: the differences between fact-finding IR and opinion mining, subtlety of sentiment, and domain context. The approaches to sentiment analysis include domain adaptation, unsupervised approaches, and relationship classification. Finally, broader implications to consider are privacy and manipulation.

In their discussion of the terms, Pang and Lee (2008) conclude that "sentiment" or "opinion" are defined with subtle differences by many researchers, however they are most commonly defined as subjective views that cannot be verified or objectively observed (p. 9). They define "polarity" of sentiment as subjective text that has either positive or negative opinion expressed (p. 10). They use the term “strength” to indicate how powerfully the opinion is expressed (p. 29). For instance, an opinion stated as “the movie was great!” has stronger sentiment than the sentence “the movie was good,” because the word “great” is more positive than “good” and the exclamation mark indicates more excitement. However, it is important to note that both contain sentiment (p. 29) because they are both expressing subjective positive views. Sentiment is not about the strongest opinion, so both must be considered when performing automatic analysis.

Pang and Lee note that extracting sentiment from a document is very different from fact-based textual analysis. Fact-based (objective) text categorization aims to classify documents by topics. Topic categorization can use term frequency, tf*idf and a number of other well-researched approaches to determine the topic of a document. However, sentiment classification requires different approaches. As Pang and Lee state, “...with sentiment classification, we often have relatively few classes (e.g., “positive” or “3 stars”) that generalize across many domains and users...In fact, the regression-like nature of strength of feeling, degree of positivity, and so on seems rather unique to sentiment categorization” (2008, p. 10). This means that given a binary classification (positive/negative), the labels are opposing (which the authors note is similar to binary topic-based relevance). However, for ordinal categories, a variety of sentiment has to be placed in a (typically) small scale that does not vary over topics (for example, using a 5-star rating system for anything from cars to washing machines to hotels).

Something that Pang and Lee (2008) discuss is the use of labeled data in testing methods. They note that the rise of labeled data gave the field of sentiment analysis a large-scale empirical evaluation tool by essentially having the “right” analysis against which to test new systems (p. 24). Esuli and Sebastiani (2004) use labeled data as a “gold standard” against which to test their approach (p. 421). Much like test collections that include relevance judgments for queries, labeled data has human-determined sentiment that allows researchers to build better systems by knowing what the ultimate outcome of their programs should be.

Sentiment analysis relies on keywords to determine sentiment, which can be a challenge. For instance, a list of words that implies negative or positive sentiment can vary between domains. What has a positive connotation for sports might have a negative connotation for politics. For example, if a shot or play was “crazy” or “nasty”, the sentiment behind that is positive for sports fans (e.g. “That shot was crazy!”), whereas if a politician or a bill is described that way, the connotation changes to negative (e.g. “It’s crazy that the Idaho Republican doesn’t know female anatomy!”). Additionally, gathering and maintaining the list of words (in a lexicon) that are labeled as positive or negative is an extreme amount of work if done manually. One way to alleviate that is to employ the use of machine learning to achieve a higher accuracy in analysis by using training data (Pang & Lee, 2008, p. 11). Not only does curating the lexicon take a long time, figuring out which terms to include can be problematic. Research is required in order to determine which words have strong enough sentiment consistently across domains to be included in a lexicon and care must be given to maintaining the list so that it excludes superfluous terms. As Pang and Lee (2008) state, “Compared to topic, sentiment can often be expressed in a more subtle manner, making it difficult to be identified by any of a sentence or document’s terms when considered in isolation” (p. 12). In one example they give, “She runs the gamut of emotions from A to B,” they note that “no ostensibly negative words occur” (p. 12) but the sentiment still has a negative connotation because it is essentially saying that the actress can *only* portray two emotions. Only using a lexicon to match keywords for a positive or negative sentiment would not produce very good results in cases like these.

This leads to another issue of sentiment analysis – domain context. Rather than textual context, which refers to understanding phrases based on the text, domain context describes the idea that the domain in which the sentiment is expressed is key to understanding. The example provided in this paper is “go read the book” as it pertains to a book review and movie review (Pang & Lee, 2008, p. 13). In one domain, it is a positive review for a book that someone is recommending. In the other, it is a negative review for a movie that did not live up to expectations. Pang and Lee (2008) go on to note that “In general, sentiment and subjectivity are quite context-sensitive, and, at a coarser granularity, quite domain dependent (in spite of the fact that the general notion of positive and negative opinions is fairly consistent across different domains)” (p. 13). While lexicons can allow for different domains’ word usage, they cannot be used as the sole source of information for sentiment if one expression is used with differing sentiment in different domains.

In order to account for differences in sentiment using the same vocabulary over multiple domains, many researchers have attempted to find effective solutions. One of the approaches to account for context is domain adaptation. Read (2005) found that standard machine learning yielded good results by relying on emoticons, while Blitzer et al. (2007) found that a Structural Correspondence Learning (SCL) algorithm using pivot features proved successful. The latter involves a source domain, which has the training data for the algorithm, and a target domain, which is the new domain to be analyzed; the data sets were then joined on the pivot features. Pivot features are frequently occurring terms present in two domains that are chosen based on their mutual information scores (Blitzer et al., 2007, p. 441). This allows for domain transfer to the target domain by using the pivot features to predict sentiment in the target domain. Additionally, using topic-extracting analysis along with sentiment analysis can also be beneficial. Pang and Lee (2008) note, "One approach to integrating sentiment and topic when one is looking for opinionated documents on a particular user-specified topic is to simply first perform one analysis pass, say for topic, and then analyze the results with respect to sentiment" (p. 43). This would help identify discrete topics within a collection that would then be mined for sentiment, as opposed to mining the entire collection for one domain-specific sentiment or even for general non-domain specific sentiment.

Two unsupervised approaches to sentiment analysis are lexicon based approaches and bootstrapping. Pang and Lee (2008) state,

> Quite a number of unsupervised learning approaches take the tack of first creating a *sentiment lexicon* in an unsupervised manner, and then determining the degree of positivity (or subjectivity) of a text unit via some function based on the positive and negative (or simply subjective) indicators, as determined by the lexicon, within it. (p. 27)

Some variations on this approach include gathering words based on whether they appear with other words (using mutual information and co-occurrence) and using seed words to determine which clusters to label as positive or negative. Bootstrapping can be defined as using the results from an initial classifier to create labeled (training) data, and applying a second algorithm to the results (Pang & Lee, 2008, p. 28). This allows algorithms to self-train themselves to provide sentiment analysis.

One of the most prevalent lexicon-based approaches to sentiment analysis is the LIWC: Linguistic Inquiry and Word Count. This is a software program that features a lexicon developed from an initial extensive study of language.[3] Development of this software began in 1993 and it has since undergone significant changes based on years of research. The word list at the heart of LIWC2007's implementation contains 4,600 words and stems, each of which defines a category or subcategory. One word can belong to more than one category or subcategory. For example, *cried* is in five categories: sadness, negative emotion, overall affect, verb and past tense verb. Using these categories is how LIWC scores the words so if *cried* appears in text, each of the scores for the categories will be incremented.[4]

Another popular lexicon-based tool available is SentiWordNet.[5] The developers of SentiWordNet attempted to create a lexicon that gives users a score for a WordNet synset (Esuli & Sebastiani, 2006, p. 417). A synset is a set of terms that are held together by a common definition. For example, *blasphemous*, *blue* and *profane* are all in the same synset because they meet the definition of "characterized by profanity" (Esuli & Sebastiani, 2006, p. 418). SentiWordNet is designed to solve three issues of sentiment analysis: determining sentiment, determining objectivity (which is the lack of sentiment), and determining polarity strength (which is determining positive/negative polarity and the strength of that polarity). To do so, each term in a synset is assigned three scores: one for positive sentiment, one for negative sentiment and one for objectivity. The scores vary from 0.0 to 1.0 and add up to 1.0 (Esuli & Sebastiani, 2006, p. 417-418). The authors had not gone through a complete evaluation of their system. Esuli and Sebastiani indicated that more work was to be done but that their benchmark tests were sufficiently effective to continue the work. The developers of this system have subsequently introduced SentiWordNet 3.0 in their 2010 paper. Changes include improvement in the algorithm involving ranking and updates to the synsets being used from WordNet 3.0 (Baccianella et al., 2010, p. 53). This lexicon is still currently available for research and at this time is still on version 3.0 (*SentiWordNet*).

Another tool that is currently available is OpinionFinder. In the introductory paper for the tool, Wilson et al. describe its purpose as “aim[ing] to identify *subjective* sentences and to mark various aspects of the subjectivity in these sentences, including the *source* (holder) of the subjectivity and words that are included in phrases expressing positive or negative sentiments” (Wilson, 2005, p. 34). This is not a lexicon, but rather a two-part classifier system that uses a few lexicons. First, the document is processed (including tokenization, parts-of-speech tagging, etc). Then the document is analyzed for subjectivity (Wilson, 2005, p. 34-35). This analysis has four components. First, a Naive Bayes classifier is applied to distinguish subjective sentences from objective ones. Next, a rule-based classifier is used to identify speech events and direct subjective expressions. Speech events are parts of the text that indicate someone has expressed a comment by using terms like “said” or “according to.” Direct subjective expressions are defined as words or phrases where an emotion or opinion is directly described, such as “fears” or “happy” (p. 35). After that, the source of the opinion is identified using a trained system that combines a tagging model and extraction pattern learning (p. 35). Finally, words that contain sentiment are extracted using two classifiers trained on the MPQA Corpus (Wilson, 2005, p. 35). Two versions of OpinionFinder are currently available, one that relies on external packages and version 2.0 that is Java-based and platform-independent (*OpinionFinder System*).

After lexicon and bootstrapping methods, classification using relationships is the last approach described in Pang and Lee. There are many relationships to take into account when classifying sentiment. One relationship to consider is user-to-user communication. Using a study of 100 responses in newsgroups, Pang and Lee found a discourse relationship consisting of opposing polarity of sentiment. Given comments for a newspaper article, each comment in one conversation has the opposing sentiment of the previous comment. For example, if a user replies to an article with a negative comment, the reply to that user would likely have a positive sentiment about the article, and the reply to the reply would likely be a negative comment about the article (Pang & Lee, 2008, p. 48-49). Understanding this trend could help build more effective analytical tools. Relationships between sentences and documents are also considered. These can be monitored to assign objectivity to certain sentences within a document or to monitor sentiments across the document (Pang & Lee, 2008, p. 47). For example, "I really enjoyed the movie. The theater where I saw it was disgusting though" shows opposing sentiment about different topics that can be extrapolated using a sentence-document relationship model. This model looks at the sub-document units (e.g. sentences) and labels them separately. Using these labels, the document can receive a more accurate label (p. 47).

2.1 Social Media

One domain rich with sentiment is social media. Many companies do customer support on Twitter and brands are reaching out to their customers to keep them engaged in their products or services through social media. Government representatives communicate with their constituents on social media. Social news sites like Digg and

Reddit have millions of users commenting on current events. Given the great amount of opinions on social media, sentiment analysis can be incredibly useful in many contexts. Politicians that interact with the public using these websites can gauge how well their efforts are paying off by using sentiment analysis on the comments. Corporations can use the comments to see whether or not the public supports a business move.

While tools and methods are being researched across sentiment analysis, social media sentiment analysis has additional challenges. Maynard et al. (2012) discuss challenges that they encountered when they built applications for sentiment analysis for this domain. Some larger sentiment analysis issues over many domains are entity extraction, which is the identifying persons, location and organizations, and event recognition, which detects a world event or topic such as ‘crisis’ or ‘economic growth’ (Maynard et al., 2012, p. 17). However, for social media in particular the issues Maynard et al. faced were relevance, target identification, negation, context, volatility, and summarization. In this domain, relevance is difficult because “Even when a crawler is restricted to specific topics and correctly identifies relevant pages...discussions and comment threads can rapidly diverge into unrelated topics, as opposed to product reviews which rarely stray from the topic at hand” (Maynard et al., 2012, p. 18). One of the solutions to this is to use clustering to pick out data centered around topics and discard points that stray. This allows users to apply a topic filer first and then do the sentiment analysis on the documents that are on topic. Target identification is another concern. Targets are defined as topics about which a sentiment is expressed. The target can be the direct subject of the sentence (such as “The beautiful flowers are on the table”) or it can be an object of a prepositional phrase (such as “Cheese is available on the beautiful table”). It is important to identify the target because the topic of a sentence is not necessarily the object of the sentiment expressed. In the example above, cheese is the topic of the sentence, but the opinion that is expressed is about the table. The example the authors provide for this issue is expressing sadness at the death of a celebrity. These Tweets would be classified as negative but only because of mourning, not because the celebrity was hated (Maynard et al., 2012, p. 18). The target of the sentiment is the death while the topic or subject is the celebrity. A solution to this issue is to simply mark documents as sentiment containing, rather than include the topic of the sentiment. This eliminates the need to sort them into a topic while still retrieving the results with sentiment.

The next issue recognized by Maynard et al. is negation, specifically as it pertains to unigrams. Unigram-based approaches only look at one word at a time and make an independent judgment about the sentiment of the word, meaning that ‘not good’ is processed as ‘not’ and ‘good’ independently of each other. Using a rule-based system allows for bigrams and n-grams to make unigrams, such as ‘isn’t helpful’ to make the pseudo-word ‘NOT-helpful’ to be processed as a unigram (Maynard et al., 2012, p. 18-19). Adding these types of unigrams to lexicons so that the algorithm has a match when parsing the text can improve performance.

Much like negation detection, another major issue for social media data mining is detecting irony, such as someone saying “great job” about an event that one does not like. In Carvalho et al.’s (2009) paper on irony detection, the authors used sentiment analysis to identify key phrases and punctuation that indicate positive sentences with ironic statements on the assumption that irony reverses the polarity of the sentiment. The authors found that identifying negative opinions is much easier than identifying positive opinions with verbal irony. In their research, Carvalho et al (2009) collected about 250,000 user posts from a popular Portuguese newspaper website (due to their research being done in Portugal). They found eight Portuguese-specific and multilingual clues for detecting positive sentiment used for a negative opinion: diminutive forms, demonstrative determiners, verb morphology, cross-constructions, laughter expressions, quotation marks, heavy punctuation, and interjections (p. 53-54). The clues they could generalize to multiple languages were interjections in specific contexts (e.g. short, positive sentiment), heavy punctuation ("!!!"), quotation marks, and laughter expressions (e.g. "haha") (Carvalho et al., 2009, p. 54-55). These are also the four markers they found to be most effective in helping to determine irony overall and were only tested on social media texts (Carvalho et al., 2009, p. 56).

Gathering context is also difficult in social media. As Maynard et al. (2012) state, "Social media, and in particular Tweets, typically assume a much higher level of contextual and world knowledge by the reader than more formal texts. This information can be very difficult to acquire automatically" (p. 20). However, Tweets in particular (and social media more broadly) have a great amount of metadata included in each post, meaning there are a number of ways to disambiguate sentiment by using this data to detect context. Metadata can also help when it comes to volatility over time for social media. As Maynard et al. (2012) point out, public sentiment can change very quickly over time and one way to address this is to use the timestamp in order to put a sentiment in the right temporal context (p. 20). For example, an opinion expressed in a Tweet supporting Barack Obama in 2008 might not represent sentiment for Barack Obama in

2014. By having the time that the opinion was expressed, readers can put it in context with the 2008 presidential election. Even more volatile is something like a hashtag that comes about quickly and ends just as fast used in sports and other current events. Hashtags can provide temporal and topical context, as well as attitude. For example, #YesAllWomen was popular during May 2014, used in reaction to perceived misogyny of a violent event.[6] If a Tweet contained this hashtag, it would be reasonable to assume it was created in May 2014 and that is about misogyny in American culture.

Agichtein et al. (2008) took on the task of finding high-quality content in social media, another important factor in sentiment analysis. Finding high-quality content is important to sentiment analysis because of the prevalence of spam and non-textual content (i.e. pictures or links). In order to get accurate sentiment analysis for a population, it is necessary to be able to weed out poor content, such as robot-created text that does not represent the views of a real person. In this study, Agichtein et al. (2008) examine the question-and-answer forum Yahoo! Answers to find content that is defined as both questions and answers that are well-written (e.g. proper grammar, punctuation and capitalization) (p. 186). This could help with content ranking, as well as with separating content from spam in order to pinpoint where to analyze for sentiment. Pang and Lee (2008) discuss reviewer quality on sites like Amazon.com that enable users to vote on reviews with a binary helpful/not helpful judgment (p. 50). Agichtein et al. seek to make similar judgments (whether a post is high quality or not) in their project using Yahoo! Answers as their social media platform. In order to get the intrinsic quality of a post (only the content), they measure punctuation and typos, syntax and semantic complexity, and grammar (Agichtein et al., 2008, p. 186). They then take into account user relationships. Agichtein et al. look at users interacting with other users (for example, noting when a user replies to a question posted by another user who has answered the first user's question) and model the relationships between users as a graph of nodes and edges. Link analysis for relationships and usage statistics are more traditional information retrieval methods of determining quality but still good indicators of popularity of answers in this project (Agichtein et al., 2008, p. 186).

The assumption that popular posts have high-quality content and that users who give good answers vote for other good answers drive the usage of these methods. A classifier is used to identify high-quality text (split into question quality and answer quality), which proved accurate for this project (Agichtein et al., 2008, p. 190-192). Furthermore, the authors believe that their work on a question/answer site can be applied to other social media domains, although their system was not tested in other domains (Agichtein et al., 2008, p. 192).

2.2 Approaches to Social Media Sentiment Analysis

Pak and Paroubek's (2010) approach social media opinion mining with a system to mine Tweets and analyze them for sentiment. They first collected Tweets using emoticons to sort them into two types, positive and negative. They additionally gathered Tweets from news agencies' Twitter accounts to make up their objective type on the assumption that they are objective (p. 1321). Due to the character limitation, it is a reasonable assumption that these news agencies post article titles or a brief description of an event using objective language. After collecting the data and tagging parts of speech, Pak and Paroubek (2010) found that "...objective texts tend to contain more common and proper nouns, while authors of subjective texts use more often personal pronouns" (p.

1322). Text processing that includes tokenization and removing stopwords follows (p. 1323). Finally, the classifier is trained on their initial corpora and later they improved accuracy by discarding frequent n-grams that do not indicate sentiment (p. 1324).

It is important to note that Pak and Paroubek (2010) performed this study under the assumption that one Tweet contains one sentiment given the character limitation (p. 1321). This differs from the approaches that Pang and Lee (2008) cover that discuss the relationship between sentence and document. In their overview they state that sub-document units, such as sentences, can often include opposing sentiment (p. 47). However in Pak and Paroubek's article, the authors explain that due to the shortness of a Tweet (limited to 140 characters), the relationship between sentence and document is more of one to one, that is one document (e.g. a Tweet) is about one topic and has one sentiment about the topic (p. 1321).

In another approach to social media opinion mining, Paltoglou and Thelwall (2012) use an unsupervised algorithm to focus on data from Twitter, Digg and MySpace. They discuss the difficulty of mining social media for sentiment as opposed to more structured data, like product reviews. While product reviews usually have pros and cons that are listed out by a user and focus on one topic (the product and distinct aspects of the product), social media interactions can be much shorter and be about any topic. Some of the reasons for the lack of a "gold standard" (as they put it) for informal sentiment analysis are that unlike product reviews, social media text tends to be shorter, do not include metadata that mirrors their sentiment (e.g. stars, ordinal rating), and generally have informal spelling and slang (Paltoglou & Thelwall, 2012, p. 66:2). In their paper, Paltoglou and Thelwall use the following features to determine sentiment:

negation/capitalization detection, intensifier/diminisher detection, emoticon/exclamation detection, and a traditional lexicon. Unlike Pak and Paroubek (2010), this approach is unsupervised because there is no need for a reference corpus. They combine multiple methods in order to come up with a robust measure of sentiment polarity and intensity in short, informal text (Paltoglou & Thelwall, 2012, p. 66:5).

Paltoglou and Thelwall use the LIWC (Linguistic Inquiry and Word Count) emotional dictionary due to its basis in psychology and for the ease with which it can be used in informal language (p. 66:6). This tool creates a neighborhood of words by considering the five words before and after the emotion word. Neighborhoods can also be sentences by themselves, denoted by a period, comma or question mark. This step helps uncover "long-distance phenomena" as in the example "I don't think this is a good movie…" (Paltoglou & Thelwall, 2012, p. 66:6). The authors created a corpus of their own labeled data in three parts. They used Twitter data that had emoticons which they assumed indicated the emotion expressed in the text (that is, ":-)" indicated a positive sentiment, ":-(" indicates negative sentiment). They also used data from Digg and MySpace that were given to human annotators for sentiment and polarity judgments (Paltoglou & Thelwall, 2012, p. 66:7-66:10).

Paltoglou and Thelwall's approach was tested against three state-of-the-art machine learning methods that all use unigrams (p. 66:11). In their analysis, they found that the polarity classification with their system performed better than the unigram approaches and proved robust enough to be recommended for other social media data. Paltoglou and Thelwall's system had an F1 score of 86.5% while the next best system's score was 80.7% (p. 66:13-66:14). For detecting subjectivity however, the proposed system performed worse than two of the three competing systems, but still proves relatively successful with an F1 score of 70.9% (p. 66:15). Overall, Paltoglou and Thelwall (2012) found their proposed classifier to be robust and perform well across a variety of social media data with no training or adjustment made to the system (p. 66:16).

Thelwall et al. (2010) introduce the program SentiStrength, a program geared towards sentiment detection in short and informal text. The goal with this software was to create a machine-learning approach to sentiment analysis that optimizes sentiment term weighing, to contribute methods for extracting sentiment from non-standard spellings in text, and to create a spelling correction method of sentiment analysis (Thelwall, et al., 2010, p. 2555). The core of the program is the list of words with sentiment scores. This list of words contains 298 positive terms and 465 negative terms (p. 2549). This list was developed using 2,600 MySpace comments and human classification of the terms. The developers used three judges to determine scores for these terms, ranging from positive 2 to 5 or -2 to -5 for negative sentiment (p. 2548). Many of the terms on this list were based on truncated words from LIWC. After the term list had been added, developers optimized their approach through repetitive sentiment analysis to stabilize word scores (p. 2549). A rule was added to determine alternative spellings of words due to repeated letters (e.g. helloooooo is identified as hello). The score of a word with two or more repeated letters is boosted by 1 because Thelwall et al. found that in their initial set of data, letters repeated twice or more shows increased emotion or energy (p. 2549). An additional list of booster words is also included in the algorithm. Booster words are terms that increase or decrease the sentiment of subsequent words. For example, “very good” would have a stronger sentiment score than “good” because of the booster word “very” (p. 2549).

These words increase the sentiment score by 1 or 2 or decrease it by 1 depending on the term. Additionally, a negating word list is included. Words on this list are considered to invert subsequent emotion words. For example, “not happy” would have an inverted score from “happy” so that if the former scored a positive 2, the latter would score a -2. An emoticon list supplements the algorithm and consists of emoticons and their associated strengths, which were determined to be either positive or negative 2. Finally, punctuation was taken into account, such as periods, exclamation marks and question marks. Thelwall et al. elected not to count negative words in questions because their pilot data did not show these questions to contain sentiment. However, positive sentiment in questions is counted based on idioms such as “what’s up?” which they have determined contains mild positive sentiment (p. 2549). Repeated punctuation received a strength boost of 1 to the emotion-bearing word immediately preceding the punctuation.

SentiStrength was developed because the authors saw a need for sentiment analysis tools in informal texts. Thelwall et al. (2010) argue that the LIWC is better for longer documents, where statistics of the program would serve the researcher better (i.e. how well people cope with bereavement) (p. 2546-2547). SentiStrength's relative success in automatic sentiment analysis is mostly due to the ability to handle nonstandard spellings and booster methods (p. 2555). Being able to decode misspelled words and correct them as well as determine whether sentiment is behind the misspelling is what makes SentiStrength a tool designed for short informal text, such as that found on social media.

Moving into further specialization of opinion analysis, Thelwall and Buckley (2013) added features to SentiStrength to optimize it for topic-specific sentiment analysis.

They introduce mood setting, which is the attempt to assign a positive or negative mood based on punctuation and repeating letters (p. 1609). Since they do not formally define mood in this paper, the reader is left to assume that mood refers to an affective state, that is to say there is some emotion in the text, based on the definition from Thelwall et al.'s (2010) introductory paper for SentiStrength (p. 2547). This feature seems to have improved performance when applied to two different Twitter data sets, although the limitations of the data only show the efficacy of this feature for some topics (Thelwall & Buckley, 2013, p. 1614). The second feature the authors tested was lexical expansion, which is adding topic-specific words to the SentiStrength lexicon's core 2,608 words and word stems (p. 1609). Though this modification did not prove as successful, the lexical expansion is recommended for social media data with a focus on a particular topic (p. 1615). An important conclusion from this study is that specialized tools and features require manual work: "Human labour seems likely to be particularly important for narrowly-focused topics for which small misclassifications may result in significant discrepancies if they are for terms that are frequently used with regard to a key aspect of the topic" (p. 1615).

Finally, Pang and Lee (2008) discuss the broader implications of sentiment analysis. Many businesses and government agencies might be interested in using these tools in order to improve service, reputation, products, etc. However, it is important to keep in mind that these tools should be used responsibly. Pang and Lee (2008) do not comment extensively on the matter of privacy (p. 55) but given allegations about the NSA that surfaced after the publication of their paper, it is a matter worth acknowledging. Americans have become wary of government and private companies collecting personal data, which can include opinions. While Twitter users in particular know they are posting on a public forum, using sentiment analysis on more private data can become unethical. For example a company might monitor their employees' emails and use sentiment analysis to see how employees feel about certain topics, such as newly acquired software systems. The company's intent might not have been malicious but the invasion of assumed privacy is problematic.

Additionally, opinion mining can lead to manipulation by companies or users. One user can create multiple profiles to make it seem as if a product is bad or a company can create multiple profiles to give itself positive reviews. As Pang and Lee (2008) state, "Indeed, there has already been a term — "sock puppet" — coined to refer to ostensibly distinct online identities created to give the false impression of external support for a position or opinion…" (p. 56). Perhaps one company can manipulate sentiment for a competitor, which can drive their stock prices down. Another possible scenario is a company who uses sock puppets to inflate their company's value based on popular sentiment. While it is not clear how exactly these manipulations would affect the economy of these companies, the point still stands that it is a possibility.

Overall, this field of sentiment analysis and opinion mining is progressing rapidly and becoming very popular. There are so many practical outcomes (market research, government use, political analysis, etc) that are scalable. Individuals can use tools and systems to analyze large-scale sentiment and do not have to read discrete reviews or posts, especially if there is a prohibitive amount. There seem to be many applications of the field to a vast number of users - private companies, government, consumers, education, religious organizations, market research companies, etc. Reviews, social media, documents and other types of sentiment analysis seem to be extremely valuable. Given big data's rise in recent years, the problem for these users is not getting the data but rather making sense of it. Sentiment analysis is definitely a field that is growing to build tools to meet the challenge of making big data useful. The next steps in sentiment analysis are to master determining irony and sarcasm and to create industry-specific tools that are at least as effective as broad approaches and human analysis.

Currently, there are many tools for sentiment analysis, both lexicons and software programs, as described above. Many methods are beginning to specialize into more specific fields as Thelwall and Buckley (2013) did with SentiStrength. Although no longer available, a sentiment analysis approach for sports was attempted with Sentibet[7] in 2012. The purpose of this software was to provide more information for sports betting. The aim of this sentiment-based forecasting service was to provide real-time sentiment analysis of social media to predict the outcome of upcoming sporting events. Additionally, sentiment analysis is expanding into other languages, as with Carvalho et al. (2009). SentiStrength also offers their dictionaries in several languages, including Irish, French, Indonesian and Japanese. Sentiment analysis is on the radar of many companies. From reviews to social media, institutions are researching how best to use sentiment analysis to their advantage. One example is the recent use of sentiment analysis in an effort to predict Oscar winners.[8] While effort could be improved as they did not take into account negative sentiment and only predicted 2 out of 5 categories correctly, Shine Communication states "using sentiment analysis in this way has become one of the agency's key resources" (Aron). This is one example of how sentiment analysis is constantly being researched and used in new ways.

[3] www.liwc.net
[4] www.liwc.net
[5] http://sentiwordnet.isti.cnr.it/
[6] http://en.wikipedia.org/wiki/YesAllWomen
[7] http://www.neurolingo.gr/en/node/175
[8] http://www.prweek.com/article/1334632/shine-communications-uses-sentiment-analysis-predict-oscar-winners

3. Methodology

The purpose of this research is to analyze how sentiment expressed on social media relates to basketball games characteristics using SentiStrength for the sentiment analysis. My research question is:

1. Is there a correlation between the average sentiment strength of fan Tweets and the outcome of a game?

More specific questions focus on where sentiment might be influenced during a game.

1a. Is sentiment stronger or weaker during close games?
1b. Is sentiment stronger or weaker during blowout games?
1c. Do sentiment strength and polarity stay the same over all quarters?
1d. Do events affect sentiment strength?

My operational definitions for this project are as follow:

- *Close game* - a game where the score has a difference of 5 points or less 1 minute before the end of the game.
- *Blowout* - final score is 15 or more points in difference
- *Average* - the average of the added output of Sentistrength's scores for each Tweet. Sentistrength's scores are given as coordinates of positive and negative numbers and I then convert them to one score by adding the two together. I take the average of these scores, which is what will be used to answer the research question. Average of both positive values and negative values will also be calculated.
- *Polarity* - whether the sentiment is positive (+) or negative (-).
- *Sentiment strength* - strength is defined on a scale from -5 to 5. -5 is the strongest negative sentiment, while 5 is the strongest positive sentiment. Strong sentiment is defined as 4 and 5 or -4 and -5.
- *Neutral* – in terms of SentiStrength, a score of -1, 1 is considered as having no sentiment and therefore is neutral (*SentiStrength)*.

I followed one basketball team for 30 games. I collected Tweets during each game that contained the official team hashtags as used on the official team Twitter feed using

Twitter's streaming API. Afterwards, I parsed out necessary information from the collected data, removed Retweets and the hashtag symbol (#), categorized Tweets by the quarter in which they were sent, and removed Tweets sent before the start of the game and after the end of the game. Once the data was ready, I analyzed Tweets using SentiStrength to determine the sentiment polarity and strength of each Tweet.

3.1 Data Analysis and Approach

Twitter Assumption. I echo Pak and Paroubek's (2010) assumption that a Tweet is about one topic, in this case the basketball game, and contains sentiment about this one topic (p. 1321) in order to analyze each Tweet as a sentiment related to the basketball game.

Team and Game Selection. I selected the Chicago Bulls because it is a popular team that I have followed for years so I was familiar with players, the coach and their fan base. Tweets from 33 out of 82 possible games of the 2014-15 season (~40%) were collected. However data corruption occurred in 3 of them, bringing the total number of games with complete data to 30, which is a common minimum for statistical analysis. Since these games are a third of the season, they are enough to represent the population (all 82 games). These games were collected during the regular season so they do not represent the post-season (playoff games). Because it is possible that there are differences between regular and post-season games, I only collected regular games.

Identifying Tweets. In order to ensure that I captured data about the games, I decided to collect Tweets using hashtags. Using keywords would have required more ambiguous terms, such as Bulls or Chicago that might have referenced other events or teams. The hashtags I used for collection are those that the official Bulls Twitter page uses for games. Because users who follow the official account are most likely fans of the team, those who Tweet using these hashtags are assumed to be mostly Chicago Bulls fans. The hashtags also differ from other teams' hashtags slightly. The Bulls use the pattern of TEAM at OTHERTEAM for their hashtags. For example, #ORLvsCHI represents a home game between the Bulls and the Orlando Magic. ORL is the official abbreviation of the Orlando Magic team, as CHI is for the Chicago Bulls. On home games the pattern is TEAM at CHICAGO BULLS, so the hashtag is #ORLvsCHI. If the game was hosted in Orlando, the hashtag would be #CHIvsORL. Other teams have similar variations to indicate which game their official page is talking about. For example, the Washington Wizards use the template #WizWarriors to indicate that the team is playing the Golden State Warriors.

I decided to use the official hashtag for each game because the hashtags do not have any sentiment and are unambiguous in identifying the specific teams playing on that date. For example, #Bulls could refer to other topics or teams (e.g. Durham Bulls or Vodacom Bulls: "Love my bar light! #friends #present #Bulls #SuperRugby“). I am also operating under the assumptions that using an official hashtag would get positive and negative emotion-bearing Tweets whereas Bulls-specific hashtags such as #BullsNation might only return positive Tweets. Finally, it was easy to adapt for different games and made sure the focus of the Tweet was, at least partially, about the game.

Tweet Collection. The first concern was to be sure to collect all Tweets for each game. Collection started before the game start time (anywhere from 30 minutes to a few hours before to provide some buffer time to check that the script was working) to make sure I collected all the Tweets from the first quarter, and ended after the game's end to ensure fourth quarter or any overtime Tweets are collected throughout.

I used the Twitter streaming API[9] to collect data. I chose not to use Twitter's REST API because this method uses rate limiting, whereas there is no rate limiting in the streaming API. Due to rate limiting searching after the event, it was essential that I collect data as the events were happening. By collecting Tweets from games as they were happening (streaming), I was able to get all Tweets with my chosen hashtags during the games. Tweets were gathered as JSON objects via a Python script using Tweepy[10] a wrapper for the Twitter API.

Data Cleaning. Once I had collected Tweets from games, I ran the resulting text file output through a Python script I wrote to extract necessary attributes and remove unwanted aspects. Specifically, the script

- Parsed out and stored time of Tweet
- Parsed out and stored text of Tweet
- Removed Retweets
- Removed the hashtag symbol (#)
- Removed non-English Tweets

Other than removing the # symbol, hashtags were not modified or separated in any way to avoid incorrectly assigning sentiment by the program to ambiguous hashtags that are multiple words. For example, #Roaracle is a hashtag used by a user during the December 6th game against Golden State Warriors in Oracle Arena. This can be split up as 'roar Oracle' and could be construed as a cheer for Golden State. However, it could also be a GSW cheer that does not need to be split up. Because of this, I decided not to modify any phrases used in hashtags, including breaking up regular words in camel case (e.g. #BullsNation or #SeeRed).

Game Metadata. After every game, I recorded the game metadata needed to answer the research questions, which consisted of

- Opponent
- Date of game
- Winning team
- Halftime score
- Final score
- Starting Time
- Whether it was a blow out or not
- Whether it was a close game or not
- Whether there was an overtime or not
- Notable events such as injuries or technical fouls
- End times of each quarter

There were several options to get the end quarter times for each game. I could sign up for alerts that text me at the end of each quarter, I could use notifications of an application and record the times each notification appeared, or I could find a source that had recorded this information after the fact. While testing several methods on games before my data collection, I noticed that both texts and notifications were slow to note the end of the quarter. I found that sources that kept a record of times in their play-by-play records only noted the time of the game clock. For example, a timeout was taken at 4:31 into the first quarter. Additionally, this did not account for commercial breaks so there was no way to figure out how long the quarter had lasted.

While looking at Twitter data before my experiment, I found that to get the end times, the official Bulls' Twitter feed was consistent and timely. When deciding the best way to note end of quarters, I found the Bulls Twitter feed to be the fastest to post about the end of the quarter than notification and texts, which took a few minutes longer. The feed posted after each quarter that the quarter was over and gave the score, usually with player statistics as well. For example, this Tweet comes from the January 25 game against Miami Heat:

At the half, Bulls trail the Heat 48-38. Rose 6pts, Gasol leads with 8pts 9reb. #MIAvsCHI.

In another example from January 30 against the Phoenix Suns, the feed does not use any hashtags:

After one, Bulls trail the Suns 26-23. Butler 7pts, Gasol 6pts 3reb, Rose 4pts.

Because the Twitter feed does not always use a hashtag or a hashtag that I have specified, it is not always collected with other Tweets, so I visited the page after every game to get the times.

Sentiment Analysis. After each game, Tweets were analyzed using the SentiStrength program (Thelwall & Buckley, 2013) described in the previous section. SentiStrength returns two sentiment scores, a positive (1 to 5) and a negative (-1 to -5), as well as an explanation of its process for each line. For example, the Tweet

Offensive rebounds are killing us. CHIvsORL.

has a score of (1, -3). SentiStrength shows how it determined this:

Offensive	[-2]
rebounds	[0]
are	[0]
killing	[-1]
us.	[0]
[Sentence=-3,1=word max, 1-5]	
CHIvsORL.	[0]
[Sentence=-1,1=word max, 1-5]	
[1,-3 max of sentences]	

Table 1. Details of sentiment analysis by SentiStrength of one Tweet with two sentences.

The program divides this Tweet into two sentences based on the punctuation. Starting with the first sentence, it gives “offensive” a score of -2 and the word “killing” gets a score of -1 based on the dictionary score of these words, while the rest of the words were found to have no sentiment. After each sentence SentiStrength gives a sentence score. The sentence score is calculated by taking the maximum emotion expressed in the sentence and adding it to the neutral score of -1 and 1. For the first sentence, the score was -3 because “offensive” had the highest negative score in the sentence (-2) and that score is added to -1, giving a sum of -3 for the sentence: “[Sentence=-3,1=word max, 1-5].” The second sentence was simply CHIvsORL, which had no sentiment according to the program. After that sentence the user sees “[Sentence=-1,1=word max, 1-5].” Because there was no positive or negative sentiment, the max score is 1 and -1 for the sentence since those are the neutral scores and the maximum score to add is 0. After the entire document is analyzed, SentiStrength gives a final score for the document (in this case a Tweet): “[1,-3 max of sentences].” The final score of a document takes the max score of the sentences and uses that score.

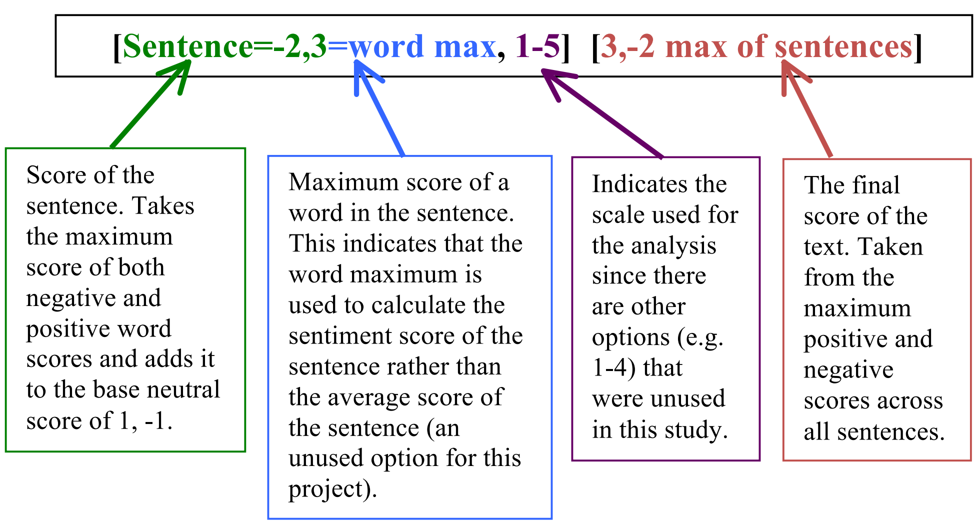


Figure 1. Explanation of SentiStrength’s scoring output per text line (one Tweet).

In another example, the Tweet

126-120 Bulls trying to close this one out against a very good raptors team CHIvsTOR is scored as (3, -2). The analysis shows booster words had an impact here:

126	[0]
120	[0]
Bulls	[0]
trying	[0]
to	[0]
close	[0]
this	[0]
one	[0]
out	[0]
against	[-1]
a	[0]
very	[0]
good	[1][1 LastWordBoosterStrength]
raptors	[0]
team	[0]
CHIvsTOR	[0]

Table 2. Breakdown of sentiment analysis by SentiStrength

Here, the negative score is easy to calculate. There is only one negative word ("against") so the score from that (-1) is added to the neutral score -1 and a final negative score of -2 is reached. For the positive score, the program uses a booster word. "Good" has a score of +1 but because the term "very" appears immediately before, the score is boosted by +1. This means that the booster word score is added to the positive term score, giving it a score of +2. This makes is the maximum (albeit only) positive term in the sentence so it is added to the neutral 1 all sentences begin with. Adding the boosted "good" gives us +3, so the final sentence score is (3, -2) as explained in Figure 1.

The version of SentiStrength used for this research does not feature an expanded lexicon from the original SentiStrength word lists. Because of this, SentiStrength evaluates text for sentiment only using the standard words and scores of the program. Therefore slang and jargon is not weighed properly. For example, the Tweet

Nastyyyyy shot Taj..... #Bulls #ChiVsHou #WindyCity is counted as negative, although in this sport it is interpreted as a positive statement, especially given that the player this person is referring to is on the team referenced in a hashtag. One of the ways to combat this for future analysis is for researchers to create topic dictionaries that can be added to expand lexicons.

Calculation of Game Sentiment. Once the Tweets from all the games were analyzed, averages were calculated to represent the sentiment associated with each game. The statistics calculated are described in Table 3.

Stat Name	Stat Definition	Stat Calculation
Positive Average per quarter	Each quarter's positive scores average	Adding each quarter's positive score and dividing it by the number of Tweets in the quarter
Negative average per quarter	Each quarter's negative scores average	Adding each quarter's negative score and dividing it by the number of Tweets in the quarter
Overall average per quarter	Each quarter's added scores average (adding together positive and negative scores)	Adding each quarter's added average score and dividing it by the number of Tweets in the quarter
Positive average per game	Each game's positive scores average	Adding each game's positive score and dividing it by the number of Tweets in the game
Negative average per game	Each game's negative scores average	Adding each game's negative score and dividing it by the number of Tweets in the game
Overall average per game	Each game's added scores average	Adding each game's added average score and dividing it by the number of Tweets in the game

Table 3. Statistics calculated for analysis to answer RQ1.

The score of each Tweet determines sentiment strength. For this project, I consider scores of 4 to 5 and -4 to -5 as having strong sentiment and Tweets with scores of 2 to 3 and -2 to -3 as having weak sentiment. Sentiment strength calculations were used to answer research questions 1a and 1b, based on the percentage of Tweets with strong sentiment in a quarter and in a game. For example, for the December 22, 2014 game against the

Toronto Raptors, about 3.4% of the Tweets during the first quarter had strong sentiment, while during the entire game about 2.9% of the total Tweets had strong sentiment.

Calculating Correlation. For this study, I used the quantitative approach of correlation analysis to identify significant relationships between sentiment polarity and strength and game characteristics. Outcome of games were coded as 0 for losses and 1 for wins. Other game characteristics included outcome of game quarters, and the occurrence of unusual events, such as overtime, or injury. Using correlation and correlation coefficients to determine relationships and strength of relationships is a classic quantitative approach (Byrne, 2007, p. 43). Correlation analysis measures the degree to which two variables are associated (Moutinho , 2011, p. 57). More specifically, I used the Pearson correlation coefficient to measure the linear association two variables. If the correlation coefficient is positive, it shows a tendency for high values in one variable to be associated with high values in the second variable (Moutinho , 2011, p. 57). Alternatively, a negative correlation shows an association between a high value in the first variable with a low value in the second variable, or that they are inversely associated. Since my data represents a sample of all Tweets about basketball games, I use the sample correlation (r), which always lies between -1 and 1 (Moutinho , 2011, p. 58). If $r = 1$, the two variables have a perfect positive linear association, whereas $r = -1$ shows a perfectly negative linear association. If the correlation coefficient is 0, no association is found. For this paper, I decided to use Quinnipiac University's strength of correlation scale.[11] According to this Pearson's Correlation scale, a correlation of r = 0.7 is considered very strong positive relationship, while r = 0.2 is considered weak positive relationship.

For this paper, I used a significance level of α = 0.05. This value indicates that there is only a 5% chance that the correlation found has been found by chance. Correlations with a significance probability of p= 0.05 or less indicate that there is sufficient confidence to reject the null hypothesis, which is that there is no correlation between two variables (Moutinho, 2011, p. 59). These measure are extremely common statistical tests used for determining the association between two variables. Because my questions ask for correlation between a number of two-variable sets, these tests are the most appropriate for the analysis.

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$$

Figure 2. Pearson Correlation Coefficient formula.

A benefit of using Twitter data for this research is that the process to collect it is well known (the Twitter Streaming API) and can be replicated. Although the Tweets will change, the process is extremely transparent and replicable. I used correlation to note the existence of any relationships between outcome and sentiment to answer my main research question (1), strength of sentiment and close games to answer research question 1a, strength of sentiment and blowout games to answer research question 1b, and sentiment and quarters to answer question 1c.

Limitations of methods: For this study, no emojis or emoticons were taken into account. Not only does SentiStrength lack the capacity to process emojis, the situation in which they are used does not always make it clear whether the emoji is positive or negative. For example, some Tweets used an emoji of an ox, which can be taken as positive or negative depending on a person's own judgment. Since it looks like a bull, it might simply indicate which team the user is supporting. Some Tweets simply have a basketball and hoop emoji that seems to signify that they are watching basketball and has no obvious sentiment.

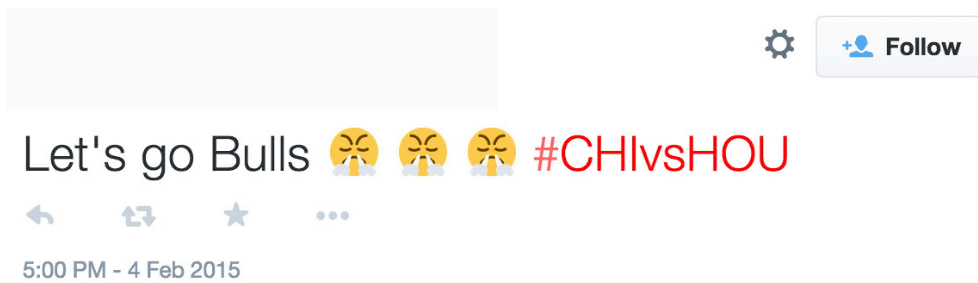


Figure 3. Example Tweet from February 4, 2015 game against Houston Rockets containing emojis.

In the example in Figure 3, the emojis look like angry faces but are coupled with a cheering sentiment. If I were to use them to extrapolate polarity, I would classify the emoji as negative because of the number of times the angry emoji is expressed. Rather than guessing at sentiment, I did not take into account emojis and therefore also had to exclude emoticons. Larger examples of some basketball related emoji are in the Appendix (Figure 4).

I did not collect all Tweets about a game because some relevant Tweets would have other hashtags and some would not have any hashtags. However, the use of other hashtags is more unpredictable, and could vary widely from game to game. Using only the official hashtag allowed me to have a consistent collection strategy across games. That means that I have only gotten a portion of all Tweets per game.

Additionally, the tool I use, SentiStrength, does not always assign sentiment. For example, while "Let's Go" can be construed as a positive sentiment to cheer on the team, SentiStrength's algorithm does not capture it as sentiment-bearing at all in this case. This is an example of domain-specific expression of sentiment. Another example is

JIMMY GETS BUCKETS AND BOARDS BKNvsCHI HopCity which was analyzed as (1, -1) meaning it does not convey sentiment (as explained in operational definition of neutral). However, the fact that the Tweet is in caps shows some excitement about Jimmy Butler (whose fans nicknamed Jimmy Buckets).

Examination of the results shows that most analyses made sense. For example, Tweets with no sentiment (1, -1) were usually updates for scores or time left in the game. For example, for the January 7th game against Utah Jazz, 53 of 101 (52%) neutral Tweets were updates on the score and players, e.g.

At the half Bulls leads the Grizzlies 51-43,

while the rest were observations:

You can hear booing at the UC. #Bulls #UTAvsCHI

and

Gordon Hayward gels his hair before the game #UTAvsCHI.

Others were Tweets with emojis and hashtags such as

#BullsNation #UTAvsCHI #WeRunThangz #Chillin,

or Tweets that simply do not contain any emotion-bearing words:

There we go Rose #CHIvsUTA.

There were also some occurrences of non-Bulls fans also using the hashtags. While it did not happen often, the hashtags I collect are predictable so there are some instances of fans rooting for the other team. The polarity would be reversed in this case much like an event commented on with negative sentiment by a Bulls fan would receive a positive sentiment by a fan of the opposing team. Some positive strong Tweets (scoring 4-5 or -4 to -5) included ones cheering on the opposing team, such as

Looks like one of the Chicago Bulls had a temper tantrum and got ejected oh well see ya #wearebrooklyn #netsonyes #BKNvsCHI".

[9] https://dev.twitter.com/streaming/overview
[10] http://www.tweepy.org/
[11] http://faculty.quinnipiac.edu/libarts/polsci/Statistics.html

4. Results

Tables showing data broken out by game are included in the Appendix. 14,440 Tweets were collected in total. Table 4 summarizes the Tweets collected, broken out by sentiment polarity and strength. Tweets scoring 1, -1 are neutral, Tweets with a score of 2 or 3 and -2 or -3 score as weak and Tweets with a score containing -4 to -5 or 4 to 5 are strong. Equal scores are Tweets scoring 2, -2 or 3, -3.

	Total	Max per game	Min per game	Average per game	Max per quarter	Min per quarter	Average per quarter
Strong (all)	387	28	2	12.9	12	0	3.225
Strong +	110	18	0	3.66	6	0	0.91
Strong -	277	20	1	9.23	10	0	2.30
Weak (all)	7444	632	59	248.13	323	5	62.03
Weak +	4892	381	39	163.06	159	2	40.76
Weak -	2538	287	20	84.6	164	0	21.15
Neutral	5954	471	51	198.46	162	7	49.61
Equal	655	62	5	21.83	34	0	5.45

Table 4. Summary table showing the number of Tweets collected (N = 14,440)

	Positive Average	Negative Average	Added Average
First quarter	1.453986455	-1.303114137	0.150872215
Second quarter	1.482129897	-1.387748036	0.094371032
Third quarter	1.468652532	-1.413873227	0.054779271
Fourth Quarter	1.526322573	-1.392188029	0.134134567
Overtime	1.537474737	-1.330404018	0.207070708
Double overtime	1.501901141	-1.467680608	0.034220532
Overall average	1.495745764	-1.383344707	0.112398871

Table 5. Positive, negative and added sentiment averages per quarter.

I collected Twitter data from 33 games. Three of the games had corrupted data and therefore were unusable: December 29, 2014 against Indiana Pacers, January 12, 2015 against Orlando Magic and January 27, 2015 against Golden State Warriors. After taking out these games, I was left with 30 games to analyze. The minimum number of Tweets per game is 123 Tweets, the maximum number of Tweets per game is 1136 Tweets, and the average number of Tweets per game was 480 Tweets.

I collected the end times of each game quarter and then divided each game's Tweets into the quarter during which they were expressed. The average number of Tweets per quarter was 115 Tweets. I classified games by my operational definitions. After looking at the end score, I determined whether or not they fit the definition of close game or blow out game. Games 7, 14, 15, 26, 27 and 30 were close games. Games 2, 4, 10, 16, 17, 20, 23 and 29 were blowout games. I also noted any important events in games 12, 18, 23, 26, 27 and 29. Since there are many events one might consider important during a basketball game, I used my subjective judgment to determine which ones might influence sentiment.

Table 5 shows the average positive sentiment, average negative sentiment and average added sentiment of each quarter as well as the sentiment averages per game. Averages for overtime and double overtime are also included.

Average number of Tweets/game	480
Average number and percent of strong Tweets per game	13 (2.66%)
Average number and percent of weak sentiment or no sentiment Tweets per game	468 (97.33%)

Table 6. Aggregated numbers of sentiment strength.

After the data calculations, I calculated correlations using Pearson's correlation coefficient and the statistical program JMP. The results are recorded in Table 7.

Twitter Characteristics	**Game Characteristics**	**Correlation**	**Significance**
Average added sentiment	outcome of game	**r = .58**	**p = .0009**
2nd quarter added average	outcome of game	**r - .45**	**p = 0.01**
4th quarter added average	outcome of game	**r = .72**	**p = .000008**
strength of sentiment (%)	outcome of game	**r = -.36**	**p = 0.05**
halftime outcome (whether the Bulls are ahead or behind at halftime)	outcome of game	**r = .36**	**p = 0.05**
total number of Tweets per game	Close games	r = .1	p =.5
Average added sentiment	halftime outcome	r = .22	p = .24
Average added sentiment	Close games	r = -.2	p =.28
Average added sentiment	Blowout games	r = .26	p = .17

Table 7. Table of correlation and significance scores. Significant results are bolded.

There was significant strong positive correlation between the average added sentiment, either positive or negative, of fan Tweets and the outcome of the game (r = .58, p = .0009). In other words, Bulls fans issued generally positive Tweets during games which their team ultimately won, and generally negative Tweets during games which their team ultimately lost. Fans reflected the outcome of the game by Tweeting about how well the game is going during wins or about aspects of the game that are not going well during losses. For example, consider this Tweet from the December 22, 2014 win against the Toronto Raptors:

I've always loved Aaron Brooks' style of play. But it's so good to see the @chicagobulls give him a decent run. @thirty2zero TORvsCHI.

This Tweet scored (-1, 4) showing it contains a strong positive sentiment about the player Aaron Brooks and the team Chicago Bulls. During a losing game, a fan expressed frustration by Tweeting

If the fucking Bulls lose to the Suns imma be real fucking pissed!! CHIvsPHX.

This scored the maximum negative score (-5, 1) and likely reflects that the game is going poorly for the Chicago Bulls.

There was strong positive correlation between fourth quarter sentiment average and outcome of the game ($r = 0.72$, $p = 0.000008$). This suggests that by the end of the game, fans could tell which way the game was heading. For example, a Tweet during the fourth quarter of a loss on January 7, 2015 against the Utah Jazz says

This might be the worst game I've watched in the Rose era. So, so awful.
Bulls Jazz UTAvsCHI.

This Tweet scored (-4, 1) for sentiment. By the fourth quarter, fans have been watching their team struggle or do well and express opinions on what they have seen not just during the last quarter but also throughout the game. Having the knowledge of 36 minutes of play before the last quarter starts gives fans context to understand where the game is going and how likely it is that their team will be successful.

There was a significant negative correlation between strength of sentiment (percentage of sentiment that is strong) and outcome of a game ($r = -.36$, $p = 0.05$). This shows that in some cases sentiment was stronger throughout the game when the team lost. As the majority of strong sentiment expressed during the games was negative (Table 4), the negative correlation makes sense.

There was no significant correlation between close games and average added sentiment ($r = -0.2$, $p = 0.28$). This means that sentiment was not significantly stronger or weaker during close games. The lack of relationship is somewhat surprising because of the intense nature of close games. Furthermore, there was no correlation between close games and total number of Tweets per game ($r = 0.1$, $p = 0.5$), meaning that the lack of correlation between close games and sentiment strength is not due to a decrease in Tweets during the game. Fans did not stop Tweeting to watch the action.

There was no significant correlation between blow out games and average added sentiment ($r = .26$, $p = 0.17$). This means that sentiment was not significantly stronger or weaker during blow out games.

Two negative events were noted during data collection. The first happened during the December 30, 2014 game against the Brooklyn Nets. One of the Bulls players, Aaron Brooks had two technical fouls and was ejected from the game in the second quarter. The second negative event was during the February 7, 2015 game against the New Orleans Pelicans. One of the Pelicans' players, Anthony Davis suffered a severe shoulder injury during the second half. In my data, I found that negative events affect sentiment strength more than positive events; for the two negative events that occurred most of the strong sentiment Tweets were about the event throughout the game. In the case of the game ejection all strong sentiment Tweets during that quarter were about the ejection and were all negative. 35 out of 93 Tweets (36%) during the quarter also referenced the player's ejection, the officiating or both showing that this particular event had a lot of reaction, although the sentiment behind the Tweets was not always strong. For the injury (of a player on the opposing team), 2 out of 8 strong sentiment Tweets (25%) were about the injury and were negative. No strong sentiment Tweets were about the positive events (career high points or All-Star selection).

A large amount (43%) of the strong sentiment Tweets were about the overall state of the game, for example:

What an ugly, ugly game. #BOSvsCHI #Bulls #Celtics.

Some were about a particular play or combination of plays and effort by certain team members. For example, *Loving the energy! Keep it up! Yes, #Gasol! #CHIvsBOS @NBA #SEERED @ChicagoBulls*

and

Loved that pass!! #BOSvsCHI.

Overall, negative sentiment was the strongest with 78% of strong sentiment Tweets being negative (302 out of 387).

5. Discussion

Based on these results, people who were Tweeting during the game, especially in the fourth quarter, matched their sentiment to the game action. That is, people were positive during games that the Bulls won and were negative during games they lost. This shows that the fans that Tweeted were invested in the games. Their participation is important to the team. Not only does fan participation keep the team relevant to fans and the league, marketing the team relies on fans being enthusiastic fans. They are paying attention to the game and are treating it as an interactive experience. It does not seem to matter whether the game is close or a blowout because people were still interested in participating in the game-watching experience. Watching the game could be more about expressing themselves to others and using it as a social context that allows them to communicate to friends or other fans.

The large amount of negative sentiment found during the games could partly be because SentiStrength is better at detecting negative sentiment than positive sentiment (Thelwal, et al., 2010, p. 2544). This could also be explained by people's tendency to express negative opinions more than positive ones. As Baumeister et al. (2001) discover, the principle that "bad is stronger is good" is prevalent in many contexts of human life and behavior. When detecting emotion in faces, Baumeister et al. found that threatening faces were detected more quickly and accurately than happy faces (p. 342). The authors also concluded that people tend to spend energy on avoiding negative experiences than pursuing positive ones. This idea could influence how people react when they see something they perceive as negative, such as a losing game of their favorite team. Baumeister et al. argue that “bad events will have longer lasting and more intense consequences than good event” (2001, p. 325). Therefore it would be expected that when negative events are perceived, people are likely to react stronger to those rather than positive events.

Fans whose Tweets were included in this study seem to be following the game because it is an interactive experience for them. They notice how well the team is playing and express frustration or happiness at how the team is doing. Given the prevalent expression of negative sentiment, many events can trigger these, while only a few can trigger positive sentiment. Since people are more likely to dwell on negative experiences, it is important to keep fans excited about the team even when they are losing.

Regarding the lack of correlation between blowout games and average added sentiment, fans might be just as excited to watch these games as any other games just because they enjoy watching the team. They might still want to express their opinions on what they are seeing even if they can tell how the game will end.

Regarding the lack of correlation between close games and average added sentiment, perhaps during close games, fans simply did not use the hashtag so the data was not collected. Including the hashtag is more time consuming than leaving it out so in order to get back to the action, perhaps fewer people included it. In the excitement of the close game, fans might want to express opinions quickly to get back to the action so they do not miss important plays. This question could be addressed in future research by expanding the data collection strategy.

Interestingly, there was some significant strong positive correlation between average sentiment of the second quarter and outcome of game ($r = .45$, $p = 0.01$)**.** There is no correlation between halftime outcome (whether the Bulls are ahead or behind at the end of the second half) and overall sentiment. ($r = 0.22$, $p = 0.24$). However, there is a moderate correlation between game outcome and halftime outcome ($r = 0.36$, $p = 0.05$). This means that there is some relationship between which team is ahead at halftime and which team wins. This correlation could explain the correlation between second quarter sentiment and outcome of the game. Some teams attempt to close out the second quarter the way they would the fourth quarter so that they are ahead after returning from halftime. This push in effort in the second quarter could be mimicked at the end of the game by the team so if the team is successful in the second quarter, fans could react to the effort in both the second and later in the fourth quarter.

6. Conclusion

Based on my observations during this project, SentiStrength is a good basic system that could take customization well. It worked well for informal text due to the ability to parse misspellings and alternate spellings. Another positive attribute of the software is the ability to add dictionaries and other supplemental lists for customization. Out of the box, however, SentiStrength did have some difficulty with the domain. I also found that the emotion word dictionary that was used weighed obvious words too strongly (e.g. sad is -4) despite the fact that the colloquial use is not so severe. This calls for more research and reform to informal emotion dictionaries and how they differ from document emotion dictionaries.

Future investigation would be improved by using a tool more tailored to Twitter and the domain of sports in general and basketball in particular. Dictionaries adapted to word use on social media would fare better than a dictionary that is meant to be applicable more generally. Furthermore, looking at language specifically used in sporting events would improve sentiment analysis of this domain. Narrowing the scope to a particular sport would result in even more accurate emotion word usage due to sport-specific slang and references. Other studies to consider would be language used during team sporting events versus individual sporting events, such as tennis.

Watching a game at home is not the same as watching it in the stadium. Based on the correlation between game outcome and fan sentiment, it is clear that fans are participating in the game experience even if they are not present at the event. Expressing these opinions brings them closer to the experience of being in the stadium, watching with fellow fans and feeling like part of the culture. Using social media, fans take a step toward inclusiveness. Since not everyone can get to the game because of budget, location, availability or a myriad of other obstacles, live-Tweeting the game gets fans participating in the experience. They are socializing through these sentiments and bridging the gap with other fans.

Sentiment analysis is rapidly growing in many industries. Entertainment, sports, manufacturers and retailers are using sentiment analysis as a tool to improve their business. Sports team owners should pay attention to these results and would be wise to consider doing their own studies with more variables and data, especially for sports that are looking to expand and engage more fans. Using sentiment analysis, teams can determine where they are engaging fans, where negative sentiment is coming from and which events can inspire positive sentiment.

7. Appendix

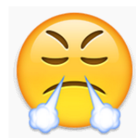
"Basketball and hoop"	"Face of Triumph"	"Ox"
		

Figure 4. Examples of Unicode emojis found in Tweets.

Table 8. Collected Game Metadata. Games 31-33 (red) represent the games with corrupted data that were not used in the analysis.

ID	Opponent	Date	Winner	Score
1	Golden State Warriors	6-Dec	GSW	112 to 102
2	Brooklyn Nets	10-Dec	CHI	80 to 105
3	Portland Trailblazers	12-Dec	CHI	106 to 115
4	Miami Heat	14-Dec	CHI	93 to 75
5	Atlanta Hawks	15-Dec	ATL	86 to 93
6	New York Knicks	18-Dec	CHI	97 to 103
7	Memphis Grizzlies	19-Dec	CHI	103 to 97
8	Toronto Raptors	22-Dec	CHI	120 to 129
9	Washington Wizards	23-Dec	CHI	99 to 91
10	Los Angeles Lakers	25-Dec	CHI	93 to 113
11	New Orleans Pelicans	27-Dec	CHI	100 to 107
12	Brooklyn Nets	30-Dec	BKN	96 to 82
13	Denver Nuggets	1-Jan	CHI	101 to 106
14	Boston Celtics	3-Jan	CHI	104 to 109
15	Houston Rockets	5-Jan	CHI	105 to 114
16	Utah Jazz	7-Jan	UTA	97 to 77
17	Washington Wizards	9-Jan	WAS	86 to 102
18	Milwaukee Bucks	10-Jan	CHI	95 to 87
19	Washington Wizards	14-Jan	WAS	105 to 99
20	Boston Celtics	16-Jan	CHI	119 to 103
21	Atlanta Hawks	17-Jan	ATL	107 to 99
22	Cleveland Cavaliers	19-Jan	CLE	108 to 94
23	San Antonio Spurs	22-Jan	CHI	81 to 104
24	Dallas Mavericks	23-Jan	CHI	102 to 98
25	Miami Heat	25-Jan	MIA	96 to 84
26	Los Angeles Lakers	29-Jan	LAL	123 to 118
27	Pheonix Suns	30-Jan	PHX	99 to 93
28	Houston Rockets	4-Feb	HOU	90 to 101
29	New Orleans Pelicans	7-Feb	CHI	107 to 72
30	Orlando Magic	8-Feb	CHI	98 to 97
31	Indiana Pacers	29-Dec	CHI	92 to 90
32	Orlando Magic	12-Jan	ORL	121 to 114
33	Golden State Warriors	27-Jan	CHI	113 to 111

Table 9. Collected Game Metadata. Games 31-33 (red) represent the games with corrupted data that were not used in the analysis.

ID	At haltime	Close Game?	Blowout?	Overtime?
1	61-49 (gsw)	no	no	no
2	51-51	no	yes	no
3	51-59 (por)	no	no	no
4	39-32 (chi)	no	yes	no
5	44-50 (atl)	no	no	no
6	52-45 (chi)	no	no	no
7	51-23 (chi)	yes	no	no
8	66-60 (tor)	no	no	no
9	40-46 (chi)	no	no	no
10	47-48 (chi)	no	yes	no
11	45-49 (chi)	no	no	no
12	55-45 (bkn)	no	no	no
13	53-42 (den)	no	no	no
14	40-46 (chi)	yes	no	yes
15	62-62	yes	no	no
16	36-32 (uta)	no	yes	no
17	42-59 (was)	no	yes	no
18	39-48 (chi)	no	no	no
19	44-50 (chi)	no	no	no
20	55-58 (BOS)	no	yes	no
21	48-39 (atl)	no	no	no
22	39-54 (cle)	no	no	no
23	40-46 (chi)	no	yes	no
24	54-51 (chi)	no	no	no
25	48-38 (mia)	no	no	no
26	48-59 (lal)	yes	no	yes
27	42-55 (PHX)	yes	no	no
28	48-57 (hou)	no	no	no
29	48-39 (chi)	no	yes	no
30	50 to 45 (bulls)	yes	no	no
31		yes	no	no
32		no	no	no
33		yes	no	yes

Table 10. Game Events table. Games 31-33 (red) represent the games with corrupted data that were not used in the analysis.

ID	Other events
1	no
2	no
3	no
4	no
5	no
6	no
7	no
8	no
9	no
10	no
11	no
12	2 techs to Aaron Brooks for arguing with refs (2nd half; 7:41 report)
13	no
14	no
15	no
16	no
17	no
18	Pau Gasol hits career high 46 points at 34 years old
19	no
20	no
21	no
22	no
23	Pau Gasol voted as starter for all-star game (notification right before game)
24	no
25	no
26	double ot; gasol's first game in la after leaving (thank you shirts/videos); jimmy butler named to all star reserves
27	drose/butler dunk at 10:28
28	no
29	Anthony davis is injured
30	no
31	no
32	OSU vs ORE game also happened at this time
33	beat #1 team in the league at the time

.

Table 11. Percentage of Strong Tweets in each Game.

ID	Total Tweets per game	total strong Tweets	percentage
1	641	20	3.120124805
2	438	13	2.96803653
3	845	28	3.313609467
4	457	12	2.625820569
5	304	2	0.657894737
6	312	6	1.923076923
7	345	5	1.449275362
8	670	18	2.686567164
9	175	4	2.285714286
10	1063	24	2.257761054
11	123	3	2.43902439
12	275	13	4.727272727
13	476	12	2.521008403
14	413	13	3.147699758
15	653	12	1.837672282
16	256	17	6.640625
17	665	23	3.458646617
18	273	3	1.098901099
19	494	10	2.024291498
20	134	4	2.985074627
21	258	12	4.651162791
22	548	19	3.467153285
23	449	9	2.004454343
24	790	15	1.898734177
25	1136	27	2.376760563
26	1101	27	2.452316076
27	340	7	2.058823529
28	390	16	4.102564103
29	192	8	4.166666667
30	209	5	2.392344498

Table 12. Number of strong, weak, and equal Tweets for each game. Equal sentiment represents Tweets that a) have emotion expressed in them and b) have an equal score for positive and negative sentiment (e.g. 2, -2 or 3, -3).

ID	strong total	weak total	neutral total	equal sentiment
1	20	304	300	17
2	13	208	197	21
3	28	435	355	27
4	12	229	197	19
5	2	153	134	15
6	6	153	144	9
7	5	164	166	10
8	18	367	253	32
9	4	100	62	9
10	24	512	471	56
11	3	59	51	10
12	13	149	101	12
13	12	275	162	27
14	13	205	177	18
15	12	397	210	34
16	17	125	101	13
17	23	317	290	35
18	3	122	139	9
19	10	285	177	22
20	4	65	60	5
21	12	127	105	15
22	19	260	242	27
23	9	212	215	13
24	15	398	344	33
25	27	632	415	62
26	27	592	431	51
27	7	197	136	13
28	16	196	156	22
29	8	98	75	11
30	5	108	88	8

Table 13. Number Tweets with strong positive, strong negative, weak positive, and weak negative scores for each game.

ID	strong positive	strong negative	weak positive	weak negative
1	5	15	199	105
2	1	12	133	74
3	10	18	326	109
4	2	10	163	66
5	0	2	79	74
6	2	4	92	61
7	1	4	129	35
8	3	15	256	111
9	3	1	79	21
10	16	8	381	131
11	1	2	39	20
12	2	11	89	60
13	5	7	162	113
14	4	9	146	59
15	2	10	280	117
16	2	15	70	55
17	3	20	183	134
18	0	3	82	40
19	0	10	194	91
20	0	4	45	20
21	1	11	81	46
22	18	1	144	116
23	3	6	150	62
24	4	11	282	116
25	7	20	345	287
26	8	19	377	215
27	1	6	100	84
28	1	15	134	62
29	3	5	77	21
30	2	3	75	33

Table 14. Quarter 1 breakdown of number of Tweets based on sentiment strength by game. Number of Tweets broken out by polarity and strength posted during Quarter 1 for each game.

ID	q1 strong	q1 strong positive	q1 strong negative	q1 weak	q1 weak positive	q1 weak negative	q1 neutral	q1 equal
1	7	2	5	99	70	29	85	5
2	2	0	2	48	28	20	50	9
3	8	1	7	99	73	26	95	5
4	3	1	2	40	27	13	45	2
5	0	0	0	41	19	22	29	3
6	1	1	0	48	33	15	48	3
7	0	0	0	24	20	4	25	0
8	1	0	1	25	21	4	25	0
9	1	0	1	25	21	4	25	0
10	4	1	3	167	131	36	147	17
11	0	0	0	26	15	11	21	4
12	1	0	1	54	27	27	36	3
13	1	1	0	58	34	24	32	5
14	2	1	1	38	25	13	52	4
15	0	0	0	61	39	22	34	6
16	0	0	0	21	14	7	20	0
17	4	0	4	66	37	29	71	7
18	2	0	2	31	20	11	39	2
19	1	0	1	59	47	12	33	5
20	0	0	0	14	12	2	16	0
21	2	0	2	32	23	9	24	1
22	6	0	6	74	52	22	62	4
23	2	0	2	42	22	20	49	1
24	3	1	2	95	75	20	70	8
25	3	1	2	91	68	23	93	7
26	2	0	2	47	28	19	47	2
27	2	1	1	36	19	17	36	3
28	4	0	4	70	57	13	47	5
29	2	1	1	12	8	4	23	3
30	1	1	0	23	14	9	25	1

Table 15. Quarter 2 breakdown of number of Tweets based on sentiment strength by game. Number of Tweets broken out by polarity and strength posted during Quarter 2 for each game.

ID	q2 strong	q2 strong positive	q2 strong negative	q2 weak	q2 weak positive	q2 weak negative	q2 neutral	q2 equal
1	3	2	1	68	42	26	79	1
2	4	0	4	5	41	24	54	2
3	9	3	6	110	85	25	85	5
4	1	0	1	46	24	22	46	6
5	0	0	0	31	18	13	37	4
6	1	1	0	29	16	13	31	3
7	1	0	1	34	28	6	42	3
8	1	1	0	13	12	1	10	0
9	1	1	0	16	12	4	7	0
10	9	2	7	138	99	39	148	18
11	2	1	1	17	13	4	12	1
12	6	0	6	52	28	24	32	3
13	1	0	1	48	22	26	29	2
14	4	2	2	32	24	8	35	0
15	5	1	4	103	74	29	48	13
16	6	1	5	20	11	9	14	3
17	2	1	1	67	39	28	65	6
18	1	0	1	23	15	8	25	2
19	0	0	0	106	69	37	53	8
20	1	0	1	5	2	3	12	1
21	4	0	4	33	19	14	21	3
22	6	1	5	77	32	45	80	10
23	1	1	0	56	42	14	56	4
24	3	2	1	85	62	23	86	9
25	6	2	4	115	61	55	79	8
26	1	0	1	49	32	17	48	7
27	4	0	4	49	28	21	37	2
28	7	0	7	56	37	19	47	7
29	2	0	2	21	14	7	13	4
30	1	1	0	14	14	0	15	1

Table 16. Quarter 3 breakdown of number of Tweets based on sentiment strength by game. Number of Tweets broken out by polarity and strength posted during Quarter 3 for each game.

ID	q3 strong	q3 strong positive	q3 strong negative	q3 weak	q3 weak positive	q3 weak negative	q3 neutral	q3 equal
1	2	0	2	53	42	11	47	2
2	3	0	3	30	20	10	46	9
3	1	0	1	71	45	26	71	4
4	4	1	3	81	57	24	55	6
5	0	0	0	34	17	17	29	2
6	2	0	2	44	22	22	39	1
7	2	1	1	30	26	4	32	3
8	1	1	0	13	9	4	12	5
9	1	1	0	13	9	4	12	5
10	5	3	2	91	61	30	98	15
11	1	0	1	6	3	3	8	1
12	5	2	3	31	20	11	19	2
13	4	0	4	61	31	30	37	11
14	1	0	1	19	12	7	16	2
15	0	0	0	60	43	17	42	8
16	4	1	3	43	20	23	30	6
17	5	0	5	86	46	40	82	13
18	0	0	0	28	16	12	29	1
19	3	0	3	50	33	17	35	2
20	2	0	2	20	7	13	12	9
21	4	0	4	26	15	11	23	7
22	3	0	3	70	37	33	69	8
23	4	2	2	70	57	13	62	5
24	1	0	1	96	69	27	95	9
25	9	0	9	102	57	45	81	13
26	1	1	0	66	34	32	52	4
27	0	0	0	49	29	20	27	2
28	1	0	1	41	25	16	30	4
29	2	1	1	25	18	7	27	1
30	0	0	0	8	4	4	11	1

Table 17. Quarter 4 breakdown of number of Tweets based on sentiment strength by game. Number of Tweets broken out by polarity and strength posted during Quarter 4 for each game.

ID	q4 strong	q4 strong positive	q4 strong negative	q4 weak	q4 weak positive	q4 weak negative	q4 neutral	q4 equal
1	8	1	7	84	45	39	89	9
2	4	1	3	64	44	20	47	5
3	10	6	4	155	123	32	104	13
4	4	0	4	62	50	12	51	5
5	2	0	2	47	25	22	39	6
6	2	0	2	32	21	11	26	2
7	2	0	2	76	55	21	67	4
8	1	1	0	46	37	9	18	4
9	1	1	0	46	37	9	18	4
10	6	2	4	116	90	26	78	6
11	0	0	0	10	8	2	10	4
12	1	0	1	30	14	16	14	4
13	6	4	2	108	76	32	64	9
14	5	0	5	68	55	13	40	5
15	7	1	6	173	124	49	86	12
16	7	0	7	41	25	16	37	4
17	12	2	10	98	61	37	72	9
18	0	0	0	40	31	9	46	4
19	6	0	6	70	45	25	56	7
20	1	0	1	26	24	2	20	6
21	2	1	1	36	24	12	36	4
22	4	0	4	39	23	16	31	5
23	2	0	2	43	28	15	48	3
24	8	1	7	122	76	46	93	7
25	9	4	5	323	159	164	162	34
26	5	2	3	143	90	53	78	11
27	1	0	1	50	24	26	36	6
28	4	1	3	29	15	14	32	6
29	2	1	1	30	37	3	12	3
30	3	0	3	63	43	29	37	5

Table 18. Overtime breakdown of the number of Tweets based on sentiment strength by game. Note that overtime and double overtime were combined in game 26

ID	OT strong	OT strong positive	OT strong negative	OT weak	OT weak positive	OT weak negative	OT neutral	OT equal
14	1	1	0	48	30	18	34	7
26	18	5	13	287	193	94	206	27

Table 19. Added average scores for all Tweets and for positive and negative Tweets posted during Quarter 1.

ID	Q1 # Tweets	Q1 Average	Q1 positive average	Q1 negative average
1	196	0.173469388	1.474489796	-1.301020408
2	105	-0.019047619	1.361904762	-1.380952381
3	207	0.169082126	1.483091787	-1.314009662
4	90	0.111111111	1.4	-1.288888889
5	73	-0.082191781	1.342465753	-1.424657534
6	100	0.18	1.43	-1.25
7	49	0.346938776	1.489795918	-1.142857143
8	119	0.285714286	1.571428571	-1.285714286
9	51	0.31372549	1.470588235	-1.156862745
10	335	0.292537313	1.534328358	-1.241791045
11	51	0.039215686	1.411764706	-1.37254902
12	76	0.236842105	1.473684211	-1.236842105
13	96	0.098904	1.494737	-1.39583
14	96	0.11458333	1.375	-1.2604167
15	101	0.178217822	1.514851485	-1.336633663
16	41	0.14634146	1.3902439	-1.2439024
17	148	-0.0540541	1.41216216	-1.4662162
18	74	-0.013513514	1.337837838	-1.351351351
19	98	0.37755102	1.653061224	-1.275510204
20	30	0.366666667	1.433333333	-1.066666667
21	59	0.203389831	1.508474576	-1.305084746
22	146	0.089041096	1.493150685	-1.404109589
23	94	-0.085106383	1.29787234	-1.382978723
24	176	0.329545455	1.573863636	-1.244318182
25	194	0.206185567	1.458762887	-1.25257732
26	98	0	1.367346939	-1.367346939
27	77	-0.064935065	1.376623377	-1.441558442
28	126	0.380952381	1.658730159	-1.277777778
29	40	0.125	1.45	-1.325
30	50	0.08	1.38	-1.3

Table 20. Added average scores for all Tweets and for positive and negative Tweets posted during Quarter 2.

ID	Q2 Total	Q2 average	Q2 positive average	Q2 negative average
1	151	0.125827815	1.397350993	-1.271523179
2	125	0.048	1.448	-1.4
3	209	0.277511962	1.578947368	-1.301435407
4	99	-0.070707071	1.383838384	-1.454545455
5	72	0.027777778	1.361111111	-1.333333333
6	64	0.015625	1.453125	-1.4375
7	80	0.275	1.45	-1.175
8	103	0	1.349514563	-1.349514563
9	24	0.625	1.833333333	-1.208333333
10	313	0.182108626	1.485623003	-1.303514377
11	32	0.25	1.625	-1.375
12	93	-0.139784946	1.483870968	-1.623655914
13	80	-0.14509	1.392405	-1.5375
14	71	0.23943662	1.4929775	-1.253211
15	169	0.326732673	1.633663366	-1.306930693
16	43	-0.1627907	1.51162791	-1.6744186
17	140	0.1	1.47142857	-1.3714286
18	51	0.137254902	1.450980392	-1.31372549
19	167	0.215568862	1.550898204	-1.335329341
20	19	-0.210526316	1.157894737	-1.368421053
21	61	-0.163934426	1.491803279	-1.655737705
22	173	-0.213872832	1.329479769	-1.543352601
23	117	0.247863248	1.478632479	-1.230769231
24	183	0.229508197	1.486338798	-1.256830601
25	209	0.038277512	1.492822967	-1.454545455
26	105	0.085714286	1.447619048	-1.361904762
27	92	-0.141304348	1.369565217	-1.510869565
28	117	0.008547009	1.521367521	-1.512820513
29	40	0.075	1.625	-1.55
30	31	0.548387097	1.709677419	-1.161290323

Table 21. Added average scores for all Tweets and for positive and negative Tweets posted during Quarter 3.

ID	Q3 total	Q3 average	Q3 positive average	Q3 negative average
1	104	0.259615385	1.538461538	-1.278846154
2	88	0.056818182	1.420454545	-1.363636364
3	147	0.136054422	1.428571429	-1.292517007
4	146	0.171232877	1.54109589	-1.369863014
5	65	-0.046153846	1.338461538	-1.384615385
6	86	-0.069767442	1.38372093	-1.453488372
7	67	0.343283582	1.567164179	-1.223880597
8	117	0.222222222	1.564102564	-1.341880342
9	31	0.322580645	1.838709677	-1.516129032
10	209	0.167464115	1.488038278	-1.320574163
11	16	-0.1875	1.25	-1.4375
12	57	0.157894737	1.631578947	-1.473684211
13	113	-0.16964	1.428571	-1.59821
14	38	-0.0263158	1.42105263	-1.4473684
15	105	0.295238095	1.561904762	-1.266666667
16	83	-0.2168675	1.46987952	-1.686747
17	186	-0.0645161	1.42473118	-1.4892473
18	58	0.034482759	1.344827586	-1.310344828
19	90	0.111111111	1.511111111	-1.4
20	37	-0.486486486	1.297297297	-1.783783784
21	60	-0.116666667	1.516666667	-1.633333333
22	150	-0.06	1.393333333	-1.453333333
23	141	0.382978723	1.617021277	-1.234042553
24	201	0.199004975	1.437810945	-1.23880597
25	205	-0.092682927	1.424390244	-1.517073171
26	123	-0.016260163	1.406504065	-1.422764228
27	78	0.076923077	1.5	-1.423076923
28	76	0.118421053	1.486842105	-1.368421053
29	55	0.290909091	1.527272727	-1.236363636
30	20	-0.15	1.3	-1.45

Table 22. Added average scores for all Tweets and for positive and negative Tweets posted during Quarter 4.

ID	Q4 total	Q4 average	Q4 positive average	Q4 negative average
1	190	-0.078947368	1.363157895	-1.442105263
2	120	0.133333333	1.516666667	-1.383333333
3	282	0.411347518	1.684397163	-1.273049645
4	122	0.229508197	1.549180328	-1.319672131
5	94	-0.127659574	1.393617021	-1.521276596
6	62	0	1.435483871	-1.435483871
7	149	0.201342282	1.469798658	-1.268456376
8	331	0.163141994	1.561933535	-1.398791541
9	69	0.492753623	1.753623188	-1.260869565
10	206	0.325242718	1.597087379	-1.27184466
11	24	0.208333333	1.541666667	-1.333333333
12	49	-0.224489796	1.469387755	-1.693877551
13	187	0.283422	1.620321	-1.3369
14	118	0.18421053	1.63157895	-1.4473684
15	278	0.223021583	1.633093525	-1.410071942
16	89	-0.1348315	1.4382025	-1.5730337
17	191	-0.0209424	1.46073298	-1.4816754
18	90	0.266666667	1.444444444	-1.177777778
19	139	0.028776978	1.482014388	-1.45323741
20	48	0.5	1.625	-1.125
21	78	0.128205128	1.435897436	-1.307692308
22	79	-0.075949367	1.417721519	-1.493670886
23	97	0.134020619	1.443298969	-1.309278351
24	230	0.065217391	1.482608696	-1.417391304
25	528	-0.020833333	1.482954545	-1.503787879
26	237	0.105485232	1.556962025	-1.451476793
27	93	-0.139784946	1.397849462	-1.537634409
28	71	-0.084507042	1.450704225	-1.535211268
29	57	0.736842105	1.894736842	-1.157894737
30	108	0.111111111	1.555555556	-1.444444444

Table 23. Added average scores for all Tweets and for positive and negative Tweets posted during overtime (OT) and double overtime (DOT).

ID	OT Total	OT average	OT positive average	OT negative average	DOT total	DOT average	DOT positive average	DOT negative average
14	90	0.177	1.5222	-1.344				
26	275	0.2363	1.5527	-1.3163	263	0.0342	1.5019	-1.4676

References

Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008, February). Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, 183-194.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R. (2011, June). Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, Portland, Oregon, pp. 30–8. Stroudsburg, PA: Association for Computational Linguistics.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *Proceedings of Language Resources and Evaluation Conference.* 10, 2200-2204.

Baumeister, R., Bratslavsky, E., Finkenauer, C. & Vohs, K.(2001). Bad is stronger than good. *Review of General Psychology,* 5(4), 323-370.

Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. *Proceedings of the 45th annual meeting of the Association for Computational Linguistics (ACL)*. 440- 447.

Byrne, G. (2007). A statistical primer: Understanding descriptive and inferential statistics. *Evidence Based Library and Information Practice, 2*(1), 32-47.

Carvalho, P., Sarmento, L., Silva, M. J., & de Oliveira, E. (2009, November). Clues for detecting irony in user-generated contents: oh...!! it's so easy;-). In *Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion,* 53-56.

Esuli, A., & Sebastiani, F. (2006, May). Sentiwordnet: A publicly available lexical resource for opinion mining. *Proceedings of Language Resources Evaluation Conference,* 6, 417-422.

Hu, M. & Liu, B. (2004). Mining and summarizing consumer reviews. *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining,* 168-177.

Iliyecheva, Maria. (2005). *"Faithful until Death": Sports fans and national discourse in Bulgarian internet forums.* Polish Sociological Review 3(151), 251-270.

Lee, L. (n.d.). Lillian Lee: Research Summary. Retrieved November 19, 2014, from http://www.cs.cornell.edu/home/llee/research.html.

LIWC: Linguistic Inquiry and Word Count. (n.d). Retrieved March 15, 2015, from http://www.liwc.net/howliwcworks.php.

Maynard, D., Bontcheva, K., & Rout, D. (2012). Challenges in developing opinion mining tools for social media. *Proceedings of @NLP can u tag #usergeneratedcontent?!,* 15-22.

Moutinho, L. (2011). Correlation analysis. In L. Moutinho, & G. Hutcheson (Eds.), *The SAGE dictionary of quantitative management research.* (pp. 57-61). London: SAGE Publications Ltd. doi: http://dx.doi.org/10.4135/9781446251119.n17

OpinionFinder | MPQA. (n.d.). Retrieved November 16, 2014, from http://mpqa.cs.pitt.edu/opinionfinder/.

Paltoglou, G., & Thelwall, M. (2012). Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *3*(4), 66.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), 1-135.

Pang, B. (n.d.). Bo.pang. Retrieved November 19, 2014, from https://sites.google.com/site/bopang42/.

Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of Language Resources Evaluation Conference,* 10, 1320-1326.

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. *ACLStudent '05 Proceedings of the ACL Student Research Workshop*. 43-48.

SentiStrength - sentiment strength detection in short texts - sentiment analysis, opinion mining (n.d.). Retrieved December 3, 2014 from http://sentistrength.wlv.ac.uk/.

SentiWordNet. (n.d.). Retrieved November 16, 2014, from http://sentiwordnet.isti.cnr.it/.

Thelwall, M., & Buckley, K. (2013). Topic-based sentiment analysis for the social web: The role of mood and issue-related words. *Journal of the American Society for Information Science and Technology, 64*(8), 1608-1617.

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology,* 62(2): 406–18.

Thelwall, M., Buckely, K., Paltoglou, G., Cai, D. & Kappas, A. (2010, December) *Sentiment strength detection in short informal text.* Journal of the American Society for Information Science and Technology, 61(12), 2544-2558.

Tumasjan, A., Sprenger, T., Sandner, P. & Welpe, I. (2010) Predicting elections with Twitter: What 140 characters reveal about political sentiment. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media,* 178-185.

Vaughan, L. (2001). Examining relationships for interval and ratio data -- Correlation and regression. In *Statistical Methods for the Information Professional.* Medford, NJ: Information Today, 93-110.

White, M.D., & Marsh, E.E. (2006). Content analysis: A flexible methodology. *Library Trends, 55*(1), 22-45.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., & Patwardhan, S. (2005, October). OpinionFinder: A system for subjectivity analysis. Proceedings of hlt/emnlp on Interactive Demonstrations, 34-35.