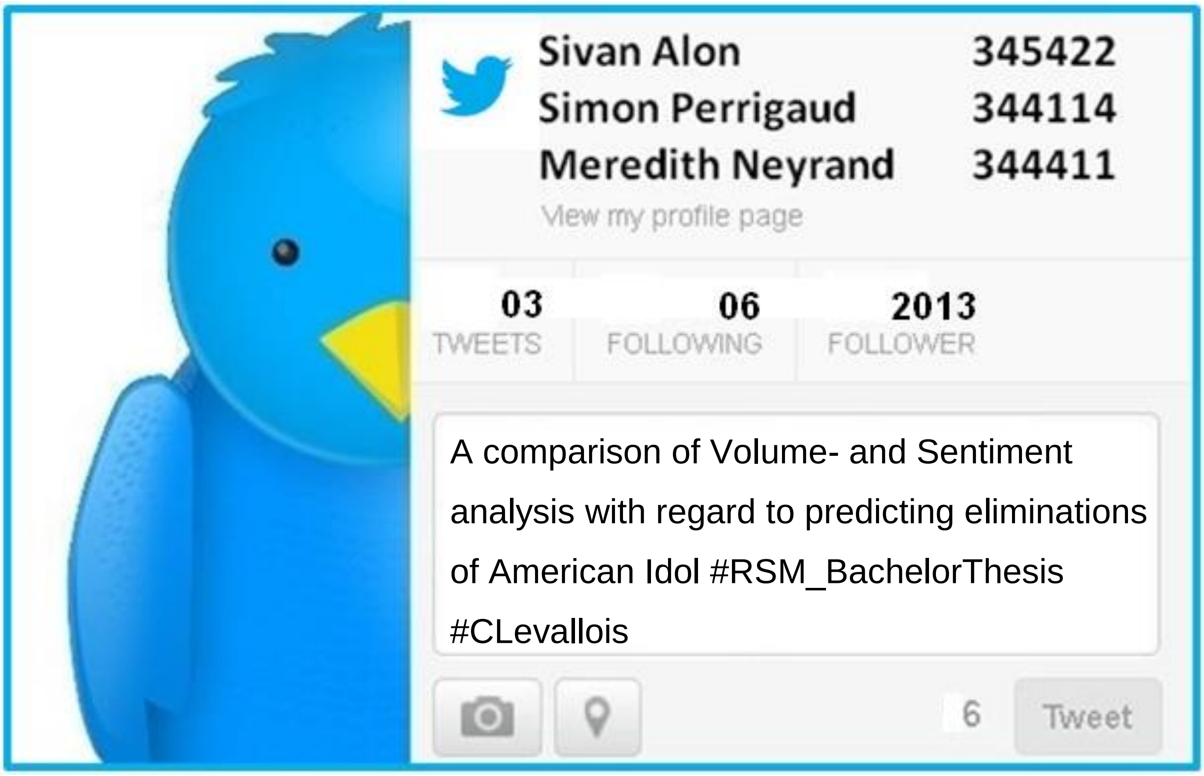
Predicting American Idol with Twitter Sentiment



Sivan Alon	345422
Simon Perrigaud	344114
Meredith Neyrand	344411

A comparison of Volume- and Sentiment analysis with regard to predicting eliminations of American Idol #RSM_BachelorThesis #CLevallois

This document is written by Sivan Alon, Simon Perrigaud, and Meredith Neyrand, who declare that each individual takes responsibility for the full contents of the whole document. We declare that the text and the work presented in this document is original and that no sources other than mentioned in the text and its references have been used in creating it. RSM is only responsible for supervision of completion of the work but not for the contents.

Summary Form

Name instructor	**Clément Levallois**
Team number	3
Name student 1	Sivan Alon
Name student 2	Simon Perrigaud
Name student 3	Meredith Neyrand
Hypothesis as reformulated by you	A greater number of positive tweets about a contestant of American Idol is more likely to be associated with a greater number of votes for that contestant
Theoretical domain	Elections
Population that was studied	US singing contest show
Case that was studied	American Idol 2013
Focal unit	Individual contestants
Independent variable	The relative share of votes received (Semi-ordinal ranking)
Dependent variable	The number of positive tweets about a contestant
Type of relation (Causal / Not causal)	There is no causal relationship. This hypothesis is based on an association between the number and nature of tweets pertaining to a contestant on the one hand (X) and the number of votes that contestant subsequently receives (Y).
Minimum effect size for having practical relevance	To have a point of reference, traditional polls included in the survey report MAEs varying from 0.4% to 1.81%. Since they are currently the best prediction tool available to forecast election results they set the golden standard
Outcome of the quantitative meta-analysis	Volume based meta-analysis MAE: 11.65% Sentiment based meta-analysis MAE: 9.53%
Your research strategy	Cross-sectional
Effect size parameter	N/A
Observed effect size	N/A
Your study's contribution to what is known about the hypothesis	Sentiment improves prediction accuracy when compared to volume based approach.
Most important recommendation for further research	One of the central issues further research should focus on is to correct for the demographic bias present within the twitter user base.

Abstract:

In an attempt to add to the body of research on the topic of electoral predictions with Twitter we have tried to predict the eliminations of contestants from American Idol 2013. To this end, we have conducted a literature review to uncover the trends in the existing literature and to isolate the factors that might improve our methodology. Building on the study by Ciulla et al. (2012) we have extracted over 40,000 tweets and constructed one prediction model based primarily on tweet volume and a second model focusing on sentiment. In line with our hypothesis, we found that sentiment improved the prediction accuracy. Furthermore, we found that the overall accuracy of our predictions were low when compared to the findings of Ciulla et al. As a result, we have outlined the possible reasons and limitations of our study and suggest four main points of improvement for further research.

Table of content

1. Introduction

Ever since its creation the Micro-blogging platform 'Twitter' has been used as a tool to predict a variety of outside variables. User generated content (in the form of 'tweets') has been analysed to predict outcomes ranging from the box-office of movies (Asur & Huberman 2010, Krauss et al. 2008), through the stock-prices of companies (Bollen et al. 2010), all the way to the outcome of presidential elections (O'Connor et al 2010).

Research on predicting the outcome of elections is based on the assumption that popularity and voting intention translates into the volume and sentiment of tweets (analysis of the tweet-content) about a party or contestant. Building on the research carried out by Ciulla et al (2012) we have extracted data from Twitter to predict the percentage of votes candidates on the popular TV show American Idol will obtain as well as their ranking (and elimination). In addition to using the tweet volume we have also tried to improve the model's accuracy by manually annotating the tweets' sentiment. As such, the results of this study can contribute to the body of literature on the subject of electoral predictions.

We believe that by including sentiment in our prediction we can improve the overall accuracy of the model. The underlying assumption of this research is that a positive sentiment strength will reflect positively on a contestant's relative number of votes and ranking.

1.1 Twitter

Twitter is a social networking service that allows its users to post content in the form of 'tweets'. These tweets are short messages containing up to and including 140 characters. Recognisable by its light-blue bird-logo and its users' tendency to classify their own content by use of hash tags, the micro-blogging platform has grown to attract the attention of consumers as well as businesses (Gallaugher 2012). Created in 2006 in San Francisco, Twitter is now available in over 20 languages (Twitter 2013) and counts over 250 million active users (McCune 2013). Increasingly, researchers have attempted to analyse and make use of the vast amounts of digital content that is continuously being created through Twitter. This has led to people claiming that an analysis of this content could be used as a tool for predictive purposes.

1.2 American Idol

Created by Simon Fuller in 2002, American Idol welcomes four popular judges to select, mentor, and guide aspiring young artists. The show can best be described as a reality-singing competition made up of several stages.

Eleven years after the airing of the first season, American Idol began its twelfth season in January of 2013. After the initial selection of candidates, the judges eliminated contestants every week until reaching the 'semi-final' stage in which only 24 to 36 contestants remained. From this stage onward, the show's viewers decided on the fate of the singers by voting for their favourite one (online, by telephone, or through SMS). Every week, the bottom 2 candidates were eliminated until only 10 remained. At this point, the final stage was introduced and 1 contestant was eliminated each week. In addition to the elimination, the show would at times also present the bottom two and/or top two contestants (the two contestants who received the lowest and highest number of votes respectively).

There were two shows per week. The first one took place on Wednesday and was dedicated to the singers' performances. At the end of the show, viewers benefited from a 2 hour window to cast their vote (Online: 50 votes per time window per user; Telephone & Text: Unlimited). The results were then announced the next day during the second show.

1.3 Predicting American Idol with Twitter

With our study we aimed to replicate research conducted by Ciulla et al. (2012). Consequently, instead of trying to predict the outcome of political elections, our research domain was more generally elections and our focal unit the individual contestants of American Idol 2013 (population: singing competition in the US).

1.4 Relevance

Although this study's focus is on American Idol, it has implications for the wider field of 'making predictions with Twitter' and more particularly 'predictions about elections'. The hypothesis being that a greater tweet volume and positive sentiment strength about a contestant of American Idol is more likely to be associated with a greater number of votes for that contestant. In this context American Idol constitutes a "well defined electoral phenomenon that each week draws millions of votes in the USA" (Ciulla et al. 2012; p.1). As a consequence, American Idol constitutes a stepping stone in making successful predictions about political elections. Ultimately, if successful, the use of Twitter might complement or even substitute traditional polls, thereby significantly decreasing costs (Tumasjan et al. 2010).

2. Literature Review

Within the body of research regarding Twitter there has been a growing number of studies pertaining to the use of information gained from Twitter to predict elections. Previous studies have attempted to predict the outcomes of eleven elections, in various countries (USA, Germany, Ireland, The Netherlands and Singapore) at different levels (presidential, federal, and state). The general underlying assumption is that more (less) popular candidates (or parties) attract more (less) attention on twitter. Thus, it seems intuitive that the actual percentage of votes candidates (parties) receive is related to twitter chatter during the election period.

Two main methods have been employed to infer votes from tweets - volume analysis and sentiment analysis. The former involves simply counting the number of tweets mentioning a particular candidate or party. Sentiment analysis, the subject of this study, attempts to infer the voting intentions from tweets.

The results of these studies, however, have remained largely inconclusive, raising the question, as to what extent Twitter can actually predict elections.

In the following literature review we will describe the evolution and quality of the most relevant research that has been conducted on the subject of predicting elections with Twitter.

2.1 Survey of the literature

The first publications to underscore the potential of using Twitter data as a proxy for measuring political opinion is the work conducted by O'Connor et al. (2010). In this study the authors examine the temporal correlation between various indices of public opinion (e.g. Gallup index) and twitter volume and sentiment score (ratio of positive to negative tweets).

With regard to electoral polls the authors found a correlation of 79% between the number of mentions of ‘Barack’ and the results of the presidential election polls in 2008. However, the volume of tweets associated with J. McCain also correlated highly with those of Obama (74%), rather than with his own tweeter ratings (no measure was reported). O’Connor et al. conclude that “topic frequency may not have a straightforward relationship to public opinion in a more general text-driven methodology for public opinion measurement” (2010, p. 7), but encourage further exploration of the subject.

Another pioneering work was published in the same year by Tumasjan et al. (2010). The authors were the first to directly examine the predictive power of twitter with regard to election results and suggest the use of mean absolute error (MAE) as a measure of performance. They claim that they were able to accurately predict the results of the 2009 German Federal election, using the mere tweet volume mentioning one of six political parties. They report a mean difference of 1.65% - a competitive performance when viewed against traditional polls (0.8%-1.48%).
This study was highly influential and is responsible for much of the hype surrounding the subject of using Twitter for electoral predictions. At the same time, it triggered a series of articles claiming the results were not based on a well-grounded methodology and thus could not be replicated in future studies. (Avello, 2011; Metaxas et al.,2011; Jungherr et al. 2011).

Jungherr et al. (2011) wrote a paper in response to Tumasjan et al. (2010) pointing out that the results of the former are highly contingent on the arbitrary methodological choices taken by the authors. They demonstrate that by changing the data collection period the model’s accuracy fluctuates between 1.51% and 3.34%. Even more striking, merely adding one more party to the subset included in the study increases the MAE to 9.3%.

Avello (2011) extended the analysis methods used by O’Connor et al. (2010) and Tumasjan et al. (2010) to tweets collected two months before the US 2008 presidential elections. By configuring Twitter’s search API to include a user’s geographical location, he was able to correctly predict the election results for 6 different states. With the exception of California (MAE = 0.42%), the prediction results largely overestimated voter support for Obama, with the mean difference varying from 7.49% to 20.34%. His findings make evident that bias in Twitter’s user base permeates research and highlights the need to correct the results according to demographics and voting behavior.

Metaxas et al. (2011) also analyzed a number of different state elections by using both tweet volume as well as sentiment analysis. The authors found that sentiment improved the prediction accuracy, producing a lower MAE than an analysis based solely on volume. Although the results for both methods were inconsistent and at times very high (MAEs ranging from 1.2% to 39.6%), this may in part be explained by Metaxas et al.’s choice of analyzing highly contested elections. Furthermore, two important conclusions can be drawn from this study. First, the author’s found that the sentiment analysis method employed by O’Connor et al. (2010) is only slightly better than chance. Second, the authors have outlined the importance of cleansing the data; denoising it by removing tweets that constitute spam, propaganda, rumours or other forms of disinformation.

More recent studies have focused less on developing criticism of current methodologies and have instead attempted to take into consideration the issues raised by previous authors. As a consequence, Bermingham & Smeaton (2011) have tried to improve the overall accuracy of their model by trying different measures of sentiment and time windows. These attempts have allowed the authors to improve their MAE from 5.51% to 3.67%. Similarly, Tjong et al. (2012) attempted to incorporate the advice of Metaxas et al. and tried to debias their data by the use of polls achieving an MAE of 1.33% and 2.00% for volume and sentiment respectively.

2.2 Meta-Analysis

2.2.1 Effect Size

As is apparent by the literature survey, the most commonly used effect size to evaluate the performance of the prediction is MAE. It measures the extent to which "a set of predicted values deviate from the actual values." (Bermingham, 2011:4). It was first used by Tumasjan et al. (2010), setting the standard for all subsequent studies. To have a point of reference, traditional polls included in the survey report MAEs varying from 0.4% to 1.81%. Since they are currently the best prediction tool available to forecast election results they set the golden standard.

The literature survey shows that Twitter prediction models' MAEs vary considerably from 1.65% to 17.1%. In order to assess the current status of the effect size in the body of literature on the subject, we have conducted a meta-analysis. Through the examination of different studies' methodological procedures we hope to uncover the sources of variations in these results.

We decided to include only studies reporting MAE to allow for comparison, therefore excluding O'Connor et al (2010) and Livne et al (2011), both of which report correlations. Finally, we have also excluded Jungherr et al (2011) as the MAE reported is unstable.

2.2.2 Meta-analysis Process

MAE is a measure of dispersion and as such does not permit meta-analytical manipulation as would be the case with common effect sizes (e.g. mean or correlation). As a consequence, in order to derive an overall value of the effect size we simply used an average of the MAE's reported. As the field of electoral predictions with Twitter is still in its early stages there is no uniform procedure with regard to the methodology employed. However, studies can generally be classified as paying attention to either volume or sentiment. For this reason we have decided to conduct a meta-analysis on the basis of volume and sentiment. We applied the model (simple average) once for each prediction method and computed two meta-analytical measures of MAE (table 2.1 and 2.2). However, we would like to note that methodological differences between studies force us to also analyse each on a case by case basis in order to deepen our understanding of its results.

2.2.3 Comparison of prediction methods

As can be seen in tables 2.1 and 2.2, the meta-analysis yields an MAE of 11.65% for the volume analysis and 9.53% for the sentiment analysis which suggests that sentiment analysis does not substantially improve the results. However, we believe that individual studies that have used *both* prediction methods provide a better setting for comparison. As presented in figure 2.1, sentiment-based analysis outperforms volume-based analysis for both Metaxas et al. (2011) and Bermingham et al. (2011). These results provide preliminary support for our hypothesis.

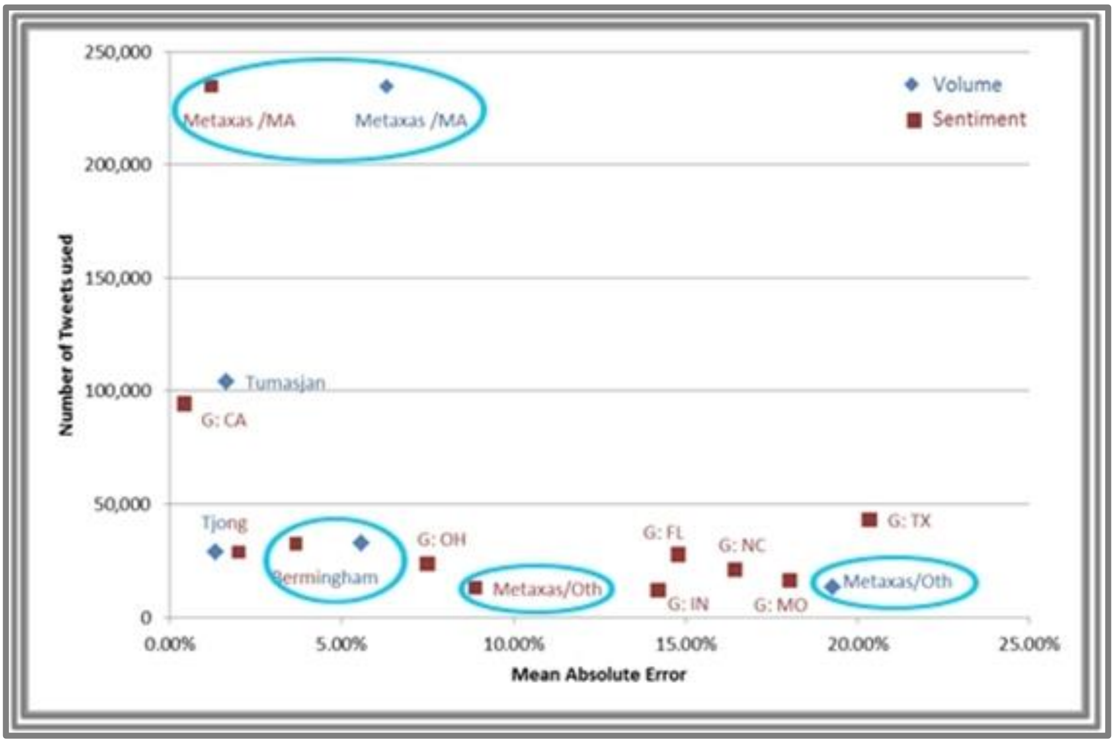


Figure 2.1: Mean Absolute Error

*Gayo Avello (2011) analysed different states and therefore is presented as G:State

Table 2.1: Volume-based Analysis

Author	Election	MAE
Tumasjan et al. (2010)	German federal election, 2009	1.65%
Metaxas et al. (2011)	US senate elections in MA, 2010	6.30%
	US senate elections in CO, 2010	24.60%
	US senate elections in NV, 2010	1.90%
	US senate elections in CA, 2010	3.80%
	US senate elections in KY, 2010	39.60%
	US senate elections in DE, 2010	26.50%
Bermingham et al. (2011)	Irish General Election, 2011	5.58%
Tjong et al. (2012)	Dutch senate election, 2011	1.33%
Skortic et al. (2012)	Singaporean election, 2011	5.23%
Overall MAE	Result	**11.65%**

Table 2.2: Sentiment-based Analysis

Author	Election	MAE
Gayo-Avello (2011)	US 2008 Presidential Election California	0.42%
	US 2008 Presidential Election Florida	14.78%
	US 2008 Presidential Election Indiana	14.20%
	US 2008 Presidential Election Missouri	18.03%
	US 2008 Presidential Election N. Carolina	16.44%
	US 2008 Presidential Election Ohio	7.49%
	US 2008 Presidential Election Texas	20.34%
Metaxas et al. (2011)	US senate elections in MA, 2010	1.20%
	US senate elections in CO, 2010	12.40%
	US senate elections in NV, 2010	4.70%
	US senate elections in CA, 2010	6.30%
	US senate elections in KY, 2010	1.20%
	US senate elections in DE, 2010	19.80%
Bermingham et al. (2011)	Irish General Election, 2011	3.67%
Tjong et al. (2012)	Dutch senate election, 2011	2.00%
Overall MAE	Result	**9.53%**

2.2.4 Systematic Overview of Methodology

In this section, we present other methodological characteristics that must be compared on a case by case basis.

Period of data collection

Table 2.3 reveals that the time frame used for analysis varies considerably between the studies. Both Jungherr et al. (2011) and Bermingham et al. (2011) have addressed this issue empirically by calculating the MAE for different time frames. Jungherr et al. (2011) demonstrated that the model accuracy varies from 1.51% to 3.34%. Similarly, Bermingham et al. (2011) found performance to fluctuate between 5.58% and 9.2%. This implies that the time period used is an important parameter. However, in both cases the choice of the time frame was not based on well-grounded rules. More specifically, the start of the data collection had no particular significance in the campaign (e.g. an important debate symbolizing the beginning of the campaign). Therefore, no further conclusions with regard to our own methodology can be drawn.

Denoising

Denoising implies the removal of tweets that constitute spam, propaganda, rumours or other forms of disinformation. Although various authors have stressed the importance of denoising (Metaxas et al. 2011; Gayo-Avello 2012), as of yet, no empirical research has found evidence of the impact it has on the accuracy of the results. Due to the advantage of analysing the data manually we were able to subject our data to the process of denoising in the course of this study.

Authors	Collection Period	End of collection
O'Connor et al. 2010	15 days	election day
Tumasjan et al. 2010	38 days	one week before election day
Jungherr et al. 2011	2 different time period	election day
Gayo-Avello 2011	2 months	election day
Metaxas et al. 2011	7 days	election day
Bermingham & Smeaton 2011	18 days	election day
Tjong Kim Sang & Bos 2012	7 days	election day
Skoric et al. 2012	30 days	election day

Table 2.3: Collection period

Sentiment analysis method

Two common approaches for sentiment annotation used in the surveyed articles are machine learning and lexicon based methods (see appendix A). These techniques offer the advantage of rapid automated analysis of the data collected. However, if not carefully chosen to fit the data type (e.g. tweets as opposed to blogs) and the subject domain (e.g. political content as opposed to text about movies) they are likely to produce less accurate results (Thelwall et al., 2011; Bermingham and Smeaton 2010).

O'Connor et al. (2010), Metaxas et al. (2011) and Gayo-Avello (2011) used a subjectivity lexicon from OpinionFinder, "a word list containing about 1,600 and 1,200 words marked as positive and negative, respectively" (2010, p.3). Metaxas et al. (2011) manually analyzed a small subset of the data and found that "the accuracy of the sentiment analysis [was] only 36.85%, slightly better than a classifier randomly assigning the same three labels." This is not unexpected as OpinionFinder was not developed to analyze to twitter data.

One would expect that the use of a customized sentiment analysis would improve the accuracy of predictions. Indeed, by excluding Metaxas et al and Gayo-Avello (2011) the Overall MAE would be improved from the original MAE of 9,53% to 2,84%. The results should however be considered with caution as the resulting overall MAE was computed from only two studies.

Debiasing

Twitter's user base is not representative of the public, as is evident by a survey conducted by Pew internet (2012) revealing that "African-Americans, young adults, and mobile users stand out for their high rates of Twitter usage". This is a critical problem since dominating demographic groups may skew the results towards a particular candidate. Debiasing refers to the attempt of limiting or removing any form of demographic bias within the data Gayo-Avello (2012).

Review of the effect size reported indicates that attempts to debias the data improved the accuracy of the prediction. Specifically, Gayo-Avello (2011) states that the attempt to reduce demographic bias by weighing tweets (for which he was able to obtain the users' age) according to age participation for the previous 2004 election enabled him to reduce the error from 13.10% to 11.61%. Similarly, Tjong et al. (2012) achieved higher accuracy by debiasing the data using pre-election polling data.

A comprehensive summary of the results of our meta-analysis can be found in Appendix B.

3. Hypothesis

Our study tests the following hypothesis: a large volume of positive tweets about a candidate is associated with a large number of votes.

First, we believe the population of Twitter users and American Idol watchers to overlap greatly. This overlap can be seen in American Idol's decision to include instant polls from audience members tweeting in the live show on screen (Stelter 2013). Furthermore, certain songs sung by the contestants were selected from submissions made by viewers on Twitter (Hernandez 2013). In the research conducted by Smith and Brenner of the PewResearchCenter, the authors found that among others African-Americans particularly "stand out for their high rates of Twitter usage" (2012). This bias is also reflected in the population of people watching American Idol; as adjusting for Ethnicity allows one to see that American Idol's Wednesday and Thursday shows are among the third and fourth most popular amongst African-Americans (compared to eight and below ten for the entirety of the US) (Nielsen 2013). Voting for American Idol is also completely voluntary. The lack of moral obligation would suggest that people taking the time to tweet about the show are more likely to also vote for the contestants.

Second, American Idol allowed us to try and predict several rounds of elimination, thereby increasing the reliability of our methodology.

Third, as pointed out by Gayo-Avello (2012) incumbency rates can play a significant role in the process of electoral predictions. By analysing American Idol we can disregard this factor as all contestants are new to the show.

However, there is a drawback with regard to the effect size. Our hypothesis was based on an association between the number and nature of tweets (sentiment) pertaining to a contestant on the one hand (X) and the number of votes that contestant subsequently received (Y). Unfortunately, American Idol did not publish the exact voting results and thus did not provide us with a dependent variable. Instead, we were forced to replicate Ciulla et al.'s (2012) methodology with regard to results, comparing our predictions to the semi-ordinal ranking published by AI (bottom two, top two). This limitation has made it impossible for us to compute an effect size, such as a Mean Absolute Error (MAE).

Ciulla et al use information gathered from Twitter to construct a rather simplistic model using the 'tweet volume' about contestants in 'American Idol' to rank them according to popularity. This model was then used in order to predict the contestants' eliminations.

Focusing on the top 10 rounds in the American Idol show, the authors first performed post-event analysis, using the volume of tweets referring to each of the contestants to rank them in descending order (ordinal data), the last contestant being most likely to be eliminated according to the model. They then used their model to predict the winner before the result was announced.

As a result of the research conducted by Ciulla et al. we constructed and tested two models:

The **first** one is a replication of the model, as devised by Ciulla et al. (2012), using only the volume of tweets about each candidate.

The **second** is an attempt to improve upon the model already provided by Ciulla et al. in order to determine whether the inclusion of further sentiment analysis renders the model more accurate. As negative and neutral tweets are less likely to result in voting intention we believed that by classifying tweets according to sentiment and only including positive tweets in our prediction would improve the model's accuracy. Furthermore, we also applied data cleansing, purifying and denoising the data.

4. Research Methodology

In this section we present the methodological steps of our study.

4.1 Data collection channel

Twitter allows for the import of tweets through the platform's application programming interface. Two channels provide an option to retrieve messages on a larger scale, 'Search API' and 'Streaming API'. Search API enables the collection of tweets matching a specific query from Twitter's global stream of tweets published during the previous 6-9 days by submitting HTTP requests to Twitter's server. However, Twitter limits the number of requests that can be sent per hour by a unique user. We have chosen to use the Twitter streaming API which enables a near real-time data import of filtered tweets by establishing a continuous connection to Twitter's servers. While it yields higher volumes of data, it still captures only about 1% of the entire traffic (higher access level requires a commercial partnership with Twitter).

As the sampling procedure is controlled by Twitter, there is no way to assess sampling bias. However there is some evidence that streaming API provides better sampling. More specifically, González-Bailón (2012) compared the differences between samples returned from Search and Streaming API and found that the former “over-represents the more central users and does not offer an accurate picture of peripheral activity” (p.1).

4.2 Data collection period

Our original plan was to follow Ciulla et al.’s methodology, according to which only tweets generated during the allowed voting window in the show, will be considered (from 8PM until 3AM EST). However, after a closer examination of the preliminary data it appears that about an hour before the start of the show (7 PM EST), the volume of tweets increased considerably, reaching a peak when the show airs at 8 PM EST (figure 4.1). During the voting window the number of tweets gradually declined, with a slight increase after the window had closed.

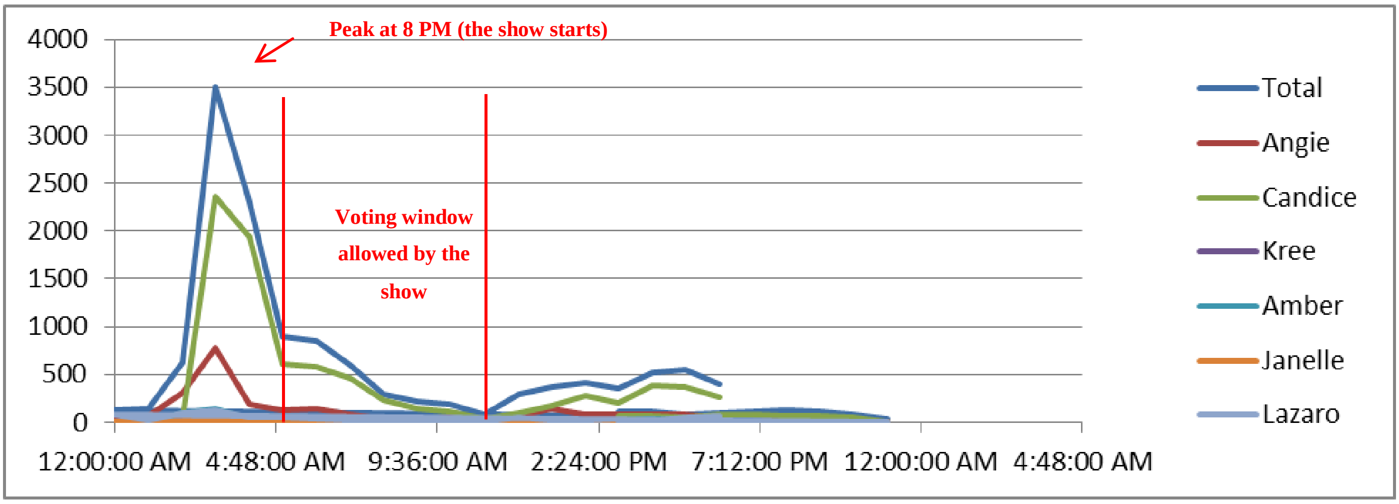


Figure 4.1: Tweet volume per hour

As a result of these findings we have decided to expand the time period for data analysis from 7PM until 3AM EST.

4.3 Data Processing and Tweet Classification

4.3.1 Data processing: Manual vs. Automatic

As was concluded in the literature survey, it appears that automatic tools’ precision depends on the context (political election vs. singing competition) and type of medium used (social media website vs. printed newspaper). We wanted to test the performance of an automated sentiment analysis tool and the extent to which its use affects the accuracy of the predictions. To this end, we relied on Umigon, a lexicon based sentiment analyzer developed specifically for Twitter by Levallois (2012), which classifies tweets as positive, negative or neutral.

First, we manually annotated 3000 tweets (each tweet was labeled by two annotators), then we measured the inter-annotators agreement using the S statistic (Bennett et al., 1954)

An inter-annotator agreement score of 0.89 was achieved, signaling ‘almost perfect agreement’ (see Appendix D for calculations and confusion matrix).

Afterwards, we compared our manual sentiment analysis with the results obtained from Umigon’s (table 4.1)

		Umigon			
		Negative	Neutral	Positive	Total
Manual	Negative	**40**	27	7	74
	Neutral	54	**229**	114	397
	Positive	207	695	**960**	1862
	Total	301	951	1081	2333

Table 4.1 : Confusion matrix

The performance of Umigon was evaluated along 3 widely accepted indicators adopted from Sokolova et al. (2009): accuracy, precision and recall (see Appendix E) Umigon appears to be relatively efficient at rating positive tweets (precision of 89% and a recall of 79%), however the precision and recall regarding negative tweets suggests that it is not reliable (13% and 33% respectively). Initial analysis of our first data set shows that about 80% of tweets are positive (see Appendix F). Furthermore it would appear that for certain contestant a large proportion of their total tweets are negative in nature. As a consequence, correctly classifying these tweets is of great importance to the accuracy of our predictions. Therefore, we decided not to make use of Umigon. Since the development of a sentiment analysis tool customized for our purposes is outside the scope of this thesis, we opted for manual annotation (thereby ensuring optimal classification of the data.)

4.3.2 Tweet Classification

After tweets had been arranged according to the contestant to whom they referred (by means of a formula in Excel), they were then subjected to manual analysis. In the context of the manual sentiment analysis and data cleansing process tweets could be classified as belonging to one of six categories. Categories one to three represented varying degrees of sentiment strength. While categories four through six consisted of different forms of tweets that had been removed in an attempt to denoise the data and to ensure its purity.

Category 1 – *Negative Sentiment:* This first category contains all the negative-sentiment tweets pertaining to a particular contestant. Tweets such as: "The world would be a much better place if Lazaro Arbos wasn't singing on national television once a week" would thus receive a rating of '1'.

Category 2 – *Neutral Sentiment:* A score of '2' is given to all tweets whose sentiment can neither be classified as positive nor negative (nor does the tweet have to removed - 4,5,6). Consequently, "I sware that Angie Miller and Miley Cyrus are the same person" would receive a score of '2'.

Category 3 – *Positive Sentiment:* Tweets are classified as belonging to category 3 when the underlying sentiment of the tweet is positive. This can take on the form of "If Candice Glover doesn't win American Idol... I quit and am moving to North Korea", as well as "Angie Miller is hot".

Category 4 – *Spam, Propaganda, or Advertising (Denoising)*: This category is made up of any tweet that seems to have been published for commercial purposes. Examples include tweets such as: "American Idol: Did Candice Glover Have the Best Performance in Show History?: The singer's version of "Loveson... http://t.co/uDNrgxsjrS", or "Candice Glover's rendition of "Love Song" is on iTunes now http://t.co/NCrFYkJzgg". Retweets of spam are not considered as spam since they show an intention of sharing information about a candidate (either positive, negative or neutral).

Category 5 – *Foreign Languages and Tweets from Outside the US (Purity):* This category amasses all the tweets written in languages other than English. In order to partially ensure the purity of our data we tried to use only tweets from eligible and prospective voters through the elimination of these tweets.The underlying assumption being that a tweet in another language is less likely to have been written by someone within the United States. Manual inspection of the data allows us to further broaden this category by including tweets in English from people who are clearly not prospective voters. Although "Wish I lived in America to vote for Candice Glover. Exception performance #idolTop6" would normally be given a high sentiment score. However, the content of the tweet makes it possible for us to place it within category 5. Another example would be: "@AngieAI12 You know, I'm from the Philippines and I'm telling everyone that Angie Miller's gonna win! Go for it!"

Category 6 – *Rubbish & Collection Errors:* This last category is a means of classifying all the defective tweets that fall within neither of the two previous categories. All three categories together thus contain the data that needs to be removed before conducting the analysis. For the most part tweets in the sixth category are the result of an error that occurred when transferring data from the streaming API to Excel (e.g. Tweets that lack the candidate's name).

4.4 Data Analysis

4.4.1 Volume Model

Models predicting share of votes with tweet volume are based on the assumption that the number of votes contestants receive is proportional to the attention they receive on Twitter. Thus by aggregating the number of tweets referring to each candidate, it is possible to calculate their share of votes according to Twitter. More specifically we will employ the following formula:

$$Vol_{share}\ (x) = \frac{Tweet\ counts \wedge candidate\ x}{Total\ tweets\ of\ all\ candidates}$$

4.4.2 Sentiment Model

The studies in the subject domain of elections have incorporated sentiment into their prediction models in different ways – employing diverse ratios to represent the sentiment distribution of the tweets (see Appendix C). Bermingham et al. (2011) distinguished between two types of sentiment measures: intra-party and inter-party measures which we will refer to as intra-candidate and inter-candidate. Intra-candidate measures consider a candidate in isolation, computing a value representing how positive or negative the collected tweets are for a given candidate (i.e. the ratio of positive counts over negative counts mentioning the candidate). In order to make prior event predictions with intra-candidate measures it is necessary to first have few instances of the dependent variable (election results) to allow for the computation of the coefficients of the regression analysis. However as each election is different it is not possible to use this method for prior event prediction. Consequently, studies that had incorporated such measures into their models relied on regression analysis to make post event predictions of elections results. Inter-candidate measures, on the other hand, reflect the relative share of positive and negative tweets between the parties. If we assume that a positive sentiment indicates an intention to vote for a given candidate, the inter-candidate measure of relative share of positive tweets may be used to predict the percentage of votes a candidate will receive before publication of the results. Since the aim of our study is to predict the elimination of candidates, this measure seems most appropriate. More specifically,

$$Sentiment\ votes_p(x) = \frac{Pos.tweets \wedge candidate\ x}{Total\ pos.\ twets \wedge all\ candidates}$$

Figure 4.2 provides a graphic depiction of the research methodology employed.

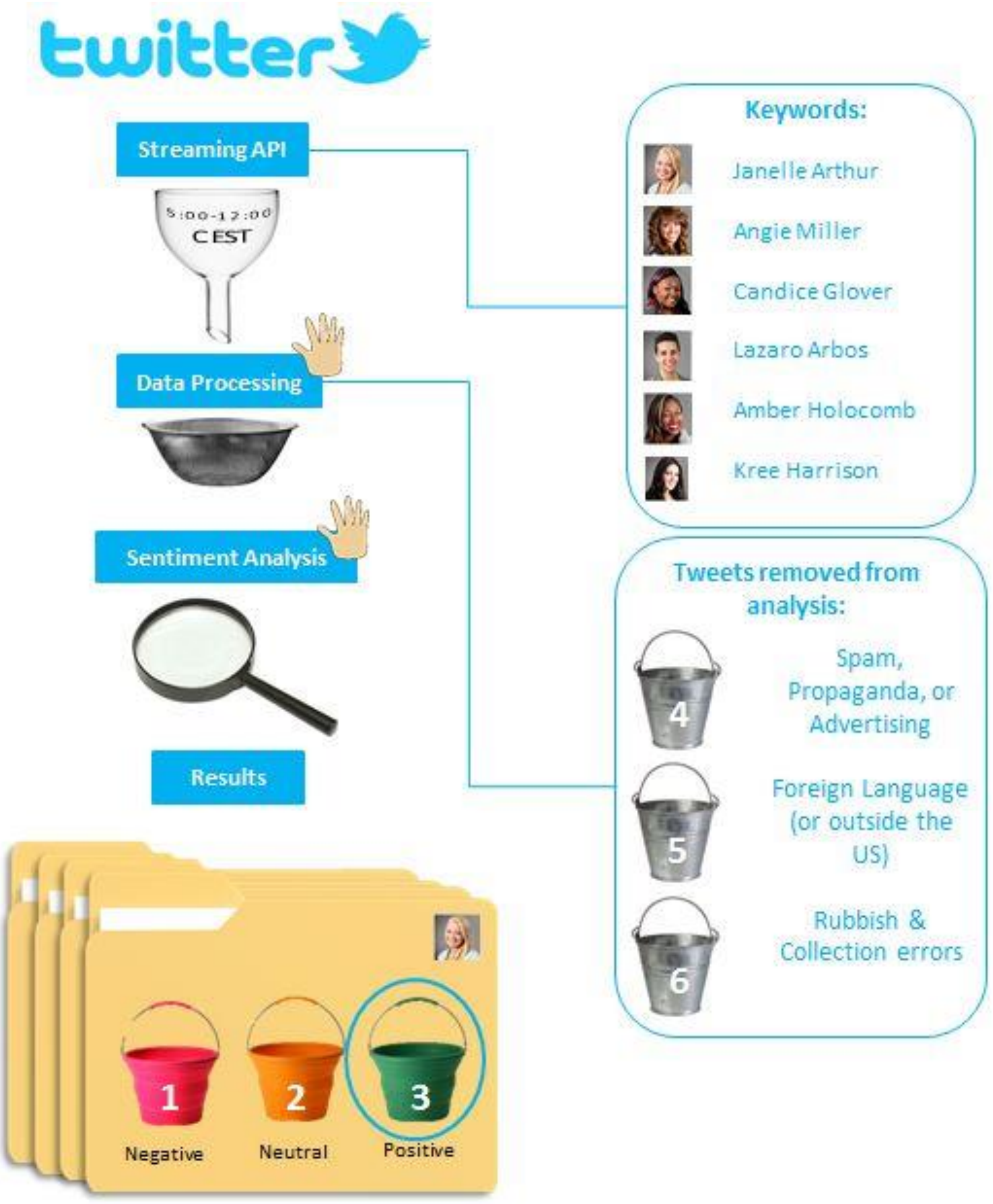


Figure 4.2: Methodology Illustration

5. Results

Table 5.1 Ciulla et al. (2012) American Idol 2012 - Results: Top 7 to Top 2

Day	Eliminated Cont.	Data Indicators	Bottom 3
May 17	Joshua	✓	N / A
May 10	Hollie	✓	N / A
May 3	Skylar	✓	(3/3)
April 26	Elise	CC	(3/3)
April 19	Colton	x	(3/3)

Table 5.2 American Idol 2013 - Results: Top 6 to Top 2 (Vol)

Day	Eliminated Cont.	Data Indicators	Bottom 2
May 9	Angie	x	N / A
May 2*	Amber	x	N / A
April 25	N / A	N / A	(1/2)
April 18	Janelle	✓	(2/2)
April 11	Lazaro	x	(0/2)

*Votes from April 25 and May 2 were combined for this elimination

Table 5.3 American Idol 2013 - Results: Top 6 to Top 2 (Sent)

Day	Eliminated Cont.	Data Indicators	Bottom 2
May 9	Angie	x	N / A
May 2*	Amber	CC	N / A
April 25	N / A	N / A	(1/2)
April 18	Janelle	✓	(2/2)
April 11	Lazaro	CC	(1/2)

*Votes from April 25 and May 2 were combined for this elimination

Table 5.1 presents the findings of Ciulla et al. (2012) for American Idol 2012 for rounds 'Top 7' to 'Top 2'. Column 3 of the table indicates whether or not the authors were able to predict the eliminated contestant ('check': successful prediction; 'x': predicted the wrong contestant; and 'CC': Too close to call- due to overlapping confidence intervals). In the last column Ciulla et al. present the number of contestants they were able to successfully predict as belonging to the bottom 3. We have summarized the results of our research (American Idol 2013) in a similar table for comparative purposes. As a consequence, Table 5.2 shows the results of applying a volume-based analysis to the collection of tweets (after removing categories 4,5,6). Table 5.3 shows the results of further classifying the volume according to sentiment and eliminating 'neutrals' and 'negatives'.

For the final dataset and classification for each week please refer to Appendix G and H.

By comparing the results of volume and sentiment analysis (table 5.2 and 5.3) it can be seen that sentiment allows us to more accurately predict eliminations and Bottom 2 contestants. For instance, in the week of April 11 classification of tweets according to sentiment reduced Lazaro's share of tweets from 5.8% to 3.4%. This is due to the fact that 24% of total tweet volume concerning Lazaro in the week of April 11 was of a negative nature. Although this presentation suggests that sentiment analysis allows making more accurate predictions it should be noted that the results obtained for both methodologies are not very accurate when compared to the results of Ciulla et al. (2012).

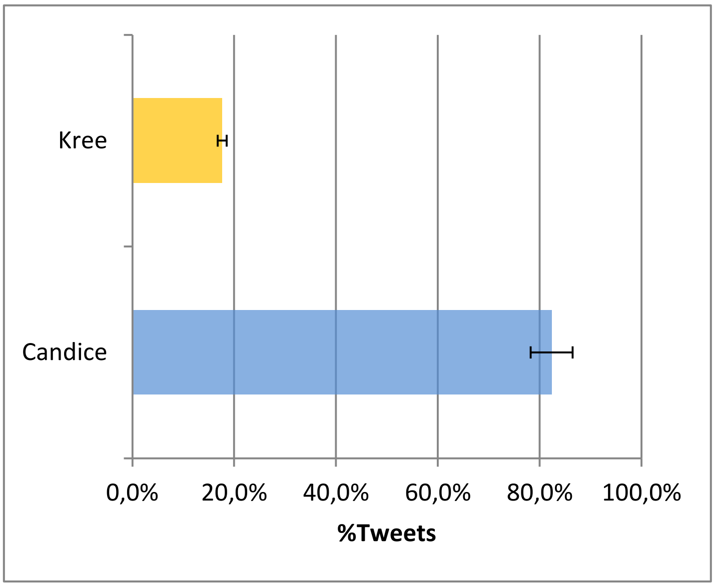
5.1 And the Winner is…

In their paper Ciulla et al. (2012) present a section entitled "And the Winner is…", in which they show how adjusting for geo-location allowed them to successfully predict the winner of American Idol. Although, Graph 5.1 shows our "accurate" prediction of the final round (Candice won American Idol), the fact that the show did not announce the share of votes leaves us unable to see to what extent our predictions were accurate. As a consequence, instead of presenting and focusing on the instances in which we were able to successfully predict the outcomes of American Idol, we would like to adopt a different tone.

First, by looking at table 5.2 and Appendix H one can see that we attempted to predict a total of five rounds of eliminations. Of those five rounds, our data would have incorrectly predicted the elimination of a contestant on one occasion; twice, the results were 'too close to call', and only two times were we able to correctly predict an elimination.

Second, we would like to stress the importance of how one presents the data. If we take for instance the week of May 9. An analysis of table 5.2 would show that we were unable to predict this particular elimination. This way of presenting the results does not, however, convey the extent to which our prediction was inaccurate. Even after classifying the tweets according to sentiment Angie (who was eliminated) still had 38.2% of positive tweets, compared to Kree's 9.9%.

Third, it would appear as if certain contestants were underrepresented on Twitter. Kree, for instance was constantly predicted as belonging to the bottom half of the contestants. Only for the two rounds in which she did end up in the Bottom 2 (April 18 and May 16) were our predictions with regard to Kree correct. The three times in which Kree was in the Top 2 our model was either unable to predict her ranking (CC) or placed her in the Bottom 2 as being eliminated. This misrepresentation may have to do with her particular profile, as she comes from Woodville, Texas and sings mostly country music. Indeed, Texas' twitter usage is among the lowest in the United States (Miller, 2011).



Graph 5.1: Relative Percentage of Tweets: Final (May 16, 2013)

Last, we have discovered that certain contestants are overrepresented in the dataset. For each of the weeks of April 11 to May 9, together, Candice and Angie have accumulated between 78.5% and 91.1% of the total positive tweet volume. With only 21.5% to 8.9% of positive tweets divided between the other contestants and confidence intervals of 5%, successfully predicting *eliminations* is thus very difficult.

5.2 The underlying assumption about voting intention

Research in the field of electoral predictions with Twitter is based on the assumption that the number or sentiment of tweets correlates with people's actual voting behavior. Sentiment-research in particular assumes that a tweet such as "I love Barack Obama" would be associated with that person voting for Obama. In order to test this assumption we have isolated over 2500 usernames of people tweeting about a contestant of American Idol. We then proceeded to sending tweets through multiple accounts asking these users which contestant they actually voted for. Even though we only received 134 answers, the results looked promising, as 112 users actually voted for the person they were supporting in their tweets (positive sentiment). Out of the remaining users, 2 tweeted a compliment to a contestant (e.g. "Kree Harrison has amazing vocals") but voted for another one and 20 did not actually vote. Although generally this attempt at empirically proving this assumption has proven successful, the results should still be considered with caution. Of the 2500 users only 134 answered, suggesting that those answering were the most active fans and that their answers may be subject to self-selection bias.

5.3 Limitations

5.3.1 Tweets from users outside of the US

Although we have tried to eliminate tweets that clearly showed that the user is outside of the US (either by content or language – categories 5), the possibility remains that many of the tweets about contestants of American Idol may still be from people unable to vote and influence the outcome of American Idol. Geo-location, as utilized by Ciulla et al. (2012), represents a possible way of solving this problem. To this end, we collected 1,500 tweets through Twitter's GetSample function (returns a small sample of all public tweets) and found that only 16 users had included their location within their profile. These results are supported by the research conducted by Syosomos Inc (2010) who found that on average only 0.23% of all tweets are tagged with a geo-location. As a consequence, research hoping to make use of geo-location would in all likelihood have to rely on commercial software (Syosomos Inc., 2010).

5.3.2 Measurements of Performance

As a result of the relative share of votes not being announced, we were unable to calculate to what extent our model's predictions deviated from the actual rates. The consequences of this characteristic were twofold. On the one hand it did not allow us to compute an effect size that would be comparable to that of previous studies conducted in this field. On the other, it has complicated our attempt at improving the methodology of Ciulla et al. Although there was some evidence of negative sentiment in tweets being inversely related to the relative share of votes, the lack of a dependent variable prevented us from empirically testing this hypothesis without resorting to a recipe-like model.

5.3.3 Decreasing Viewership

A possible explanation for the difference in accuracy with regard to the predictions made by Ciulla et al. (2012), could be a 39% decline in viewership in the age group 18-49. This decline, attributable to the success of the new TV show 'The Voice' may be responsible for a demographic shift (City Press, 2013).

5.3.4 Twitter as a social phenomenon

An increasing number of research papers have described the predictive power of data extracted from Twitter. It is unclear to what extent this information has impacted the way in which users and companies use this social media platform. Particularly, the paper by Ciulla et al. (2012), has captivated a considerable amount of media attention. For the first time in eleven years, this season of American Idol every contestant on the show had an official Twitter page (Oblivious A.T., 2013). It is possible that this is a result of this media attention and an increasing number of blogs claiming that the mere number of Twitter followers can predict the winner. This may very well constitute such a change of behavior, and may thus be partly responsible for the overall low predictive power of our model when compared to the results obtained by Ciulla et al. (2012).

5.3.5 Positive Sentiment Bias

The high level of positive sentiment we found was striking, representing over 80% of the tweets in our dataset (Appendix F). As a result, the impact of the negative tweets on the results is greatly diminished, possibly reducing the effect of sentiment on the prediction. This large positive bias may be partly explained by the psychology of sharing positive image among followers (Wong et al., 2012), but more likely due to the nature of the show. Namely, the onscreen conflict level between American Idol's contestants is rather low; they appear like a unified group, supporting each other before eliminations and posting pictures together on Facebook and Twitter. It is possible that the effect of sentiment analysis is stronger in political contexts, where the conflict level is high and each candidate represents an ideology that is often time at odds with that of the other candidate. As a result, Twitter users identifying with a particular stand are more likely to overtly support their candidate and object to the opposing view.

5.4 Conclusion

In this study we have extracted data from Twitter on the subject of American Idol in order to predict the show's eliminations and contestants' rankings. We employed two prediction methods: The first one was based on the work of Ciulla et al. (2012) who attempted to predict the eliminations of contestants from American Idol 2012, by counting the number of tweets mentioning a contestant (Volume). The second method aimed to improve the accuracy of the model by analyzing the sentiment of the tweets. The assumption underlying this second method was that positive tweets are more closely associated with voting intention.

We have found that a classification of the tweets according to sentiment, allows for more accurate predictions with regard to the subsequent ranking of a contestant. These findings are in line with the conclusions we have drawn from our Meta-analysis, in which we showed that studies using sentiment, rather than volume were able to reduce their MAE.

Although we found evidence in support of our hypothesis that the inclusion of sentiment would improve the model's accuracy, we would nonetheless like to stress that the overall predictive power of our model was low. We believe this to be a consequence of a lack of representativeness in the Twitter user base. As this is a problem that other researchers in this field have also encountered (Avello, 2012), we would suggest that future research be centered around more fundamental issues with regard to using Twitter to make predictions:

1) Although research has been conducted to compare the samples generated by Streaming – and Search API (Gonzales-Bailon et al., 2012), no study has examined the representativeness of either of these methods compared to the entirety of the Twitter data. Researchers tend to regard these methods as 'black boxes', assuming that their sample is representative of the entire Twitter population.

2) An increasing amount of evidence has pointed to the demographic bias present within the Twitter user base (PewResearchCenter, 2012; 2013). Although we believed the demographics of the Twitter users on the one hand and viewers of American Idol on the other to correlate more strongly than the population of an average election, this study has shown that certain contestants are grossly under- and over-represented. Unable to predict the eliminations of American Idol we believe political elections to constitute an even more difficult domain. As a consequence, correcting for demographic bias should be the main focus of attention for any paper trying to make predictions with Twitter. In a recent study DiGrazia et al. (2013) make an attempt to address this issue. They investigate the correlation between the share of Republican candidate name mentions and the corresponding vote margin controlling, among other factors, for demographic variables (race, gender, and age). While the authors report an adjusted R-squared of 87%, they fail to explain how the demographic data about Twitter users was obtained (a critical issue as these details are not publicly disclosed), and how the data can be adjusted *ex-ante*.

3) In this study we have attempted to question and test the underlying assumption that a positive tweet about a contestant/candidate will result in a vote. Due to a low response rate only limited conclusions can be drawn from this attempt. As a consequence, we encourage further research examining this assumption upon which much of today's research is based.

4) In order to ensure optimal sentiment-classification we have analyzed the dataset manually. This has allowed us to pre-process the data with regard to spam. As demonstrated in our literature review this pre-processing positively impacts a model's accuracy. As a consequence, we believe the programming of software that reliably recognizes spam, propaganda, and misrepresentation to be essential to the development of this field. Similar functions can already be found over the web. For example, StatusPeople.com allows its users to perform a "Fake Follower Check" analysis of Twitter accounts. The incorporation of similar applications in the domain of election can prove very useful.

Finally, we believe that the use of more advanced algorithms and new software innovations may be the key to more effective data analysis. For example, cross referencing users' profiles with other social media platforms can reveal information, which can be used to correct for demographic bias. At the same time, we would like to note that these advancements are likely to go hand in hand with infringement upon users' right to privacy.

Acknowledgements

We would like to thank Fabio Ciulla for providing us with additional information with regard to the methodology employed and advice on which rounds of the show to focus on. We would also like to thank Clement Levallois for helping us in collecting the data and offering us constructive feedback along this research.

Bibliography

Asur, S., & Huberman, B.A., (2010). 'Predicting the Future with Social Media'. [Online] Available at: < http://www.ibisworld.com/industry/default.aspx?indid=1246> [Accessed 1 Febuary 2013].

Bermingham, A, Smeaton A (2011). 'On using twitter to monitor political sentiment and predict election results'. In Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology, Asian Federation of Natural Language Processing, , Chiang Mai, Thailand, pp. 2–10.

Bollen, J, Mao, H, Zeng, X-J, (2010) 'Twitter mood predicts the stock market'. [Online] Available at: < http://arxiv.org/abs/1010.3003 > [Accessed 3 Febuary 2013].

Bryman, A., & Bell, E., (2007) *Business Research Methods*. 2nd ed. New York: Oxford University Press.

City Press, (2013). 'American Idol brings in record low TV audience'. [online] available at: <http://www.citypress.co.za/entertainment/american-idol-brings-in-record-low-tv-audience/ > [Accessed 20 May 2013]

Ciulla F, Mocanu D, Baronchelli A, Perra N, Gonçalves B, Vespignani A (2012). 'Beating the news using social media: the case study of American Idol'. [online] Available at: <arXiv:1205.4467> [Accessed 19 May 2013]

DiGrazia, J., McKelvey, K., Bollen, J. and Rojas, F., (2013). 'More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior'. [online] Available at: <http://ssrn.com/abstract=2235423 or http://dx.doi.org/10.2139/ssrn.2235423> [Accessed 30 May 2013]

Gallaugher J.2012, 'Information Systems: A Manager's Guide to Harnessing Technology", version 1.4 Lodon: Flat World Knowledge

Gayo-Avello D (2011) 'I wanted to predict elections with Twitter and all I got was this Lousy Paper' - A balanced survey on election prediction using Twitter data.' [online] available at: <arXiv:1204.6441.> . [Accessed 19 May 2013]

Gayo-Avello D (2012) 'A meta-analysis of state-of-the-art electoral prediction from Twitter data.' [online] available at: <arXiv:1204.6441.> . [Accessed 19 May 2013]

Gonzalez-Bailon, S., Wang, N., Rivero, A., Borge-Holthoefer, J. and Moreno, Y., (2012). 'Assessing the Bias in Communication Networks Sampled from Twitter'. [online] Available at: <http://ssrn.com/abstract=2185134 or http://dx.doi.org/10.2139/ssrn.2185134> [Accessed 19 May 2013]

Hak, T., (2012), Course Book: Research Training & Bachelor Thesis Course 2012-2013, 4th ed., [online] Available at: < http://bblp.eur.nl/webapps/portal/frameset.jsp?tab_tab_group_id=_2_1&url=%2Fwebapps%2Fblackboard%2Fexecute%2Flauncher%3Ftype%3DCourse%26id%3D_5800_1%26url%3D> [Accessed 20 January 2013].

Hernandez, BA, (2013), ‘'American Idol' Chooses Top 5 Themes From Twitter Submissions’. [online] Available at: <http://mashable.com/2013/04/11/american-idol-twitter-themes-vote/> [Accessed 20 April 2013].

Jungherr, A, Jürgens, P, and Schoen, H, (2011) “Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, t. o., sander, p. g., & welpe, i. m. "predicting elections with twitter: What 140 characters reveal about political sentiment"”. Social Science Computer Review, 2011. [Online] Available at: <http://www.sciencemag.org/content/338/6106/472.full.pdf > [Accessed 3 Febuary 2013].

Krauss, J, Nann, S, Simon, D, Fischbach, K, and Gloor, P., (2008), "Predicting Movie Success and Academy Awards through Sentiment and Social Network Analysis", ECIS 2008 Proceedings, Paper 116, [Online] Available at: http://aisel.aisnet.org/ecis2008/116 [Accessed 1 February 2013].

McCue, TJ, (2013), ‘Twitter Ranked Fastest Growing Social Platform In The World’, Forbes Online Magazine. [online] Available atL <http://www.forbes.com/sites/tjmccue/2013/01/29/twitter-ranked-fastest-growing-social-platform-in-the-world/> [Accessed 15 May 2013]

Metaxas P, Mustafaraj E, Gayo-Avello D (2011) How (not) to predict elections. In: Privacy, security, risk and trust (PASSAT), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (SocialCom). IEEE Press, New York. pp 165-171

Nielsen, 2013, ‘Top 10 List for Prime Broadcast Network TV – United States, Week of May 6, 2013’ and ‘Top 10 List for Prime Broadcast Programs Among African Americans, Week of May 6, 2013’. [online] Available at: <http://www.nielsen.com/us/en/top10s.html /> [Accessed 15 May 2013]

Oblivious A.T (2013). ‘Can Twitter Predict The American Idol Winner? [online] Available at: <http://obliviousat.com/2013/03/22/can-twitter-predict-the-american-idol-winner/> [Accessed 19 May 2013]

O’Connor B, Balasubramanyan R, Routledge B.R, Smith N.A (2010),“From tweets to polls: Linking text sentiment to public opinion time series,” Proceedings of the fourth international AAAI conference on weblogs and social media. AAAI Press, 2010, pp. 122–129.

Pew Internet (2012). 'Twitter use 2012'. [online] available at: <http://pewinternet.org/~/media//Files/Reports/2012/PIP_Twitter_Use_2012.pdf> [Accessed 19 May 2013]

Pew Internet (2013). 'Twitter Reaction to Events Often at Odds with Overall Public Opinion'. [online] available at: <http://www.pewresearch.org/2013/03/04/twitter-reaction-to-events-often-at-odds-with-overall-public-opinion/> [Accessed 19 May 2013]

Sagolla D, (2009). 'How Twitter was born' Available at: < http://www.140characters.com/2009/01/30/how-twitter-was-born/ > [Accessed 17 February 2013]

Tjong, E., & Bos, J. (2012). 'Predicting the 2011 Dutch senate election results with Twitter.' [online] Available at: < http://ifarm.nl/erikt/papers/sasn2012.pdf> [Accessed 19 May 2013]

Stelter, B 2013, 'Now on 'Idol', Viewers can tweet while contestants sing', New York Times online Magazine. [online] Available at: < http://www.nytimes.com/2013/02/27/arts/television/american-idol-fans-can-now-tweet-their-views.html?_r=0> [Accessed 19 May 2013]

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. P. 178-185. [online] Available at: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1441/1852 [Accessed 19 May 2013]

Twitter, (2013). 'About us'. [online]. Available at: <www.twitter.com/about > May 2013]

Appendix A: Sentiment Analysis Tools

Sentiment analysis refers to the process used to characterize the attitude, opinion, or emotion conveyed in text units.

Most Sentiment analysis tools involve classifying the polarity of the sentiment expressed in the text content, distinguishing between ‘positive’, ‘negative’ or ‘neutral’. Some also identify the potency – the degree to which the text unit is positive or negative. While other tools use multi-dimensional classifications that enable a richer annotation of human mood, such as ‘happy’, ‘calm’, or ‘alert’.

There are two main approaches to automatically extract sentiment from text:

Lexicon based method –starts with a lexical database storing polarity information for words, phrases or sentences, which are then identified in the text message and annotated according the database’s sentiment score. The individual scores are subsequently aggregated into an overall sentiment score/tag for the text (Taboada et al., 2011).

Machine learning – a corpus of manually labeled instances of words, phrases or sentences from a particular field is used to train an algorithm to identify properties and patterns that are associated with positive, negative or neutral categories. The computer uses the algorithm to search for these patterns in new texts and annotates their sentiment accordingly (Thelwall et al., 2011).

There are numerous types of lexicon based methods and machine learning. These techniques offer the advantage of rapid automated analysis of the data collected. However, if not carefully chosen to fit the data type (e.g. tweets as supposed to blogs) and the subject domain (e.g. political content as supposed to text about movies) they are likely to produce limited accuracy. For instance, a study by Bermingham and Smeaton (2010) about sentiment analysis in Twitter found that unsupervised methods, like the use of sentiment lexicons, tend to be less accurate than machine learning. Similarly, Thelwall et al. (2011) report that the performance of machine learning tools drop when used in domains different from the ones they were trained on.

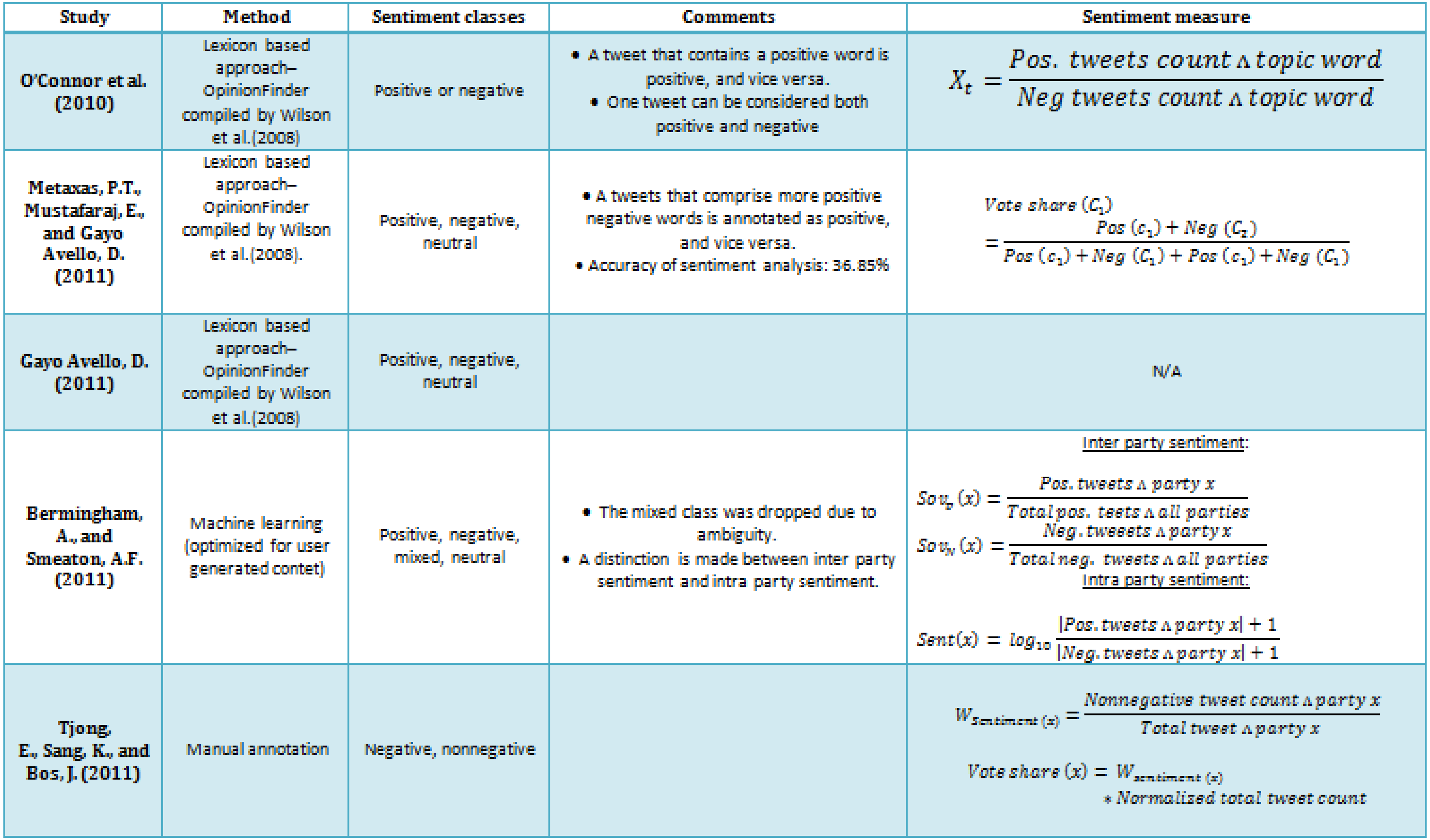
Several scholar studies as well as commercial companies specializing in brand monitoring have reported a relatively high accuracy (70%-80%) in predicting polarity of texts generated in Twitter or other social media sources using automated sentiment analysis (e.g. Bandwatchm, unknown; Bermingham, 2010; Asur and Huberman, 2010). However, evaluation of a number of sentiment tools by specialized blogs revealed that their accuracy “is largely driven by their ability to correctly label neutral posts” which typically comprise a significant amount of the tweets (Dunay, 2011; Rhodes 2010; Grimes, 2012). When only testing the annotation of tweets marked as positive and negative, accuracy rates dropped to around 30%.

Appendix B: Summary of the meta-analysis

Authors	Date	Election	Sampling Method	Data Cleansing	Prediction Method	Effect Size Measure	Result
O'Connor et al.	2010	US Presidential Election (Nov. 4, 2008)	Streaming API "Garden hose" (10% of the stream)	No Cleansing	Volume of tweets	Pearson's r correlation	74%
Tumasjan et al.	2010	German Federal Election (Sept. 27, 2009)	N/A	No Cleansing	Volume of tweets	Mean Absolute Error (MAE)	1.65%
Jungherr et al.	2011	German Federal Election (Sept. 27, 2009)	Streaming API	No Cleansing	Volume of tweets	MAE	1.5%, 3.3% (Different time periods)
Gayo-Avello	2011	US Presidential Election (Nov. 4, 2008)	Search API	Geolocation, in addition to debiasing according to usage	Lexicon based Sentiment Analysis (Opinion Finder)	MAE	13.1%
Metaxas et al.	2011	US State Elections (2010)	Streaming API "Garden hose"	No Cleansing	Volume of tweet.	MAE	17.2%
					Lexicon based Sentiment analysis (Opinion finder)	MAE	7.6%

Authors	Date	Election	Sampling Method	Data Cleansing	Prediction Method	Effect Size Measure	Result
Bermingham & Smeaton	2011	Irish General Election (Feb. 25, 2011)	Third-party application	Data was partially cleansed by eliminating tweets that were ambiguous/irrelevant	Volume of tweet.	MAE	5.58%
					Machine Learning Sentiment analysis	MAE	3.67%
Tjong et al	2012	Dutch Senate Election (May. 23, 2011)	N/A	**Sentiment**:Attempt to debias data according to political leaning by using pre-electoral polling data	Sentiment Analysis	MAE	1.45%
Skoric et al	2012	Singaporean General Election (May 7, 2011)	N/A	Only use of tweets located in Singapore	Volume of tweet.	MAE	5.23%

Appendix C: Overview of sentiment analysis methods



Study	Method	Sentiment classes	Comments	Sentiment measure
O'Connor et al. (2010)	Lexicon based approach– OpinionFinder compiled by Wilson et al.(2008)	Positive or negative	• A tweet that contains a positive word is positive, and vice versa. • One tweet can be considered both positive and negative	$X_t = \frac{Pos.\ tweets\ count \wedge topic\ word}{Neg\ tweets\ count \wedge topic\ word}$
Metaxas, P.T., Mustafaraj, E., and Gayo Avello, D. (2011)	Lexicon based approach– OpinionFinder compiled by Wilson et al.(2008).	Positive, negative, neutral	• A tweets that comprise more positive negative words is annotated as positive, and vice versa. • Accuracy of sentiment analysis: 36.85%	$Vote\ share\ (C_1) = \frac{Pos\ (c_1) + Neg\ (C_2)}{Pos\ (c_1) + Neg\ (C_1) + Pos\ (c_1) + Neg\ (C_1)}$
Gayo Avello, D. (2011)	Lexicon based approach– OpinionFinder compiled by Wilson et al.(2008)	Positive, negative, neutral		N/A
Bermingham, A., and Smeaton, A.F. (2011)	Machine learning (optimized for user generated contet)	Positive, negative, mixed, neutral	• The mixed class was dropped due to ambiguity. • A distinction is made between inter party sentiment and intra party sentiment.	Inter party sentiment: $Sov_p\ (x) = \frac{Pos.\ tweets \wedge party\ x}{Total\ pos.\ teets \wedge all\ parties}$ $Sov_N\ (x) = \frac{Neg.\ tweeets \wedge party\ x}{Total\ neg.\ tweets \wedge all\ parties}$ Intra party sentiment: $Sent(x) = log_{10} \frac{\lvert Pos.\ tweets \wedge party\ x \rvert + 1}{\lvert Neg.\ tweets \wedge party\ x \rvert + 1}$
Tjong, E., Sang, K., and Bos, J. (2011)	Manual annotation	Negative, nonnegative		$W_{Sentiment\ (x)} = \frac{Nonnegative\ tweet\ count \wedge party\ x}{Total\ tweet \wedge party\ x}$ $Vote\ share\ (x) = W_{sentiment\ (x)} * Normalized\ total\ tweet\ count$

* "Where C1 is the candidate for whom support is being computed while C2 is the opposing candidate; pos(c) and neg(c) are, respectively, the number of positive and negative tweets mentioning candidate c".

Appendix D: Inter-annotator Agreement

To increase the validity of our results, each tweet was labeled by two annotators. The labels annotator A assigned to the tweets were compared to those of annotator B, producing the following confusion matrix (table 1):

		Annotator A			
		Negative	Neutral	Positive	Total
Annotator B	Negative	**58**	7	0	65
	Neutral	0	**354**	36	390
	Positive	7	123	**1748**	1878
	Total	65	484	1784	2333

Table 1: confusion matrix

Interpretation of S (Viera, 2005)
< 0 Less than chance agreement
0.01–0.20 Slight agreement
0.21– 0.40 Fair agreement
0.41–0.60 Moderate agreement
0.61–0.80 Substantial agreement
0.81–0.99 Almost perfect agreement

Since human judgment is involved, we cannot directly assess for its correctness (Fort, 2011), for example the tweet: "Damn Angie Miller!" can be interpreted as positive or negative depending on the tone perceived by the coder. Instead, we check for the consistency with which annotators agree with each other using the S statistic (Bennett et al., 1954):

$$S = \frac{A_O - A_e}{1 - A_e}$$

Where A_o denotes observed agreement, and A_e denotes expected agreement

This coefficient takes into account the chance factor, assuming that all categories (Positive, Neutral and Negative) are equally likely to be chosen by both annotators. It yields a value between 0, indicating chance agreement, and 1, indicating total agreement (Artstein et al., 2008).

Calculations:

Observed Agreement:

$$A_o = \frac{58 + 354 + 1748}{2333} = \frac{2160}{2333} = 0.926$$

Expected Agreement:

$$A_o = q \times \left(\frac{1}{q}\right)^2 = \frac{1}{3} \times \left(\frac{1}{1/3}\right)^2 = \frac{1}{3}$$

Inter Annotator Agreement:

$$S = \frac{A_0 - A_e}{1 - A_e} = \frac{0.926 - 0.333}{1 - 0.333} = \frac{0.59}{0.67} \approx 0.89$$

Appendix E: Umigon's performance

Umigon's performance according to 3 widely accepted indicators adopted by Sokolova et al. (2009)

$$Accuracy = \frac{true\ pos. + true\ neg. + true\ neut}{all\ tweets} = \frac{40 + 229 + 960}{2333} = 0.527$$

$$Precision\ positive = \frac{true\ pos.}{true\ pos. + false\ pos.} = \frac{960}{1081} = 0.89$$

$$Recall\ positive = \frac{true\ pos.}{true\ pos. + false\ neg.} = \frac{960}{1221} = 0.79$$

$$Precision\ negative = \frac{true\ neg.}{true\ neg. + false\ neg.} = \frac{40}{301} = 0.13$$

$$Recall\ negative = \frac{true\ neg.}{true\ neg. + false\ pos.} = \frac{40}{121} = 0.33$$

Appendix F: Positive Tweets

We observed a high level of positive sentiments, comprising 80% of the total tweets (table below). For the most part it seems that twitter users prefer to express their support by posting complements and sharing videos, rather than 'trashing' the other contestants.

	Positive	Neutral	Negative	Total
Candice	1396	185	6 (0.4%)	1587
Angie	314	97	8 (1.7%)	418
Kree	62	28	2 (2.1%)	92
Amber	53	39	4 (4.1%)	96
Lazaro	31	20	50 (49.5%)	101
Janelle	11	27	1 (2.6%)	39
	1867 (80%)	396 (17%)	72 (3%)	2333

Appendix G: American Idol 2013 – Classification of Tweets per week

Apr 11

	Positive	Neutral	Negative	4,5,6
Angie	1047	285	58	356
Candice	4190	279	45	2051
Janelle	156	79	6	46
Amber	219	138	12	98
Kree	201	72	6	84
Lazaro	203	112	102	87

Apr 18

	Positive	Neutral	Negative	4,5,6
Angie	1276	503	53	384
Candice	764	342	31	312
Janelle	89	54	14	82
Amber	237	92	13	87
Kree	234	61	16	95

Apr 25

	Positive	Neutral	Negative	4,5,6
Angie	2162	423	63	897
Candice	621	187	23	285
Amber	389	76	32	252
Kree	298	53	13	164

May 02

	Positive	Neutral	Negative	4,5,6
Angie	1742	420	87	948
Candice	1365	614	53	1132
Amber	542	502	132	584
Kree	536	143	16	335

May 09

	Positive	Neutral	Negative	4,5,6
Angie	1592	251	102	706
Candice	2163	189	41	785
Kree	413	52	7	319

May 16

	Positive	Neutral	Negative	4,5,6
Candice	3471	409	73	1012
Kree	743	187	26	339

Appendix H: American Idol 2013 – Predictions (Volume and Sentiment)*

Prediction based on SENTIMENT Analysis

April 11	Mentions	Share	95% UB	95% LB	Prediction	Result
Angie	1047	17,4%	18,4%	16,4%	Top 2	Safe
Candice	4190	69,6%	70,8%	68,5%	Top 2	Top 2
Janelle	156	2,6%	3,0%	2,2%	CC	Safe
Amber	219	3,6%	4,1%	3,2%	CC	Bottom 2
Kree	201	3,3%	3,8%	2,9%	CC	Top 2
Lazaro	203	3,4%	3,8%	2,9%	CC	Eliminated

April 18	Mentions	Share	95% UB	95% LB	Prediction	Result
Angie	1276	49,1%	51,0%	47,2%	Top 2	
Candice	764	29,4%	31,1%	27,6%	Top 2	
Janelle	89	3,4%	4,1%	2,7%	Eliminated	Eliminated
Amber	237	9,1%	10,2%	8,0%	CC	
Kree	234	9,0%	10,1%	7,9%	CC	Bottom 2

April 25	Mentions	Share	95% UB	95% LB	Prediction	Result
Angie	2162	62,3%	63,9%	60,7%	Top 2	Top 2
Candice	621	17,9%	19,2%	16,6%	Top 2	Bottom 2
Amber	389	11,2%	12,3%	10,2%	Bottom 2	Bottom 2
Kree	298	8,6%	9,5%	7,7%	Bottom 2	Top 2

Prediction based on VOLUME Analysis

April 11	Mentions	Share	95% UB	95% LB	Prediction	Result
Angie	1390	19,3%	20,2%	18,4%	Top 2	Safe
Candice	4514	62,6%	63,7%	61,5%	Top 2	Top 2
Janelle	241	3,3%	3,8%	2,9%	Bottom 2	Safe
Amber	369	5,1%	5,6%	4,6%	Safe	Bottom 2
Kree	279	3,9%	4,3%	3,4%	Bottom 2	Top 2
Lazaro	417	5,8%	6,3%	5,2%	Safe	Eliminated

April 18	Mentions	Share	95% UB	95% LB	Prediction	Result
Angie	1832	48,5%	50,1%	46,9%	Top 2	
Candice	1137	30,1%	31,5%	28,6%	Top 2	
Janelle	157	4,2%	4,8%	3,5%	Eliminated	Eliminated
Amber	342	9,1%	10,0%	8,1%	CC	
Kree	311	8,2%	9,1%	7,4%	CC	Bottom 2

April 25	Mentions	Share	95% UB	95% LB	Prediction	Result
Angie	2648	61,0%	62,5%	59,6%	Top 2	Top 2
Candice	831	19,1%	20,3%	18,0%	Top 2	Bottom 2
Amber	497	11,5%	12,4%	10,5%	Bottom 2	Bottom 2
Kree	364	8,4%	9,2%	7,6%	Bottom 2	Top 2

* “CC” in the prediction section refers to ‘Too close to call’. This occurs whenever the Confidence Intervals (between the Upper Boundary (UB) and Lower Boundary (LB)) of two or more contestants overlap, making it impossible to determine which class the contestant is likely to belong to. The colours indicate to what extent our predictions were accurate. First, ‘Green’ shows that we correctly classified a contestant (top 2, safe, or bottom 2); Second, ‘Orange’ signals that our prediction is off by one category (e.g. classifying a contestant as ‘safe’ instead of ‘top2’; Third, ‘Red’ indicates that our prediction is off by two categories (e.g. ‘bottom 2’ instead of ‘top 2’); Last, ‘Blue’ was used when the classification was not announced or overlapping confidence intervals (CC) made it impossible to classify a contestant.

Prediction based on SENTIMENT Analysis (cont.)

May 02*	Mentions	Share	95% UB	95% LB	Prediction	Result
Angie	3904	51,0%	52,1%	49,9%	Top 2	Eliminated
Candice	1986	25,9%	26,9%	25,0%	Top 2	
Amber	931	12,2%	12,9%	11,4%	CC	
Kree	834	10,9%	11,6%	10,2%	CC	

May 09	Mentions	Share	95% UB	95% LB	Prediction	Result
Angie	1592	38,2%	39,7%	36,7%	Top 2	Eliminated
Candice	2163	51,9%	53,4%	50,4%	Top 2	Top 2
Kree	413	9,9%	10,8%	9,0%	Eliminated	Top 2

May 16	Mentions	Share	95% UB	95% LB	Prediction	Result
Candice	3471	82,4%	83,5%	81,2%	Winner	Winner
Kree	743	17,6%	18,8%	16,5%	Runner-up	Runner-up

Prediction based on VOLUME Analysis (cont.)

May 02*	Mentions	Share	95% UB	95% LB	Prediction	Result
Angie	4897	46,7%	47,6%	45,7%	Top 2	
Candice	2863	27,3%	28,1%	26,4%	Top 2	
Amber	1673	15,9%	16,6%	15,2%	Bottom 2	Eliminated
Kree	1059	10,1%	10,7%	9,5%	Eliminated	

May 09	Mentions	Share	95% UB	95% LB	Prediction	Result
Angie	1945	40,4%	41,8%	39,0%	Top 2	Eliminated
Candice	2393	49,8%	51,2%	48,3%	Top 2	Top 2
Kree	472	9,8%	10,7%	9,0%	Eliminated	Top 2

May 16	Mentions	Share	95% UB	95% LB	Prediction	Result
Candice	3953	80,5%	81,6%	79,4%	Winner	Winner
Kree	956	19,5%	20,6%	18,4%	Runner-up	Runner-up

*No contestant was eliminated on April 25, instead the votes of the week were added to those of the May 02 week.