Paper 131-2012

Predicting Electoral Outcomes with SAS ® Sentiment Analysis and SAS ® Forecast Studio

Jenn Sykes, SAS Institute Inc., Cary, NC

ABSTRACT

With the wide proliferation of text-based data on the Internet, there are more opportunities for organizations to validate the claims made in their structured data. Here, I use a combination of SAS® Sentiment Analysis and SAS® Forecast Studio to predict the outcomes of popular elections when polling data is not readily available. I also use these tools to validate the outcomes of elections and check for potential instances of fraudulent election administration. I use examples from the real world of politics and the popular television show, *American Idol*.

INTRODUCTION

Imagine you are the grunt in the State Department who has to decide whether to evacuate Egypt. Or that you are at the United Nations and need to decide whether to deploy troops to monitor an upcoming election. Most of your work will come after the process has already begun. What if you could determine if a process was about to begin? Although there are avenues to do this with structured data, the addition of social media data could be of great benefit to bureaucrats trying to best allocate resources in response to a crisis. At a more mundane level, social media can be used by governments to help with mobilization after an earthquake or to determine what quantity of flu vaccines need to go where. Private corporations are already making use of social media in this way. Now it is time for governments to use social media as well.

Thanks to the rise of social media as a common tool for expression of public sentiment, businesses, governments, and other organizations have developed serious interests in using social media to help determine how consumers and potential consumers respond to their products and services. Research has shown that Twitter mirrors the aggregate public sentiment in terms of identifying stock market trends, citizen happiness, and candidate ideology. It appears that social media can improve upon the work done by traditional market researchers to maximize profits or predict the next big issue in the eye of the public. However, most social media applications rely on real-time monitoring instead of forecasting to help optimize decision making .

There are certain drawbacks to using social media to analyze consumer sentiment. In the United States (U.S.), over a quarter of Twitter users are not eligible to vote, and the younger people who are power users on social media sites are only now developing purchasing power in the marketplace. Recent work has been conducted on the veracity of posts on review sites, such as Trip Advisor and Yelp.

However, the benefits of this new data source certainly deserve some exploring. Elections are something many people are interested in, especially the results. For some people, elections help determine what policy changes might take place. These policy changes might require budget changes. For other people, having a good idea of the outcome of an election can stabilize markets or resource allocation. Elections are also a process that takes time and often do not have a great set of structured data to forecast from. In this paper, I will use SAS® Sentiment Analysis and SAS® Forecast Server to analyze two cases involving popular elections in the U.S.: the current Republican Presidential primaries and the 2011 *American Idol* television contest. While these cases involve elections, the lessons learned from them can be applied to other business cases, as I will discuss. The main goal is to use this combination of applications from SAS® to:

- forecast electoral outcomes before the election occurs
- compare public sentiment as expressed in social media to the actual election results
- review the pitfalls of using social media data to generalize for the population at large
- explain how social media forecasting can help with standard business processes

METHODS

In both examples, I use a four step-process to extract, validate, analyze, and predict electoral outcomes from the relevant social media data.

- *Extract* a set of Tweets about the candidate of interest with SAS® code.
- *Filter* the Tweets to ensure that the keyword pulls are relevant through SAS® code.

- *Analyze* the Tweets for positive or negative sentiment around a candidate using SAS® Sentiment Analysis.
- *Predict* contest winners based on the aggregate sentiment scores for the candidate of interest over time using SAS® Forecast Server.

U.S. PRESIDENTIAL ELECTIONS

BACKGROUND

For at least two years, the Republican Party of the United States (or GOP) has been trying to select a candidate to compete against current Democratic President Barack Obama. During this time, the press and public have labeled different candidates as the "frontrunner" for the nomination, with former Massachusetts Governor Williard Mitt Romney usually carrying that mantle.

The first actual elections in the nomination contest were the Iowa Caucuses that were held on January 3, 2012. These caucuses were followed closely by the New Hampshire primary on January 10, and then the primaries in South Carolina and Florida at the end of January. A number of other states have also held primaries or caucuses between January and "Super Tuesday" on March 6. The winner of each caucus or primary is simply the candidate with a plurality in the state.[1] The Iowa Caucuses began with essentially a tie between former Governor Romney and former Pennsylvania Senator Richard John "Rick" Santorum, but Santorum was declared the winner just before the South Carolina primary. Governor Romney won the New Hampshire contest, but the next week, former Speaker of the House Newton Leroy "Newt" Gingrich won the South Carolina primary. After that, Governor Romney won Florida and Nevada, and Senator Santorum won three contests the next week. From there, the race became something of a two-man show with Romney as the more moderate candidate and Santorum representing the right-wing of the party. Both of these candidates have won a number of contests in February and March.

CAVEATS

In the case of the GOP nomination, there is reliable structured data to compare against the results of Twitter pulls. It is also possible to use the Tweet pulls to determine national versus statewide preference for a candidate. However, the breakdown by state via Twitter has limited use. There are two reasons for this. The first is that some U.S. states have more representative samples of likely voters that appear in Twitter. Second and more importantly, while SAS tools are able to extract the location of the Tweet author, the number of authors giving legitimate locations is too small to use that feature reliably.

In addition, from a national sample, there are two demographic trends to be aware of with Twitter data. The first is the "lovefest" for Ronald Ernest "Ron" Paul from the expected Twitter user--a younger person. For some reason, a septuagenarian obstetrician is popular with the youth of America. Secondly, there are a number of reporters and media sites tweeting facts, not opinions, about the individual candidates. This means when a candidate does something newsworthy, like propose a moon colony, the candidate is likely to be mentioned in a neutral way in the "Twitterverse" by the media.

With the candidates and social media, there is some coordination that could bias the results. Politicians using Twitter has become so popular that Twitter has created special features for politicians. Many staffers have Twitter accounts to serve as an echo chamber for the candidates. Blasts go out from the candidates asking supporters to retweet. In a sense, the GOP primary has some coordination involved to game the market. The question is how effective is it.

METHODS AND ANALYSIS

The Republican candidates analyzed initially were those that were still in the race by the time of the Iowa Caucuses on January 3. They are, in alphabetical order: Minnesota Congresswoman Michele Bachmann, Newt Gingrich, former Ambassador to China Jon Huntsman, Jr., Texas Congressman Ron Paul, Texas Governor James Richard "Rick" Perry, Mitt Romney, and Rick Santorum. I did not include Herman Cain or Stephen Colbert and their short-lived adventure in South Carolina, though information about these two candidates does appear in the sample. For this paper, the data goes through February 28. For the presentation, the data goes through March 7.

To collect the Twitter data, I wrote a query that would collect Tweets if there was an occurrence of any of the candidates' last names in the body of the Tweet.[2] These Twitter pulls began on January 1 and cover a national sample. Because Bachmann, Huntsman, and Perry dropped out before the South Carolina contests, I was unable to do a full set of analyses on them. This means that for analysis in SAS® Forecast Server I examined only Gingrich, Paul, Romney, and Santorum.

[1] This does not necessarily mean the candidates who win receive the delegates from that state for the national nominating convention, but that is not important for this exercise.

[2] Except for Ron Paul and Rick Perry, whose full names were used to avoid confusion with many people named "Paul" and the artist Katy Perry.

I wrote a SAS® Sentiment Analysis project to identify each candidate by their names, nicknames, relatives' names, and associated slogans and super political action committees (PAC). I then used the SAS Natural Language Processing (NLP) approach to sentiment analysis to write linguistic rules to determine if what was being said about these candidates was positive, negative, or neutral. In some cases, this involved looking for a positive or negative word used in close proximity to a candidate's name. (Examples would be "Ron Paul is amazing" or "Mitt Romney is a liar"). In other cases, there are certain keywords associated with only one candidate. (Michele Bachmann, for example, is the only candidate referred to as "The Iron Lady.")

SAS® Sentiment Analysis is flexible enough in how individuals choose to write linguistic rules to allow for a great deal of customization, as was done for this project. I began with a standard list of positive, negative, and neutral keywords. Then I added and removed certain words or phrases that take on new meaning in a political context. For example, the word "impeach" appears as a negative in most contexts. However, in politics it is neutral and refers to the act of removing an official from office.

In addition, words can have multiple meanings in a political context. "Endorse" can be relevant for detecting positive sentiment (someone endorses a candidate), but might be neutral in other situations (for example, the candidate endorsed a new policy). SAS® allows users to account for this context, as done here. Mentions of other people endorsing, supporting, or voting for one of the nominees are marked as positive Tweets about that candidate, but references to the candidate voting for something are marked as neutral, factual statements about the candidate's record.

With both the U.S, Republican primary and the television show *American Idol*, a similar linguistic predicament appears with factual statements about who won, media coverage asking who will win, and genuine excitement from supporters about a candidate winning. By using a feature called "Intermediate Entities" in SAS® Sentiment Analysis Studio, I am able to differentiate between the following sentence types:

- Who will win the Michigan primary? (neutral)
- Mitt Romney is the winner of the Michigan primary. (neutral)
- I am glad Mitt Romney won the Michigan primary. (positive)

In most cases, people assume "win" is a word with a positive connotation. However, when people are reporting on who won a contest there is no sentiment – they are simply stating the facts. In the last example, the word "glad" is the main word to detect positive sentiment about Mitt Romney. Even with the appearance of a positive word with "win," I picked up on negative sentiment about candidate victories. For example, if a person says "I am not happy Mitt Romney won the Michigan primary," then they are "not happy" about the result, which implies that they do not support Mitt Romney.

One of the most difficult words to deal with in both the U.S. Republican primary example and *American Idol* is "bomb." If "bomb" is used as a verb, it can refer to blowing something up or doing a poor job on stage. For the GOP candidates, there were many references to one candidate "bombing" in a debate, which is obviously negative. The problem is that in these debates many of the candidates discussed possibly "bombing" a foreign country. To get around this, I constructed a rule that looked for any conjugation of the word "bomb" in close proximity to a country name. Also, if "bomb" is used as a noun, it could be referring to "the bomb," meaning an atomic bomb. In modern slang for a person or thing to be "the bomb" that the person or thing is in some way good. In the GOP case, there are few instances of "bomb" referring to the slang term for cool, but in the *American Idol* case, it abounds. For *American Idol,* I wrote a rule that if an article preceded the word "bomb," it was a positive but if the word appeared without an article, it was a negative. For the GOP, few people are referring to any of these candidates as "the bomb" but instead are referencing defense policy in some way.

In political reporting, stories of horse race elections and momentum shifts are common campaign narratives. Traditional momentum measures have included horse-race polling and pundit proclamations. For the past two Presidential cycles, gauging momentum has focused on social media measurements. Many news outlets have and are using the actual volume of Tweets to gauge momentum, though the volume says nothing about the positive or negative feelings of the public. Individuals that use sentiment analysis tools on social media data about candidates usually do not do the best job at estimating the time series models necessary to do proper forecasts.

SAS® Forecast Server offers a way to estimate a variety of models and work with the most accurate to properly measure the time series for momentum and forecast the final outcomes. Here I use SAS® Forecast Server to help approximate who has momentum and who does not based on the aggregate sentiment results per candidate. To get the aggregate sentiment on each candidate, I score each Tweet about a candidate as positive or negative. Then I take the daily average sentiment for each candidate as the total number of positive Tweets that day minus the total number of negative Tweets. The aggregate daily sentiment score is then put into SAS® Forecast Server as the input variable. I include variables for if a candidate won a primary or caucus that day and the previous day as well.

CANDIDATE SENTIMENT

Most of the candidates are discussed in relatively neutral terms or are equally positive and negative. As can be seen in Figure 1, with the exceptions of Ron Paul, most candidates have a relatively close number of positive and negative Tweets about them. Paul has almost twice as many positive Tweets about him relative to negative Tweets. Given Paul has not won a single contest, this should be surprising. However, two things are worth noting about Ron Paul and the average Twitter user. The average Twitter user is going to be younger and therefore less likely to be conservative. Ron Paul is more libertarian than conservative, which has some appeal to young people. Ron Paul's supporters have also set up informal infrastructures to raise money and maintain his organization since 2008 primarily through social media outlets. Seeing Ron Paul as more favorable than the rest is not surprising given Twitter's demographics.

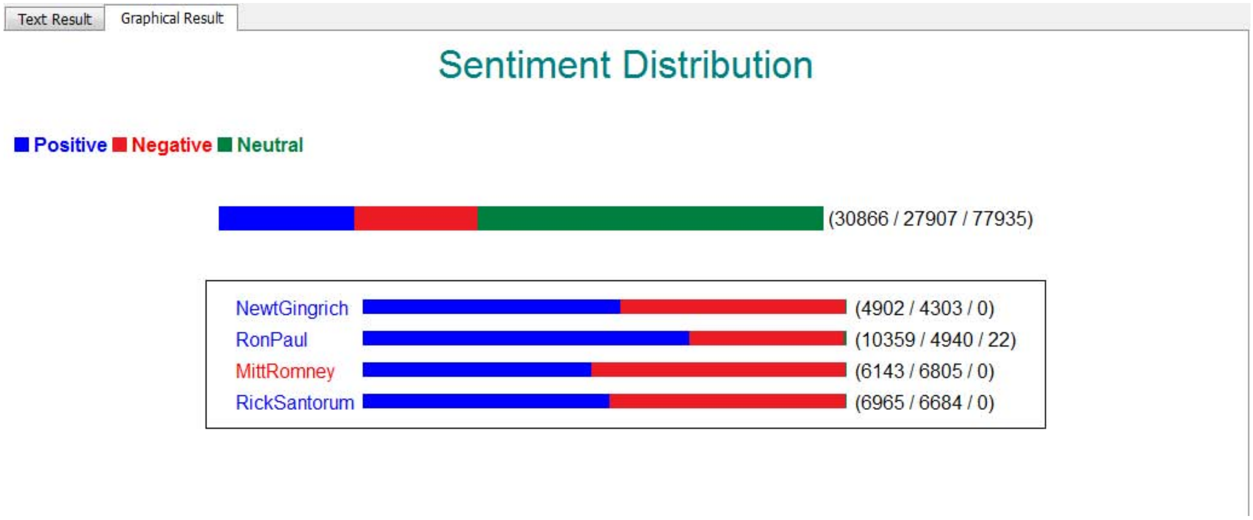

Figure 1. Candidate Sentiment

As mentioned, one of the features of SAS® Sentiment Analysis is the ability to drill into features around different candidates within the same document. For example, take the following Tweet: #Santorum is #Unelectable. #RonPaul2012. Here Santorum is identified as negative ("unelectable") while Paul is identified as being positive via the supporter hashtag "#RonPaul2012." In order to do this, all one needs to do is look for positive or negative words that could be used to describe a candidate within close proximity of an identifier for that candidate.

MOMENTUM

If the election were based on Twitter users, Ron Paul would win the GOP primary in a landslide. Throughout the primary season, he is the only candidate to consistently have positive sentiment, as seen in Figure 2. While the time series and forecast might look a bit odd, it is worth explaining that the high points in it are related to states with caucuses. Ron Paul has an explicit strategy to work hardest in caucus states to get delegates for the convention.

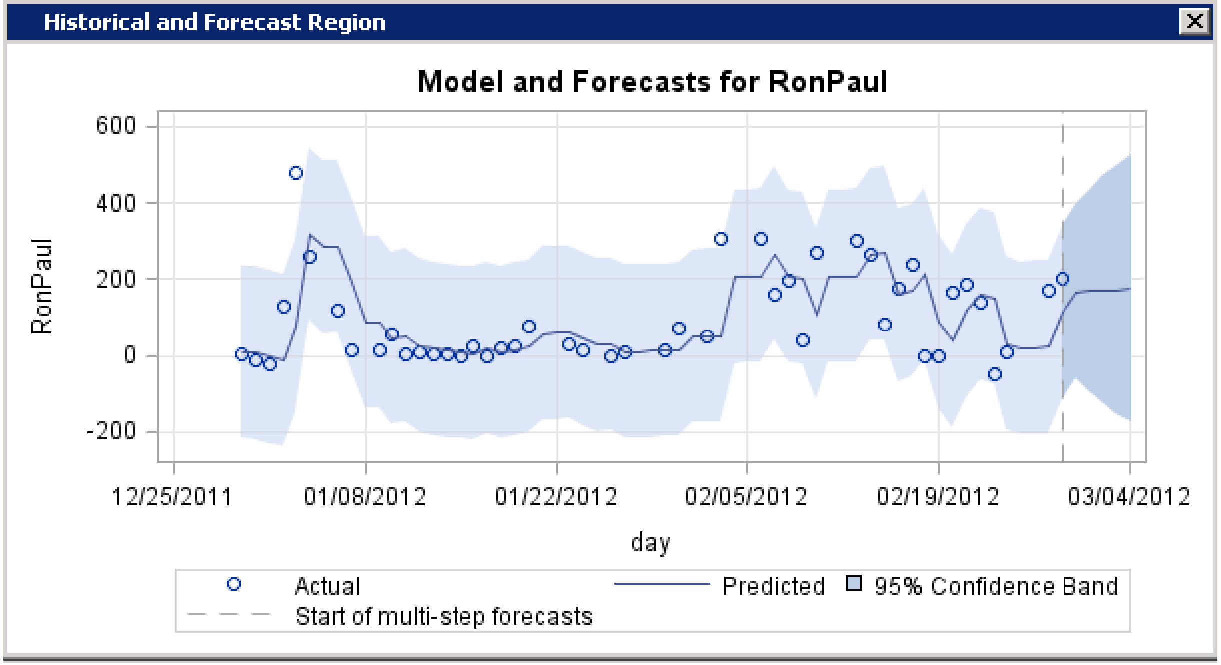

Figure 2. Ron Paul's Sentiment Score over Time

Of the remaining contenders, there is no clear indication of who will do well. In part, this is because GOP supporters are less likely to be posting their views through social media outlets. Another part of it deals with how volatile this campaign has been. By looking at some of the Tweets individually, it is possible to see that Mitt Romney has a giant problem connecting with the average voter. In SAS® Text Miner, I explored the data a bit and discovered that some of the keywords associated with Romney related to him being “robotic." There is also the problem with Rick Santorum, who in the latter part of February became the main contender against Mitt Romney. As seen in Figure 3 though, there have been some problems with Santorum coming to the forefront.

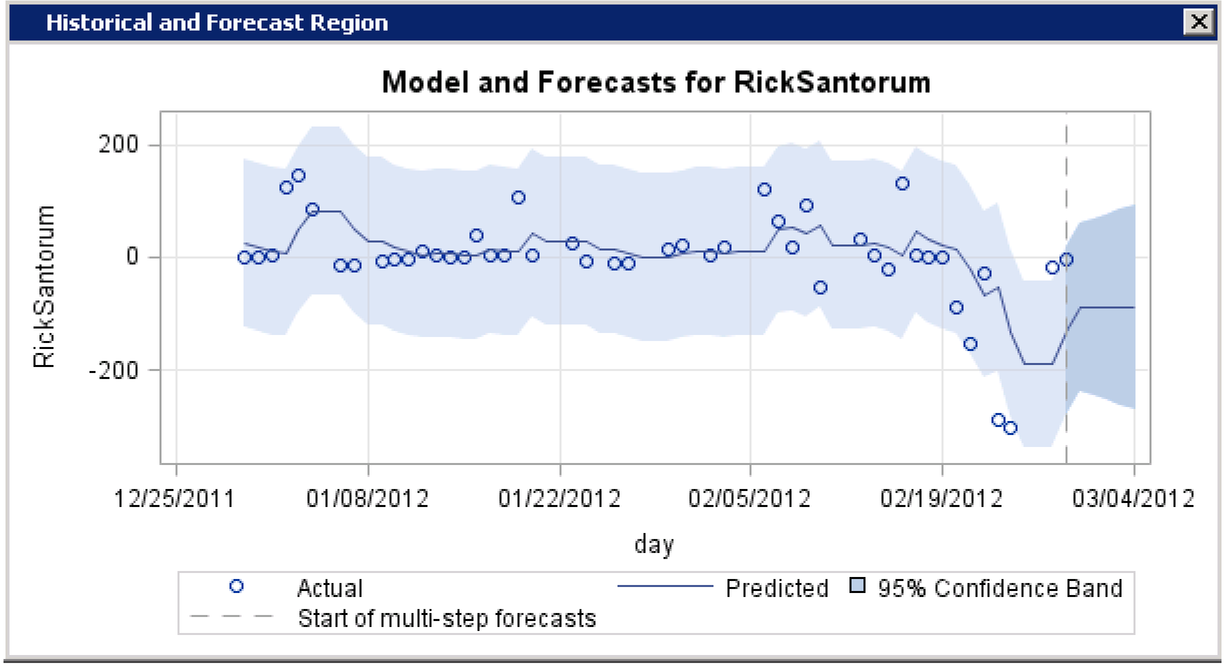

Figure 3. Rick Santorum's Sentiment Score over Time

For Santorum, there have been spikes in the time series around his victories, but a certain downward trend starting on February 20 is visible. Around this time, Santorum came under greater scrutiny from the mainstream media. What you can see in Figure 3 is that Santorum has begun to develop a clear downward trend as more information comes to light about his views on social issues. Also of note is that Santorum's camp has started back-peddling on these views; while the moderation has helped counteract some of the negative sentiment around Santorum it has not done enough to close the gap.

The question remains about who will be the GOP nominee. While we cannot say with any certainty about who it will be from analyzing social media about the candidates, we can draw some conclusions about the value of social media in predicting who it will be. First, the sentiment around Ron Paul is positive and growing over time. Ron Paul's support spikes around caucus time. This is consistent with the candidate's delegate-grabbing strategy and implies that, by convention time, Ron Paul might be in a position to play "kingmaker" at a brokered convention. Though not shown here, I will also note that support for Mitt Romney in the social sphere is relatively flat. Third, Rick Santorum appears to be on his way out. Unless his strategy to moderate his social views works, which would be evidenced by a break in the downward trend, Santorum's days are numbered.

AMERICAN IDOL

BACKGROUND

In 2011, 13 contestants made it through the "Hollywood Round" of the popular television show *American Idol*. The television show, which is modeled on other talent contests from Europe, has the public vote for which singer will win a recording contract with a record label. At the end of each week, one contestant is usually voted off, and the last person standing wins the contract. The 2011 season of *American Idol* was won by a teenager named Scotty McCreery, who had been a frontrunner throughout the season. He competed in the final round against another teenager named Lauren Alaina. Other notable contestants from that season (who have been analyzed for this project) were Haley Reinhart, James Durbin, Jacob Lusk, Casey Abrams, and Pia Toscano. Except for Pia Toscano, the other contestants were eliminated in reverse order.

CAVEATS

Before beginning the analysis, it is worth noting a few facts about the data from *American Idol*. First, the show's demographics skew older than that of U.S. Twitter users. Twitter users in the U.S. are teenagers while American Idol's main viewers are in the 18-49 demographic. The sample used here, while capturing a good number of Tweets about the show, does not capture all of them. Also, people can vote multiple times for an *American Idol* contestant, essentially stuffing the ballot box. During this cycle, there was little evidence of coordinated efforts for one candidate to receive multiple votes. There are some restrictions placed on voting in *American Idol*: some of the online forms for voting limit a person to 50 votes per show, people can vote only within a two-hour window after the show, and there are certain fees associated with voting from mobile phones. However, relative to the GOP, this contest is more representative of the demographics of the voting public.

METHODS AND ANALYSIS

From the beginning of the voting rounds (March 1, 2011) until the day after the season ended (May 25, 2011), I pulled as many Tweets as I could by searching for the contestant's full names, American Idol-issued Twitter user names, hashtags involving the candidates' names, and the American Idol hashtag. From there I assessed the sentiment on each of the individual contestants by identifying them in SAS® Sentiment Analysis through the contestant's name, nicknames, and song choices. Then I compiled an aggregate sentiment score per day for each contestant by subtracting the number of negative Tweets about a contestant from the number of positive Tweets. I used the average sentiment score as my time series variable in SAS® Forecast Server.

There are a few notes on the sentiment and forecast projects that anyone using social media should be aware of. First, context matters. As mentioned earlier, in the *American Idol* contest "the bomb" is positive. In this example, saying a contestant "killed it" is not a negative thing. It means that the contestant did a great job on a song. This phrase appears a few times when describing a performance by Haley Reinhart of "House of the Rising Sun." Secondly, I made extensive use of slang words and abbreviations to detect sentiment on the Tweets. Third, most of the grammatical rules in the SAS Natural Language Processing approach to sentiment estimation cannot be used here.

CONTESTANT SENTIMENT

Some contestants were mentioned much more than others. For example, Scotty McCreery was one of the most popular people on Twitter during the 2011 television series. Not only did McCreery have the highest number of

Tweets about him, the sentiment of those Tweets was almost four more times positive than negative, as seen in Figure 4.

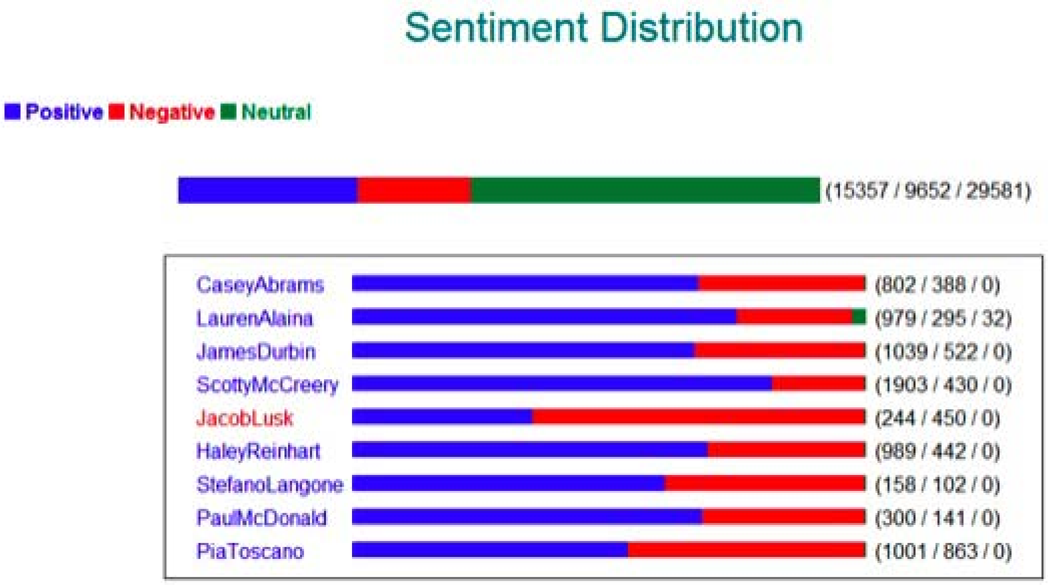

Figure 4. Overall Contestant Sentiment

Other contestants were mostly positive as well, though with a lower volume of mentions on Twitter.

MOMENTUM AND ANOMALIES

Most of the contestants on *American Idol* followed a stationary time series model, or a flat one with no trends or momentum. However, there were two exceptions. One was Scotty McCreery, who trended the entire time of the series. This is evident in Figure 5.

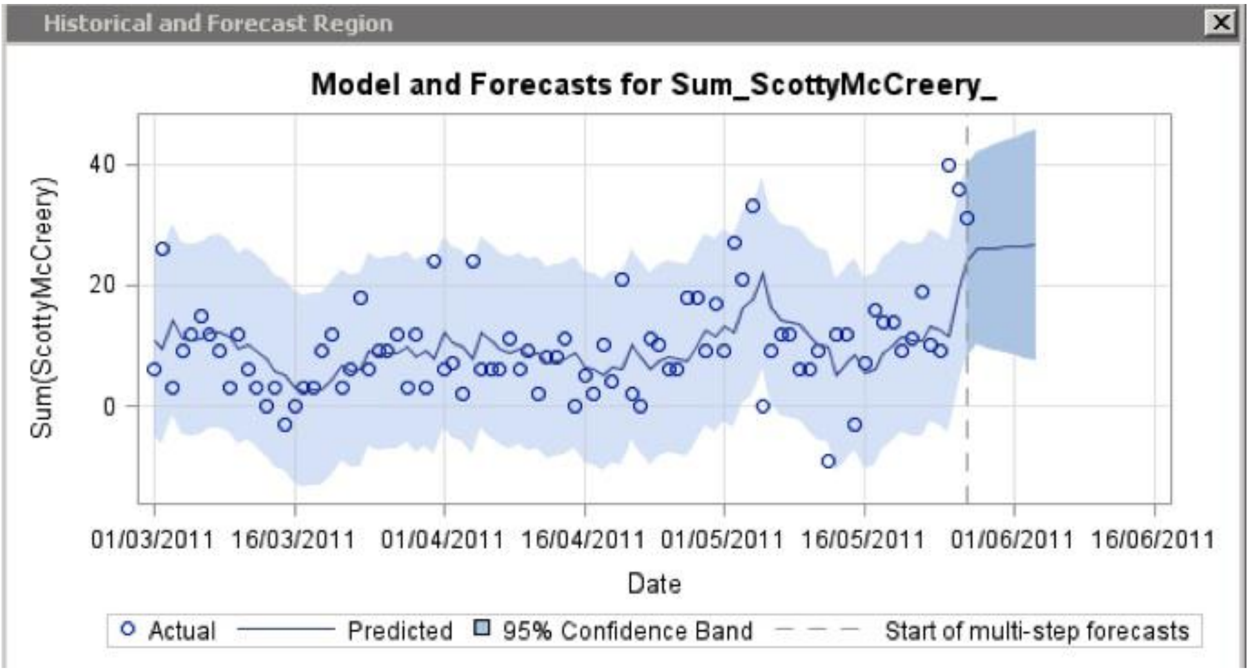

Figure 5. Scotty McCreery Forecast

The other is James Durbin. James did not appear in the “bottom three" on the show at any point before his elimination. He, like Scotty, was trending upwards with regard to the public sentiment. However, he placed fourth on the show to the surprise of many viewers. The residuals show that his series was unusually difficult to fit in SAS® Forecast Server, as shown in Figure 6. One of the big controversies early on was the elimination of Pia Toscano from the show. Both Pia and James have data points well outside the expected range from the nights they were eliminated. James in particular had a very well received performance the week he was eliminated. It was followed up with extreme backlash about his elimination.

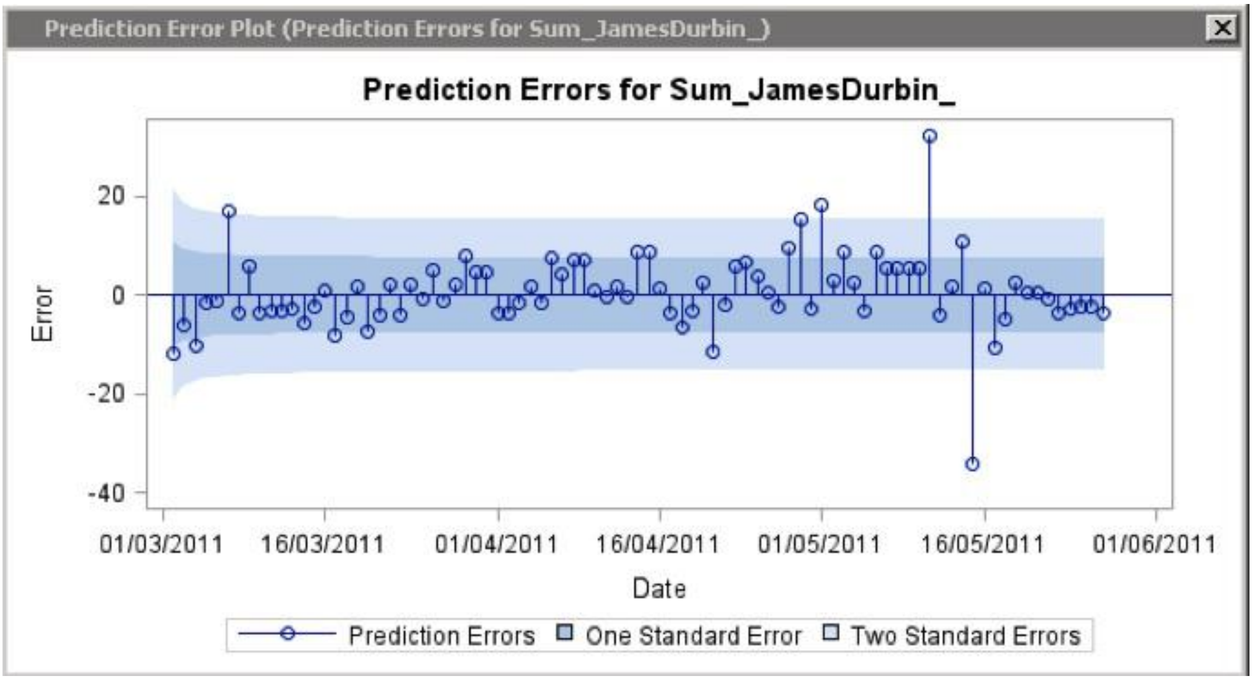

Figure 6. James Durbin Forecast Errors

EXTENSIONS

QUALITY CONTROL

Recently a large international company asked people to post their favorite memories about products from the company. Much to the company’s chagrin people began to Tweet about horrible experiences with the company’s products. Given the integration of social media and everyday life, many people will go online and post if they have a positive or negative experience with a product, as they obviously did in this example. Companies can use social media to help determine if there is a possible problem that might be emerging with their products. If a company were to monitor the daily sentiment around its products and see a trend develop, the company would have an early warning and be able to recall a product before having legal problems or forced recalls. This could save money on not only liability and lost stock, but it could also help the company save face with the public for noticing and correcting the problem early on.

NEW PRODUCT FORECASTING

Traditionally businesses have relied upon focus groups or test markets to determine how a new product will perform. Only a few people were allowed to see the prototype, and the opinions of these few people drove product design and deployment. Now with a combination of SAS® Sentiment Analysis and SAS® Forecast Server, companies can monitor the pre-release chatter about their current products and prototypes to help determine where improvements might need to be made, what the product life cycle might be, and what might be expected after the product is released. SAS® Sentiment Analysis can assist companies in seeing specifically what features current products are lacking and what features of a new product are already generating buzz among the masses. By combining that information with SAS® Forecast Server, these same business users can get a better idea of what consumer demand will be for the new product or service based on the social chatter.

COMPETITOR MONITORING

Keeping abreast of the competition is theoretically easier now with social media data. Companies monitor themselves to see how actual and potential consumers are responding to their products and services, but they can also look at what those same people are saying about their competitors. Instead of relying on expert opinions, insider information, and trade publications, companies can use social media to the same effect. Think about mobile phones. Few companies manufacture mobile phones anymore, and there is a great deal of talk on the Internet about new releases of new smartphones. A company can look not only at themselves, as described in the previous subsection, but also at their main competitors. The company can get some idea of what features their competitors have that they do not that are popular with consumers. They can also see if there is a competitor that might be on their way toward bankruptcy because of a feature on their phones that people do not like.

RETAIL DEMAND

Far too many winter holiday specials in the U. S. involve some parent desperately trying to find the dream toy for his or her offspring at the last minute. Inevitably, retailers have had a run on the dream item for that year. Unfortunately, there is some truth to the scenario that Hollywood has dreamed up. Gauging the buzz around what the next big toy will be relies on experts watching sales trends over time. With the advent of social media, retailers can first look at which items are becoming more popular based on what people are saying online. If people are excited about the new My Little Pony, they might start Tweeting about it. Retailers can use this information to help forecast the demand for these items, which could reduce the likelihood of sell outs and runs on the most popular items. The additional information from social media sources can augment what the typical purchasers for stores are looking at by extending the sample of discussants on the products. In turn, retailers can stock up their stores with the right number of items and spend more time focusing on where to place other items to increase demand for complimentary goods. For example, if people are buying lots of Flutteryshy ponies, retailers might be able to also sell those customers the Rainbow Dash pony.

BOX OFFICE SUCCESS

Since 2010, various sources have been using Twitter to help predict how well movies will do during their theatrical runs. By 2010, people in Hollywood were able to see the value added from having information about Twitter to forecast how well a movie will do. In this capacity, social media plays a dual role for studios: the first is to gauge hype before the release, and the second is to foster further hype through informal word-of-mouth campaigns. New York magazine even has a nickname for mixing the sentiment from social media with makeshift forecasts to predict what will be the “next big thing" out of the film industry.

ACKNOWLEDGMENTS

The author is grateful to Mary Osborne, Kathy Lange, and Bob Messier of the Americas Technology Practice at SAS Institute Inc. for their valuable assistance in the preparation of this manuscript.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author:

Jenn Sykes
SAS Campus Drive
SAS Institute Inc.
Jenn.Sykes@sas.com