Group Members: Branden Diniz, Xuanyu Li, Dennis Silva, Chuxin Wei

**DS 501: Case Study 1 Report**

Due to this case study focusing on data gathering, our group wanted to gather twitter data when tweets would be in high volume. Thus we decided to focus our data gathering during Super Bowl 50 which occurred Sunday February 7th. This year was the 50th Super Bowl, and since last year’s Super Bowl was said to have had over 28.4 million tweets during its broadcast according to CNBC, we believed there would be an abundance of available data [1]. Since most of our group were not huge football fans, we decided to focus our attention on the advertisements during the game. These advertisements are known to be some of the most watched commercials during the year, and as such several companies attempt to make advertisements that increase brand awareness and advertise new products. In order to avoid rate limiting on twitter’s Streaming API, we decided to split our data collection into two sections. The first would use the Streaming API to gather data on a specific advertisement during the Super Bowl, and the second would utilize twitter’s REST API to gather tweets about a variety of advertisements as well as tweets directly about the football game itself. In this paper, the specific advertisement will be discussed in Section 1 and our overall results from all advertisements will follow in Section 2.

**Section 1.1: Background**

Companies commonly will release their advertisements online weeks before the actual Super Bowl airs both in order to foster excitement as well as to reach individuals that don’t normally watch the Super Bowl. We found that according to The Wall Street Journal, the leading pre-release Super Bowl ad was the Pokemon Company’s ‘Pokemon 20’ ad with about 14.9 million views on YouTube as of Sunday midnight [2]. For reference, Pokemon is a multi-media franchise that is owned and developed by Nintendo. The franchise exists as several video games, a trading card game, and television shows involving creatures called Pokemon that are collected and trained by Pokemon Trainers to battle one another. The demographics for Pokemon have originally been around the ages of 7-12, but have extended well into the 20-30 year old demographics as well. The ‘Pokemon 20’ ad is a sentiment to Pokemon’s 20th anniversary occurring this year, encouraging people young and old to “train on” in whatever he or she is passionate about. Some of our group members grew up playing the original Pokemon games, and we believed that looking into Pokemon and its connections to the Super Bowl would be an interesting investigation.

**Section 1.2: Methodology**

Our group decided to gather data during the Super Bowl as we could plan in advance what, how, and when tweets could be collected. Super Bowl 50 began at 6:30 pm EST, but we started to collect data from 6:00pm EST. Pokemon is a commonly discussed topic on twitter, and we wanted to only collect tweets about Pokemon or the Pokemon Ad directly related to the Super Bowl. Using twitters operators in the query search we were able to, mostly, gather collected relevant tweets. We set our query search to '(pokemon OR #Pokemon20 OR #TrainOn) (#SB50 OR #SuperBowl OR super bowl OR commercial OR ad)' which we then converted to Unicode. Through this specific query search we only collected tweets that had some key term related to Pokemon as well as the Super Bowl. To avoid rate limiting, every 15 or so we manually ran our search to collect a maximum of 1,000 tweets. We rarely hit this limit and with each search we received an average of 800 new tweets. We continued this until 11:00 pm EST after the games conclusion. In total we collected 13,164 tweets. All tweets were saved a json in text files. We originally wanted to directly utilize twitter’s Streaming API, however we realized we had utilized twitter’s REST API when acquiring tweets in real time. Thus the tweets collected were in fact from the desired time period, but additional processing and storage was done on them first by twitter and then accessed by our group. Ultimately, the data should have been the same collected using either API.

**Section 1.3: Results**

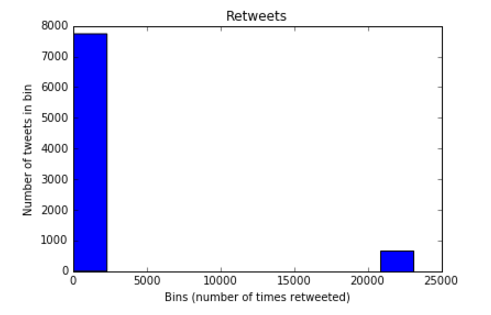
In total we collected 13,164 tweets relating to Pokemon tweeted during the Super Bowl. We calculated the top 30 words found in the tweets. We attempted to remove all of the stop words stored in the Natural Language Toolkit (nltk) package, various puncuations, and some common words used in tweets such as “rt” (re-tweet). For some reason, some stop words were unable to be removed directly by our code. In addition, some common words had punctuation attached to the word. For example, “@Pokemon” and “@Pokemon:” were two words that we deemed to be the same. The top 30 words with additional stop words removed and the same words consolidated are shown below.

|  |  |
| --- | --- |
| Top 10 Words | |
| Word | Count |
| super | 7837 |
| bowl | 6874 |
| Pokemon | 6014 |
| #SB50 | 5429 |
| @Pokemon | 4366 |
| #Pokemon20 | 4322 |
| commercial | 3930 |
| ad | 2045 |
| Spot | 1531 |
| #TrainOn | 1077 |

Due to the nature of our search queries, we expected several of the top words to be one of the required search terms. We see “super” occurring more than “bowl,” which odd assuming they should be together forming “super bowl.” The adjective “super” may have been used to describe the Pokemon advertisement itself. The only word not occurring in our required search terms, besides users or hashtags, is “spot.” Its position in the lineup was at first confusing, but we made sense of it when most popular tweets were analyzed. In addition, we see that the user @Pokemon is abundant. This is the official Pokemon twitter account, and its popularity is made clear when looking at the most popular tweets. The most popular tweets were described as those with the largest counts of retweets. The top 10 most popular tweets are shown below.

|  |  |  |
| --- | --- | --- |
| Top 10 Most Popular Tweets by Retweet | | |
| Count | **Screen Name** | **Text** |
| 23133 | Pokemon | RT @ Pokemon: Get a first look at our #Pokemon20 Super Bowl spot and tell us how you plan to #TrainOn in 2016! <https://t.co/aUIST5Hzob> |
| 1702 | \_Snape\_ | RT @\_Snape\_: Did Lady Gaga just win the Super Bowl #SB50 |
| 453 | F\_zamora3 | RT @f\_zamora3: Now we wait for the @Pokemon commercial during #SB50 <https://t.co/odfyDf10TV> |
| 362 | Kate\_Eglen | RT @Kate\_Eglen: Pikachu cupcakes to celebrate Pokemon in the Super Bowl !! #Pokemon20 |
| 340 | FunnyPokemon | RT @FunnyPokemon: Retweet if you’re excited for the new Pokemon Super Bowl commercial! <https://t.co/doyJSA4bf4> |
| 311 | Rudymancuso | RT @rudymancuso: If you drink, walk. Don’t drive. #GiveADamn #SB50 @Budweiser #ad <https://t.co/UCyYAlL6Fd> |
| 281 | TheFakeESPN | RT @TheFakeESPN: Nothing says Super Bowl like Pokemon |
| 222 | EW | RT @EW: Celebrate 20 years of #Pokemon with this epic #SB50 commercial: https:t.co/g2tS0o8TKB |
| 205 | USATODAY | RT @USATODAY: Pokemon turns 20 this year: Watch an extended cut of the #SB50 commercial #admeter #pokemon20 <https://t.co/kQHOAXcmJy> |
| 193 | AshKetchum151 | RT @AshKetchum151: Who’s waiting for the Pokemon commercial? #SB50 |

The most common retweet comes from the official Pokemon screen name @Pokemon, telling tweeters to express how they will train on this year. It should be said that our group captured several of the same retweets originally by @Pokemon, and we decided to remove repeats from the table above. We created a histogram of tweet counts by the number of retweets they had. We see the 23,133 retweets of @Pokemon’s tweet are captured on the far right. From the total 13,164 tweets collected during the Super Bowl, about 800 were @Pokemon’s retweet.



|  |  |
| --- | --- |
| Top 10 User Mentions | |
| User Mentions | **Count** |
| Pokemon | 4743 |
| FunnyPokemon | 746 |
| F\_zamora3 | 417 |
| Kate\_eglen | 333 |
| AshKetchum151 | 325 |
| TheFakeESPN | 244 |
| YouTube | 223 |
| EW | 213 |
| TheJWittz | 189 |
| USATODAY | 183 |

Looking back at the table of retweets, we see several of these retweets have links within them, as is common. Most of these URLs link to a version of the “Pokemon 20” ad. Kate\_Eglen links to a picture of homemade cupcakes in the shape of the Pokemon mascot Pikachu. USATODAY, the official screen name for the news network USA Today, joins in on the Pokemon celebration tweeting about an extended version of the “Pokemon 20” ad lasting 1:10 minutes rather than the 30 second ad shown during the Super Bowl. Our group then looked at the most popular tweet hashtags and user mentions in our collection. The tables for the top 10 of these tweet entities are shown to the right and below.

|  |  |
| --- | --- |
| Top 10 Hashtags | |
| Hashtag | **Count** |
| SB50 | 5555 |
| Pokemon20 | 4811 |
| TrainOn | 1095 |
| SuperBowl | 617 |
| Pokemon20 | 409 |
| Pokemon | 268 |
| Pokémon | 214 |
| admeter | 198 |
| Pokémon20 | 176 |
| collegelife | 174 |

We see that most users tweeting about Pokemon during the Super Bowl used the official hashtag #SB50 more than the official Pokemon 20th anniversary hashtag #Pokemon20. It is interesting to see that only 1095 tweets contained #TrainOn, as this was the specific hashtag that makes an appearance in the Pokemon ad itself and is contained in the most retweeted tweet from @Pokemon. One can see there are two different spellings of Pokemon or Pokémon. The official spelling is Pokémon, yet most individuals drop the accent. Our group chose not to group these hashtags together like we did with words as hashtags themselves link other tweets using the same hashtag. If an individual clicked on #Pokemon rather than #Pokémon, they would be brought to a different pool of tweets. The occurrence of #admeter refers to USA Today’s Ad Meter website where users can use this hashtag to vote for their favorite ad. The ranking of the ads are then tabulated and displayed in [3]. It is also interesting to see #collegelife reflecting the popularity of Pokemon with the college community. Referring back to the most popular retweets and user mentions, the group looked at the friends and followers of @Pokemon. The results are below.

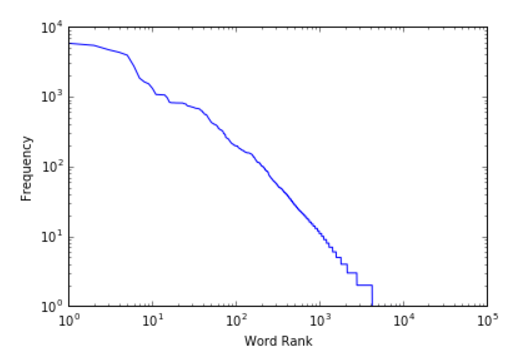
|  |  |
| --- | --- |
| 20 Example Followers of @Pokemon | |
| ID Numbers | Screen Names | |
| 156008448 | Rynmac14 | |
| 2748289025 | Orioxified | |
| 579282946 | Josh3258 | |
| 4798451716 | jokool\_aid | |
| 4884797445 | HontaroMeru | |
| 386150409 | Tuffin96 | |
| 1549543436 | elblanco490 | |
| 812126222 | DawsonAlbin | |
| 4203025752 | gaming\_hobbit1 | |
| 350580755 | Jaden\_sellers | |
| 3588616215 | OliverHaagensen | |
| 2949341210 | Porter\_madsen | |
| 996530203 | dickypermana111 | |
| 97610074 | mrJcruz88 | |
| 353294341 | Jesysariana | |
| 2448402464 | MattThill1 | |
| 4880732195 | LeviLo1 | |
| 3281508389 | Milutwo | |
| 594729000 | SilHoran\_ | |
| 4888467496 | xijAlpaca | |

|  |  |
| --- | --- |
| All Friends of @Pokemon | |
| ID Numbers | **Screen Names** |
| 2260759812 | PokemonFR |
| 2260755366 | PokemonIT |
| 2296634058 | PokemonDEU |
| 168754736 | pokemonworlds |
| 2260749054 | pokemon\_ES\_ESP |
| 385175199 | Junichi\_Masuda |

One can see that the twitter account @Pokemon only has six friends. Four of these friends PokemonFR, PokemonIT, PokemonDEU, and Pokemon\_ES\_ESP, are the official twitter accounts for France, Italy, Germany, and Spain respectively. We do not see any of these screen names in earlier analysis of popular users and retweets most likely due the American Super Bowl is not very popular in these areas. The user @pokemonworlds is the official Pokemon world championships twitter page. These world championships include both the Pokemon Company’s video and trading card game lines. @Junichi\_Masuda is the only twitter account owned by a single individual. Junichi Masuda was a major director, designer, programmer, and producer of the original Pokemon video games that began this major franchise. Some of the example followers of @Pokemon have video game related screen names such as gaming\_hobbit. User Milutwo seems to be a play on one of the popular Pokemon characters Mewtwo. Although @Pokemon has several thousand followers, since @Pokemon is only friends with its alternative language accounts and Mr. Masuda, the overlap between their sets of friends and followers is the empty set. We repeated the friends and followers analysis for several other popular user mentions, such as FunnyPokemon and F\_zamora3, but they all too had no overlap in friends and followers.

The team also did some analysis on the actual words, screen names, and hashtags found within our collection of tweets. The average word count per tweet was roughly 15 (14.9389). Although our group failed to find an accurate average statistic of word count across all tweets, 15 words seems to be fairly small when compared to the maximum 140 characters at a user’s disposal. The lexical diversity of the collected words, screen names, and hashtags were 7.46%, 8.08%, and 3.13% respectively. These relatively small lexical diversity scores could be attributed to both our restricted search queries as well as the high volume of the same retweets from @Pokemon. We also find that Zipf’s law fits relatively well with the vocabulary of the tweets. In the graph below, we see the largest perturbation around words with a rank of ten.

**Frequency by Word Rank in tweets collected  
 about Pokemon during Super Bowl 50**



**Section 2.1 Background**

For our data exploration we chose to investigate tweet volume during the Super Bowl. Tweets we were interested in had to involve either Super Bowl commercials or the teams involved in the game, the Carolina Panthers and the Denver Broncos. We chose this topic because Super Bowl commercials are notorious for being quirky and funny. The phenomenon has become such a craze over the past decade that people often claim to watch the game “only for the commercials.” We decided to put this to the test. By investigating keywords about the game, we looked for evidence that more people were interested in the commercials than the game itself.

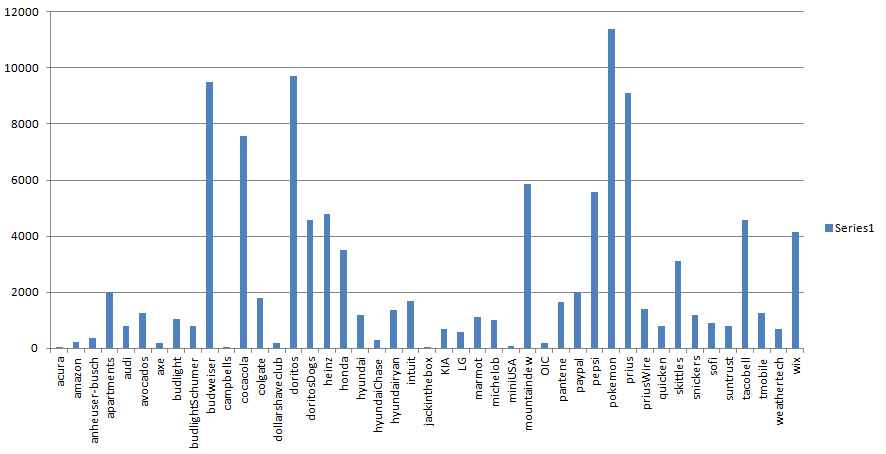
**Section 2.2 Methodology**

Instead of using Twitter’s Streaming API, we used the REST API to gather data. This benefited us in two ways. First, it allowed us to circumvent Twitter’s rate limiting during the narrow window of the Super Bowl to gather many more tweets. Second, it allowed us to gauge which commercials went viral on social media so we could do a more targeted analysis. A feature of the REST API, that is absent from the Streaming API, is being able to search for stored tweets given a time range by using the ID numbers of individual tweets. Twitter stores about 6-9 days’ worth of tweets to be accessed by the REST API. By finding the IDs of tweets posted at the start and end of the Super Bowl, 6pm and 11pm respectively, we guaranteed that all tweets gathered were posted during that time frame. Additionally, if a particular keyword had too many tweets and resulted in rate limiting, we could perform our searches over smaller time intervals to prevent this. We also used keywords and hashtags when we gathered data from the API. If we were looking for commercials, we would use the format '(company OR keyword) (#SB50 OR #SuperBowl OR super bowl OR commercial OR ad)' for ensuring that all tweets we had were referring to Super Bowl ads. When searching for teams, we would use '([team name] OR [city]) (#SB50 OR #SuperBowl OR super bowl)' as the format, again making sure that all tweets were super bowl related. This format limited the tweets we gathered as only responses that mentioned the Super Bowl were collected. Something like “PuppyMonkeyBaby,” Mountain Dew’s ad that went viral during the game, produced many more hits as a standalone search than it did when combined with the Super Bowl related keywords. For the sake of consistency, we chose to take only results with direct Super Bowl references. This way, tweets about companies or commercials by people who were not watching the Super Bowl would not be reflected in our results. After looking into the companies who had ads in the Super Bowl, we compiled a list of companies whose ads had some sort of viral impact and collected tweet volumes for each. Using the approach described above, our results are described below

**Section 2.3 Results**.

Below is a graph of all of the results of our harvesting related to television commercials. We collected 111,021 tweets about ads in total.

**Company-related Ad Mentions in Tweets during Super Bowl**



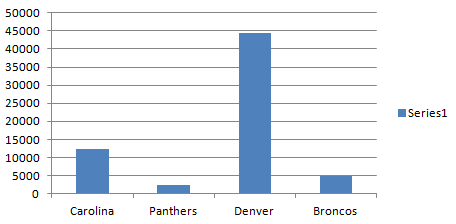
The top five most tweeted about commercials were the following:

|  |  |
| --- | --- |
| Top 5 Most Tweeted Commercials | |
| Commercial | **# of Tweets** |
| Nintendo – Pokemon | 11397 |
| Doritos – Ultrasound | 9695 |
| Budweiser - #GiveADamn | 9492 |
| Prius – The Longest Chase | 9091 |
| Coca Cola – Coke Mini (Hulk vs. Ant-Man) | 7588 |

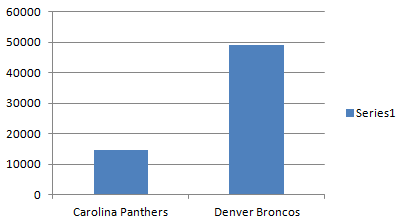
Even though the Pokémon commercial had the most tweets for any individual commercial, Doritos won the evening for the most tweets total. Its first commercial called “Ultrasound” made a huge social media splash, but it also had a second commercial titled “Doritos Dogs” that was a hit as well. Between the two, Doritos had 14275 tweets about its commercials.

Below are the graphs for results relating to the teams that played in the Super Bowl.

**Tweets Collected by Teams and Location**



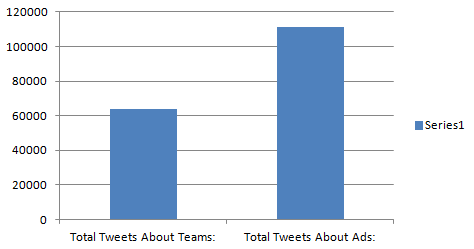
**Total Tweets Collected by Official Team Name**



The Broncos had significantly more tweets about them than the Panthers did with the vast majority of these tweets during the hour of 10pm to 11pm. This was the time around which people started to know the Broncos were the winners followed by the time that they actually won. There was such a high density during this hour that data collection had to be done at 15 minute intervals to prevent rate limiting. Worth noting as a point of interest is that most tweets came in the form of references to the teams’ respective home towns rather than to the team names.

In total, there were 49,193 tweets about the Broncos and 14,779 tweets about the Panthers. This leads to a total of 63,972 tweets between the two teams. In comparison, tweets about ads totaled 111,021 which are almost 72% greater than tweets about teams. This evidence would suggest that when people claim they’re only watching the Super Bowl “for the commercials” they’re telling the truth. At the very least, those people certainly prefer to talk about the commercials.

**Total Tweets Collected: Teams vs Ads**



**Citations**

1. Gibbs, Alexandra. “CNBC: Super Bowl XLIX smashes Twitter records.” <http://www.cnbc.com/2015/02/02/super-bowl-xlix-and-social-media-most-tweeted-nfl-game-ever.html>. (Accessed 2/5/16)
2. Tadena, Nathalie. “The Wall Street Journal: Pokemon’s Super Bowl Ad Leading Race for Pre-Game Views.”<http://www.wsj.com/articles/pokemons-super-bowl-ad-leading-race-for-pre-game-views-1454874453> (Accessed 2/7/16)
3. “USA Today: Ad Meter.” <http://admeter.usatoday.com/results/2016>. (Accessed 2/9/16)