**GitHub:** https://github.com/AntonyWalsh/S2CA2

**Twitter Data**

I was unable to get access to the twitter APIS’s so I then decided to use the data stored at <https://archive.org/details/twitterstream?sort=-publicdate>

The size of each month for 2022 was over 80 gigs so I decided to get the data for 2016 which was more manageable at about 40 gig per month.

In all I downloaded almost 500 gig of twitter data for 2016.

I did a google search for the ten most popular hashtags of 2016. There were a lot of political hashtags as the US presential election between Donald Trump and Hillary Clinton took place in 2016. Two other popular hashtags were the computer game Pokomon and the TV show Game Of Thrones. I am a big Game of Thrones fan and so decided to use that as my topic.

The twitter data at archive.org is the “Spritzer” version of twitter data grabs. The Spritzer stream is a sample of 1% of the tweets on a given day. The Sprinkler version is 10% of tweets and the Garden Hose version is 100% of tweets.

A paper by (Kergl, et al., 2014) investigated the Spritzer stream and found that sample was taken from the same 10 milliseconds from the twitter time stamp. The paper supported the randomness of the sample. However, (Pfeffer, et al., 2018) show how the sample is open to deliberate manipulation and that technical artifacts can skew the samples.

I find it unlikely that twitter data on a TV show is manipulated so will take the data at face value.

**JSON Parser**

To parse the data, I used a parser developed by Baylor University (Been, 2018) under MIT licence. I fixed a bug with the parser where it was opening the files as a bytes object and then comparing that to a string. I also added some logging.

This parser will search a directory of bz2 compressed files for a hashtag and save any match in a csv file.

**Development Process**

The python development was done in an iterative fashion and different parts of the product were developed in parallel. There were three main strands:

1. Batch processing and getting the data prepared
2. Spark environment, MongoDB and YCSB
3. Modeling (Sentiment analysis and Time series)

The process followed was similar to CRISP DM

To get started I downloaded and parsed 3 months of data and used that to do the initial development on sentiment analysis and time series.



Figure 1: CRISP DM

**Batch Process**

There was a batch processing element to this project.

I processed the data in monthly batches. To download 10 gig of twitter data took me about 1 hour, sometimes this took a lot longer. To parse 10 gig of twitter data also took me about 1 hour.

Once the monthly twitter data was processed, I then deleted the raw twitter data to save disk space.

**Spark and MongoDB**

The resulting csv files from the JSON parser contained a lot of false positives. The parser matches a string for any part of the tweet. I only want a match where GameOfThrones (any case) matches the actual hashtag. As part of the batch process, I used Spark to connect to MongoDB and then read a monthly csv file.

I then filtered out the false positives using Spark SQL from the DataFrame and then appended to a MongoDB database. For this I converted the tweet to lowercase and matched on that. A match on #GameOfThrones with any combination of upper/lower case is correct.

**Sentiment Analysis**

The dataset is non labelled so I decided to use a Lexicon based sentiment analysis approach. I used the Valence Aware Dictionary and sEntiment Reasoner (VADER) sentiment analysis tool for this project. VADER has a number of advantages it is a lexicon that is tuned for and works well on social media posts as shown by (Hutto & Gilbert, 2014). VADER is simple to use as it automatically takes care of stop words and other edge cases. VADER also has the advantage that it can analyse jargon, emojis and excessive punctuation and it can identify more nuanced forms of sentiment, such as sarcasm and irony, that may be missed by other sentiment analysis tools. A study by (Ali, et al., 2022) showed a sentimental analysis on twitter data using a VADER approach was successful in showing insight into the 2020 US presidential election.

The output of the sentimental analysis was a new column was added to the DataFrame called “sentiment”. For every tweet in the DataFrame a new value was inserted containing a value of ‘positive’, ‘negative’ or ‘neutral’. These values are the actual sentiment for each tweet.

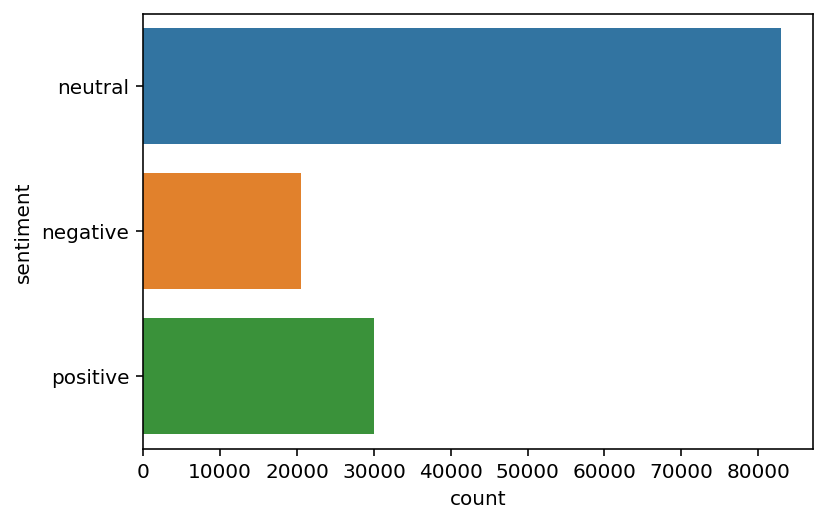


Figure 2: Sentiment Bar Chart

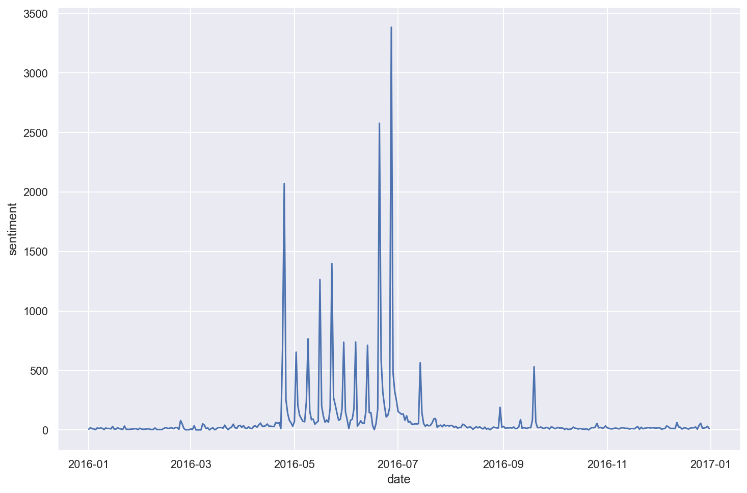


Figure 3: Positive Sentiment for 2016

**Missing Data**

I did a check for missing dates and found that there were 9 missing days. As the input to the time series analysis is the output of the sentimental analysis, I completed the sentimental analysis before investigating missing data.

During 2016 Game of Thrones aired season 6 which consisted of 10 episodes. These were initially aired Sunday evening in the US and then usually Monday evening in the rest of the world. I personally watched the episodes on Monday evening. The airing dates for 2016 were:

April 24th, May 1st, May 8th, May 15th, May 22nd, May 29th, June 5th, June 12th, June 19th and June 26th.

It is no coincidence that the twitter traffic peaked around those dates.

The 6 dates with missing data are:

'2016-02-11', '2016-02-12', '2016-02-13', '2016-02-28',

'2016-02-29', '2016-03-05', '2016-03-06', '2016-03-07',

'2016-06-03', '2016-06-04', '2016-06-05', '2016-06-06'

I decided to approach the missing data using two different strategies. One strategy for dates around the airing dates and one strategy for all other dates. This was to keep the pattern of the data as best as possible.

I did a count ‘groupby’ on the date to get the positive and negative sentimental per day creating two new DataFrames. This was the granularity I would use going forward. I added the missing dates into the DataFrames.

For all missing data dates except for the June missing dates, I filled in the positive and negative values with the value from the previous row.

For the June missing dates to keep the pattern of the data I back filled from the day of the previous week.

**Time Series**

I experimented with two different time series models the Autoregression model and the Autoregression Integrated Moving Averages model.

**Autoregression model**

With Time Series analysis one important concept is that the measurement of a value at a time period depends on the measurement of that value at the previous time period, time period before that and so on and on. The number of positive tweets in the past can affect the number of tweets today. The partial autocorrelation function shows the correlation directly from one time period to another.

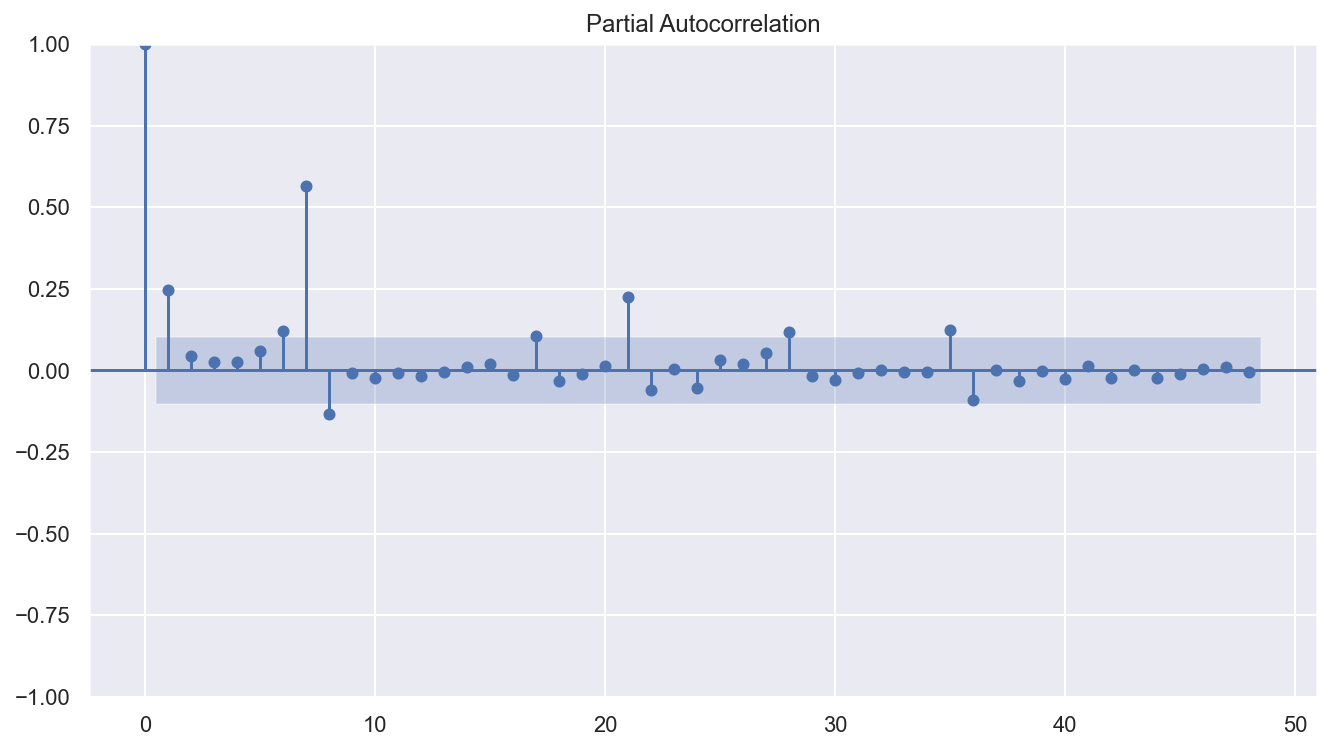


Figure : PARTIAL AUTOCORRELATION POSITIVE SENTIMENT 48 LAGS

The Partial Autocorrelation plot above show a statically significant results for lags 1, 7, 8 and 21. The rest of the lags are insignificant or barely significant.

An Autoregressive model is a Machine Learning model which attempts to predict future values based on previous values. The Autoregressive model works on the direct effects or direct correlation of previous time values on the time value today.

I experiment in predicting positive tweets one month ahead with different

**Experiment**

**Emmys**