OMIS 670: Social Media Analytics for Business

**Social Media Analysis on  Black Friday data 2018**

By

Venkat Phani Saroj Bodepudi

Siva Kiran, Thatikonda

Date of Submission: **12/09/2018**

**TABLE OF CONTENTS**

1. Executive Summary…………………………………………………………...…………. 3
2. Introduction………………………………………………………………………......……4
3. Data Extraction…………………………………………………………………….……...5
4. Exploratory Data Analysis………………………………………………………………...5
5. Sentiment Analysis……………………………………………………………………….6
6. Visualization………………………………………………………………………...........7
7. Word Cloud……………………………………………………………………………….8
8. Topic Modeling……………………………………………………………………………9
9. Inferences………………………………………………………………………………...11

**Executive Summary**

This project is about social media analysis on Target Black Friday data – 2018. Target is one of the largest retailers in USA. As a matter of fact, 45 to 55% of its annual sales happen during the holiday shopping and Black Friday being the most prominent day. As a team we thought to perform social media analysis on Target Black Friday deals, especially on the day of black Friday 2018. Target will unveil special deals for every Black Friday. People will post their opinions and shares posts on social media and talk a lot about the offers. They tend to compare the deals with its competitors and explain about their shopping experience during this season. So, we thought it would be a good project to do social media analysis on people’s opinions and draw key inferences and submit the report to the Target executive team. This could help in designing the future deals and improving any kind of discomforts customers face while shopping during the holiday season at target.

**Introduction**

* **Languages and Tools used:**

Before start of the project we thought to only extract live streaming data rather than taking archival data. Twitter is one of the famous social media platforms which provides API’s to users so, we have chosen twitter as our platform to extract data. Python is one of the powerful languages which provides libraries to extract data from the twitter. So, we used python (Tweepy library) to extract data from the twitter. We have extracted all the fields which we thought are necessary to perform analysis on this data. As we know the data from the social media platforms is not organized data, we stored them in Excel files. Cleaning disorganized data is a huge task for any analyst. So, we have concentrated only on cleaning text data and using it for the analysis.

* **Visualizations:**

In this project we used Microsoft Power BI (Intelligence tool) to perform visualization on data which we obtained from the twitter. Power BI has capabilities to visualize the data from the spreadsheets. We can also create new columns and transform the existing columns using BI.

* **Techniques followed:**

In this project we performed Sentiment Analysis, Word Cloud and Topic Modeling to draw inferences from the raw twitter data. We have considered using latent Dirichlet allocation (LDA) algorithm for these techniques. In [natural language processing](https://en.wikipedia.org/wiki/Natural_language_processing), latent Dirichlet allocation (LDA) is a [generative statistical model](https://en.wikipedia.org/wiki/Generative_model) that allows sets of observations to be explained by [unobserved](https://en.wikipedia.org/wiki/Latent_variable) groups that explain why some parts of the data are similar.

**Data Extraction**

API’s are the gateways to have permission to scrape the data. Twitter provides API’s keys for the users, if they have development account on twitter platform. To extract data from the twitter we used API’s keys which twitter has provided. We imported Tweepy library in to our python code and used API’s keys to extract live streaming data from the twitter platform. We gave **#Target** as the key word to extract live streaming tweets so, the tweets which has #target in the text of the tweet are extracted.

Each tweet is having nearly thirty attributes. For this project we concentrated only on created\_at and text attributes of the tweet. We ran the code for 7 hours (8.30 pm - 3.30 am) on the day of black friday to collect the tweets. We have saved all the tweets in excel file in our local computer.

**Exploratory Data Analysis**

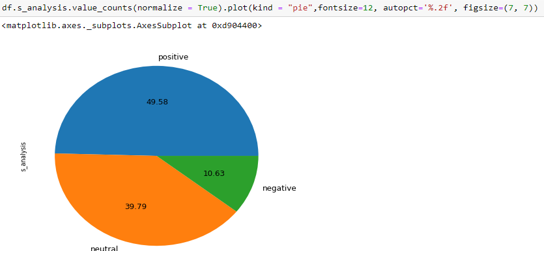
Python is known for its powerful libraries. One of the libraries in python which is used extensively for data analysis is Pandas. Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. We used Pandas to perform data analysis on the extracted data. First, we imported pandas into our python workspace. Using pandas, we read the excel file into our workspace and convert the file into Data frame. We need some framework to do things very fast and do it efficiently use a distributed, persistent and pipelined processing. Once the data is in Dataframe it will be easy to perform data analysis on the data. We have removed all null valued rows and performed simple data analysis to understand the data. Such as total no of rows present after the data cleaning and hominy fields are ready to be used with out any kind of cleaning and transformation. Once this step is done, we have proceeded to the next phase of the project, i.e. using the collected data to perform sentiment analysis.

**Sentiment Analysis:**

Sentiment analysis is a technique which is used to analyze and categorize opinions or feelings expressed in the text format by the people on digital platform. It categorizes the text into three states of sentiments such as positive, negative and neutral.

In this project after exporting data into Dataframe we performed sentiment analysis using python (TextBlob library) a library which is used to calculate polarity and subjectivity score for the given text data. In the project we have created a new column named sentiment in our Dataframe and stored all the polarity values in it. Later, to perform further analysis on the polarity score we wrote a conditional statement (IF score < 0 “negative”, > 0 “positive”, else “neutral”) in our code and a created a new column named s\_analysis and stored all segregated data. Now we have separated all the positive and negative tweets and stored in separate variables. The reason to segregate the data is, if we perform word cloud on the entire data we can’t differentiate on what topics people were positive and negative. So, we have speared negative and positive tweets and performed word cloud and topic modeling separately.

To check the segmented sentiment analysis on total tweets we used pandas to plot the pie-chart.

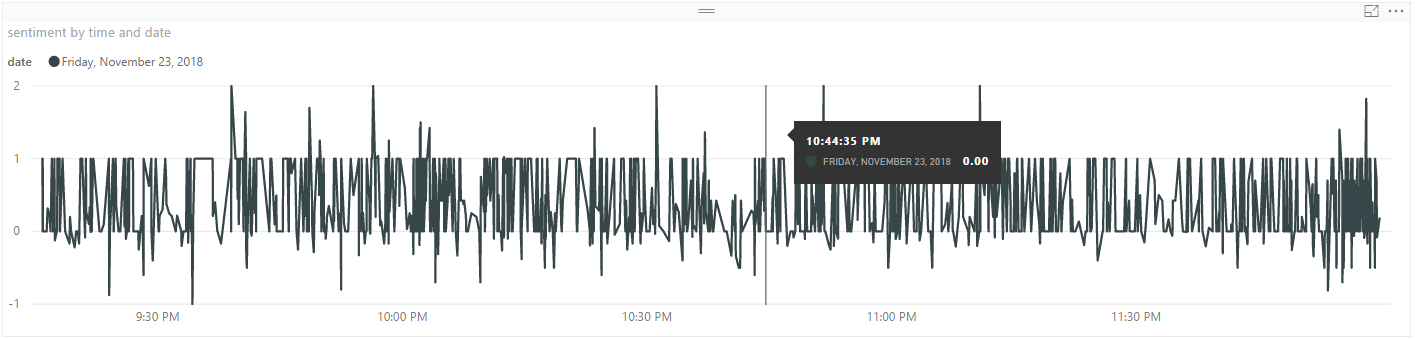


If we look at the above pie-chart we can depict 10.63% of total tweets of Target which we have extracted from tweeter are negative.

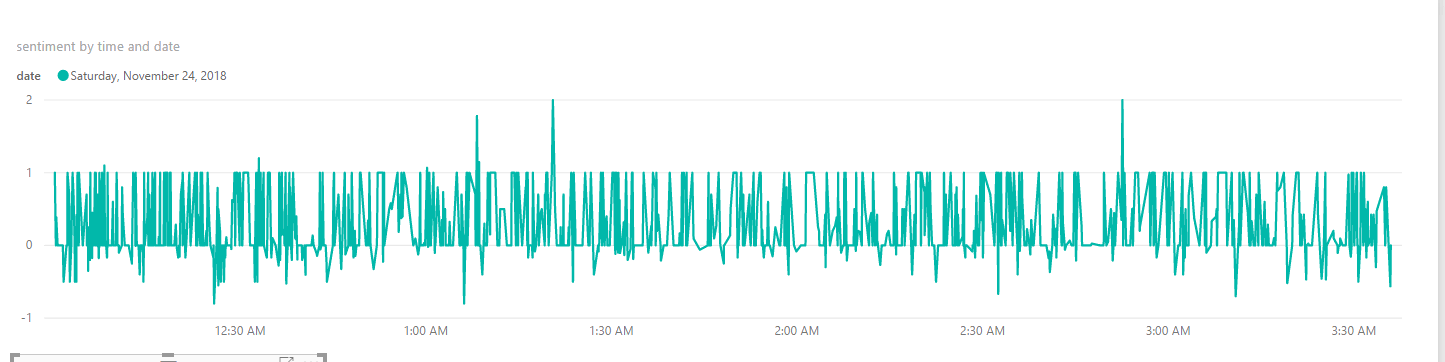
**Visualization**

We have performed visualization using Microsoft Power BI. We have imported the excel file which contains the cleaned data from twitter. After we have imported the file, we have used edit query option in Power BI to transform the created at field in the excel. This field has the date, time, day and month included in it. We need time and day in two different columns in order to get the visual appropriate. We have used split column option to transform the columns as required. Once the editing is completed, we had three different rows with date, day and time.

We have plotted sentiment score of tweets vs time for each day, and the results look like the below figures.



Sentiment on 23rd of Nov from 9:00 Pm to 11:59 PM.



Sentiment on 24th of November from 12:00 AM to 4:00 AM.

From the above figures we can infer that target received most negative comments between times 11:30 pm to 1:15 AM. We can also observe that from 9:30 PM to 11:15pm it received high number of positive spikes.

**Word Cloud**

Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. Word clouds are widely used for analyzing data from social network websites. For generating word cloud in Python, modules needed are – matplotlib, pandas and word cloud. We have divided the tweets into two categories, Positive and Negative. We have performed world could separately on both the sets. This was done to get the idea of topics to perform topic modeling. Below pictures represent the obtained world clouds for the tweets.



Word cloud for negative tweets.

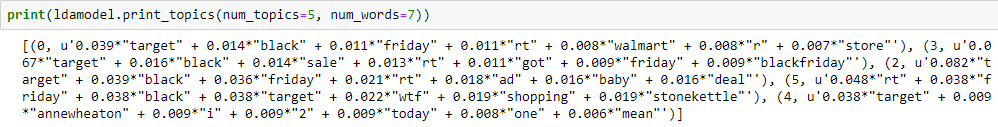


Word cloud for positive tweets.

From the above we have obtained the most talked words about the target during the time when we pulled the tweets. We can find interesting topics such as ***Walmart, WTF, Employee, Baby Deals*** in them. When we observe positive tweets, we can observe ***Chance, Win, Pop, Original Funko*** more prominently appearing in the picture.

**Topic Modeling**

We have performed topic modeling for both the positive and negative tweets. We performed topic modeling in python using the NLTK and Gensim Library. We did it using the LDA algorithm. We perform topic modeling in order to get the topic which are more spoken about. We thought of taking 5 topics each in both positive and negative Tweets. We have selected 7 words for each topic and 200 iteration for the topic. Below pictures are the result what we have obtained from topic modeling.



Negative topics from the tweets collected.

**First topic:** Walmart

People are comparing Target to Walmart. This can be observed from the obtained topic modeling results. We have observed the tweets for more inference, funniest what we have found is “We are shopping in Walmart with Target Carts”.

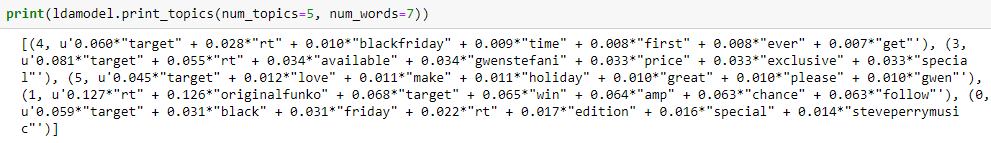
**Second Topic**: Stonekettle & Annewheaton

It happened that target have used both the authors pictures in their mobile application with out their concern. So, this raised a major backlash in the social media community.

**Third Topic**: Baby Deals.

Twitter community did not like the baby deals, they were not well received.

Topic modeling for the positive tweets is as follows.



**First Topic**: Gwen Stefani.

Target have released an exclusive edition of Gwen Stefani album for holiday session. It is well received from the social media community. Many people have twitted about this with a positive note.

**Second Topic**: Original Funko.

Original Funko is a toy manufacturer, it makes exclusive toys for target among them Pikachu and Nefller models are the most popular one. These were well received from the customers.

**Third Topic**: Chance Win

Target have announced a special offer for customers to win exclusive gifts. This is also most spoken in the twitter.

**Inferences:**

From the above we can make 4 key inferences about the Target Sales and black Friday deals during the time when collected the data.

1. Exclusive albums for the holiday season by pop singers is very well received by everyone. This can increase the brand value of target and make it most spoken by the people.
2. Exclusive collection of toys is very well received by everyone. Target should consider make arrangements to form these kinds of exclusive offers with popular brands to attract customers to their stores. These exclusive categories should be decided based on the sales and trends present at that time.
3. Using popular people pictures without their concern have caused huge backlash in the social media community. This topic has dominated the black Friday deals, which should be talked most during the holiday season. Target should take extra precautions to avoid these kinds of incidents happening during the holiday season.
4. Baby Deals were not well received by all during our period of observation More care should be taken for the next holiday season while designing the baby deals. In depth analysis should be conducted for the reason and because which caused this back lash.
5. People are comparing Target to Walmart. Target should perform in depth analysis to find out in what topics they are lagging Walmart..