Author Declaration for Group Assignments

Title of Assignment: Group 4 Mid-Term Essay
Module Number: CS7IS4
Word Count: 2396

March 8, 2020

1 Contributions

Student Number	Student Name	Nature of Contribution	%
19302263	Aishwarya Ravindran	Performed literature Review by reading previous research papers regarding Sentiment Analysis, Twitter data and weather data. Contributions to Group Essay - Introduction, Literature Review and Methods. Fulfillment of the responsibilities of *recorder*, including producing minutes of group meetings and for maintaining the file of these so that they may be consulted by group members or by lecturing staff.	20%
06374611	Cian Johnston	Collation of a dataset of textual weather forecast data from both official and unofficial sources, and preprocessing of the dataset according to the project requirements. Acquisition of a large dataset of Tweets from a publicly available source, and preprocessing of the dataset according to the project requirements. Contributions detailing the above to the Group report, available in the Git repository located at `https://github.com/johnstcn/cs7is4group4`. Fulfillment of the responsibilities of *group chair*, including arranging meetings at times suitable for all members, charing these meetings, producing agendas for these meetings, and communicating with the Lecturer.	20%
19305272	George Chavady	Attended meetings regularly to discuss and progress with the essay. Read articles on sentiment analysis performed on Twitter data. Contributed to the preparation of the Group Report. Fulfilled the responsibilities of an *Accountant*, by keeping track of the time devoted and associated contributions of each member of the group to the group project.	20%
19302270	Sameer Karode	Literature review of existing implementations of sentiment analysis done over twitter corpus. Preliminary text analysis of tweet corpus collected from Twitter API. Pre-processed the sample tweets in order to perform Sentiment Analysis. Contributed to the Group Report with above mentioned implementations. Fulfillment of the responsibilities of *verifier*, ensured that weekly responsibilities of the chair, recorder and accountant are met.	20%
19304269	Shravani Deepak Kulkarni	Referred existing papers to understand the handling of twitter data for sentiment analysis and researched on the tools available in Python for sentiment analysis. Cleaned the tweets to remove unwanted information like @mentions, retweets, hyperlinks and hashtags and transformed the tweets to replace emojis with text and contractions with its original forms for easy analysis of text. Visualized the polarity obtained from the sentiment analysis tool using box plots and word clouds. Added the implementation details in the group essay and also completed the responsibilities of *ambassador*.	20%

2 Declaration

We have read and we understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at: `http://www.tcd.ie/calendar`

We have also completed the Online Tutorial on avoiding plagiarism ‘Ready, Steady, Write’, located at `http://tcd-ie.libguides.com/plagiarism/ready-steady-write`

We declare that this assignment, together with any supporting artefact is offered for assessment as our original and unaided work, except in so far as any advice and/or assistance from any other named person in preparing it and any reference material used are duly and appropriately acknowledged. We declare that the percentage contribution by each member as stated above has been agreed by all members of the group, and reflects the actual contribution of the group members.

3 Signatures

Signed and dated: March 8, 2020

1. Cian Johnston Cian Johnston
2. Aishwarya Ravindran Aishwarya
3. George Chavady
4. Sameer Karode
5. Shravani Deepak Kulkarni KShravani

Sentiment Analysis of Twitter Data based on Implicit Categories in Textual Weather Data

Johnston, Cian Ravindran, Aishwarya Chavady, George
Karode, Sameer Kulkarni, Shravani Deepak

March 8, 2020

1 Introduction

Sentimental mining is a critical and essential area because sentiment fundamentally relates to a person's emotions, impression and attitude. Analysing every individual's sentiments from a text analytics point of view accurately is challenging. We are basically trying to comprehend the thoughts and assumptions of an individual regarding a concept or topic with some context be it positive, negative or neutral[1]. It is estimated that nearly 2.5 Quintillion bytes of data is generated each day[2] and there is a plethora of information relating to a person's messages, tweets, documents, emails, chats, conversations and comments available. Sentiment analysis aids us in analysing this huge amount of data efficiently.

In this paper, we focus on performing sentiment analysis on the weather based on data retrieved from Twitter[3]. Twitter, being an acclaimed stage for social networking possess and people to express their interests and thoughts has plenty of information from posts known as "tweets". With over 500 million tweets per day pertaining to almost anything, it is an excellent platform due to its accessibility and real-time analysis of the data which is crucial for sentiment analysis. It is not uncommon that weather plays an important role in an individual's mood and its consequences because of the decisions made.

RESEARCH QUESTION - "Can we identify a correlation between Implicit categories of Textual Weather data and Sentiment Analysis of Twitter data for the inhabitants of Ireland?"

In this paper, we present an analysis of tweets for the year 2018 in the geographical area of Ireland. At a broader level, we analyze the tweets season-wise to understand the general sentiment in the tweets for each of the four seasons that are experienced in Ireland (winter, spring, summer and autumn). For instance, we compare the emotions in the tweets in summer vs the tweets

[1] *Why is sentiment analysis important from a business perspective.* `https://blog.aylien.com/why-is-sentiment-analysis-important-from-a-business-perspective/`.

[2] *How much data do we create every day - the mind-blowing stats everyone shoudl read.* `https://blazon.online/data-marketing/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/`.

[3] *Mining twitter data.* `https://towardsdatascience.com/mining-twitter-data-ba4e44e6aecc`.

in winter. Additionally, at a deeper level, we analyze the tweets for each day and correlate it with the weather on that day. In this case, we find the trends in the sentiments on sunny, cloudy, wet, dry or cold days. Furthermore, we also compare these trends on weekdays and weekends. For these analysis, we make use of visualizations to gain more insights into finding patterns between the weather and sentiment of the population within a given geographical region.

2 Related Work

As the accessibility to real-time information or otherwise is increasing, performing analysis on this data is also becoming viable. There has been a lot of notable research done on sentiment analysis of twitter data in general and studies emphasizing on correlating weather with people's mood and sentiments from Twitter data by measuring temperature, humidity and atmospheric pressure[4]. These weather variables seem to have a satisfactory impact on one's mood. Another paper[5] focuses on implementing sentiment analysis on Twitter data using the Twitter API's and a wealth of available libraries. Twitter is known for having small texts and abbreviations used by people in their posts or tweets making it challenging to extract polarity of the texts and hence researchers resort to utilizing deep learning and machine learning techniques.

Researchers have performed sentiment analysis for various reasons and different areas like elections, politics, movie ratings and fashion to name a few. In 2015, there was a noteworthy research done in the vision of predicting future crime on each area of a major city, Chicago, Illinois of the United States using GPS tagged twitter data[6]. They aimed to predict the time and location during which a specific type of crime is expected to occur by applying lexicon-based sentiment analysis on categorized weather data combined with kernel density estimation of historical crime incidents and were successful. Hannak et al.[7] concentrate on using a Twitter specific sentiment extraction methodology and explore a corpus of over 1.5 billion tweets. With the help of machine learning techniques on Twitter corpus correlated with the weather at a particular time and location of the tweets, it was concluded that aggregate sentiment follows different climate and seasonal patterns.

3 Methodology

Our hypothesis is that there exists a significant correlation between the overall sentiment of Tweets at a given time, and the features of a given weather forecast for the same specified time on which that Tweet was posted. The null hypothesis is that no significant correlation exists. In order to address the hypothesis, two sources of data were procured: a collection of posts from Twitter, and a

[4] Kunwoo Park et al. "Mood and weather: Feeling the heat?" In: *Seventh International AAAI Conference on Weblogs and Social Media*. 2013.

[5] Hamid Bagheri and Md Johirul Islam. "Sentiment analysis of twitter data". In: *arXiv preprint arXiv:1711.10377* (2017). [Accessed 2020-02-20].

[6] Xinyu Chen, Youngwoon Cho, and Suk Young Jang. "Crime prediction using twitter sentiment and weather". In: *2015 Systems and Information Engineering Design Symposium*. IEEE. 2015, pp. 63–68.

[7] Aniko Hannak et al. "Tweetin' in the rain: Exploring societal-scale effects of weather on mood". In: *Sixth International AAAI Conference on Weblogs and Social Media*. 2012.

collection of textual weather data. Both of these corpora were assembled and filtered such that they both relate to the same geographical area, and to the same timespan. The geographical area in question was restricted to the Republic of Ireland, and the timespan in question was limited to the year of 2018.

3.1 Twitter Data

3.1.1 What to measure

To measure the sentiment of a Tweet, we elected to use the `TextBlob`[8] library, which internally utilises a Naive Bayes classifier pre-trained on a corpus of movie reviews. Applying this classifier to a text yields a tuple of two real-valued integers in the range [-1.0, 1.0], denoting both the *polarity* and the *subjectivity* of the text.

3.1.2 Data Collection

Twitter provides an API for developers to both read data and interact with users. This has a number of limitations, including rate limits[9], and a requirement to request an API key. In order to access a useful volume of data, this would require a large number of HTTP requests to Twitter. Thankfully, the Internet Archive provides a large dataset of posts on Twitter for the year of 2018 in TAR format[10]. While these datasets are essentially a subset of the *global* content of Twitter, and technically much larger than required, they are hosted using BitTorrent and are thus much more straightforward to download. For the purposes of this paper, we elected to use a subset of the Twitter archive data from the year 2018.

3.1.3 Data Processing

The large archives detailed above needed to be preprocessed before any meaningful analysis can be performed. To this end, a small program was written to consume the entire TAR archive and filter out Tweets matching certain criteria, serializing a subset of the data to CSV format. This was done to avoid extracting the entire archive to disk, which is a time-consuming operation. The following criteria were used for extracting Tweets of interest:

- Having geo-location information within Ireland, or
- Posted by a user with user-specified location containing the string *Ireland*.

Note that not all Tweets posted by users in Ireland will necessarily have geo-location information attached, and not all Tweets posted by users claiming to be located within Ireland are necessarily so.

[8] *textblob documentation.* `https://buildmedia.readthedocs.org/media/pdf/textblob/latest/textblob.pdf`.

[9] *Rate Limiting - Twitter Developers.* `https://developer.twitter.com/en/docs/basics/rate-limiting`.

[10] *Archive Team: The Twitter Stream Grab.* `https://archive.org/details/twitterstream`.

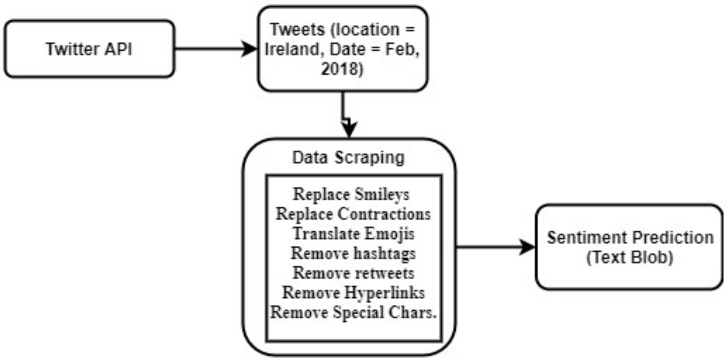

Figure 1: Twitter data process flow.

3.1.4 Data Cleaning

The tweets are 140 characters long. Considering this constraint, initial analysis from the chunk of tweets revealed that, users tend to use more contractions and emojis to express their thoughts rather than using appropriate grammar. The sentiment score given by the existing implementations like TextBlob is affected by the use of such text. We implemented a python code to do the preprocessing of the data obtained, which is obtained using Twitter API.

The filtering process removes the URLs, usernames, and hashtags. In some situations, it is known that hashtags can provide instant insight as to what the users are feeling. However, most hashtags that we encountered contain meaningless text or sentences for tags instead of keywords. The process flow is illustrated in Figure 1.

- We removed hashtags, retweets, @mentions, hyperlinks and special characters using regular expressions.

  for example,
  Sample Input:
  Hello there, This is sample tweet with mentions@john, #hashtag and link https://google.com
  Processed Output:
  Hello there, This is sample tweet with mentions, hashtag and link

- Next, we replaced all the contractions in the tweets to its original forms. We used a dictionary of all the contractions with its original form and replaced them in the text.

  for example,
  Sample Input:
  Hello, how're you doing?
  Processed Output:
  Hello, how are you doing?

- Lastly, we translated all the emojis/smileys to texts. For the smileys using special characters, we used the same approach as in the previous step (a dictionary containing the smileys and their corresponding meaning). We utilized the emoji library available in python for this conversion.

  for example,
  Sample Input:
  Sample tweet with :-) and <3
  Processed Output:
  Sample tweet with happy smiley and love

3.2 Weather Data

3.2.1 What to measure

For the purposes of this hypothesis, given a textual weather forecast and a finite set of potential features of weather forecasts, we map each textual forecast to a subset of those potential features. The features of each individual forecast text are determined by the presence or absence of a set of manually selected tokens.

3.2.2 Data Collection

Some historical weather data was collected from MET Éireann, the Irish National Meteorological Service.[11]. Unfortunately, historical textual weather forecast data is not available from MET Éireann; to work around this, we used the Internet Archive's Wayback Machine[12] to access previously published versions of the MET Éireann homepage which contains daily textual forecast data.

To access previously versions of the website more conveniently, the tool `wayback-machine-scraper` was utilised[13]. This is a command-line utility that interfaces with *Wayback Machine* and allows a user to download a number of snapshots of a website for a specified date range. For the purposes of this paper, we fetched all the saved snapshots of the MET Éireann homepage for the year of 2018. Note that this is a sparse dataset, and snapshots of this data is not available for every day.

As an alternative source of textual forecast data, we turned to a more unorthodox source – Boards.ie is a discussion board with a wide range of fora which, naturally, includes the topic of weather. One particular thread of interest on this sub-forum has almost daily forecasts provided by an amateur meteorologist with the moniker 'M.T. Cranium'[14]. The relevant print versions of the thread spanning the year of 2018 were saved, and the relevant daily forecasts were extracted using a Python script. The process is illustrated below in Figure 2.

[11] *Met Éireann Forecast - The Irish Meteorological Service.* `https://www.met.ie`.

[12] *Internet Archive: Wayback Machine.* `https://archive.org/web/`.

[13] *GitHub: sangaline/wayback-machine-scraper: A command-line utility and Scrapy middleware for scraping time series data from Archive.org's Wayback Machine.* `https://github.com/sangaline/wayback-machine-scraper`.

[14] *Your daily forecasts from Boards.ie weather forum (NO CHAT) - boards.ie.* `https://www.boards.ie/vbulletin/showthread.php?t=2055579971`.

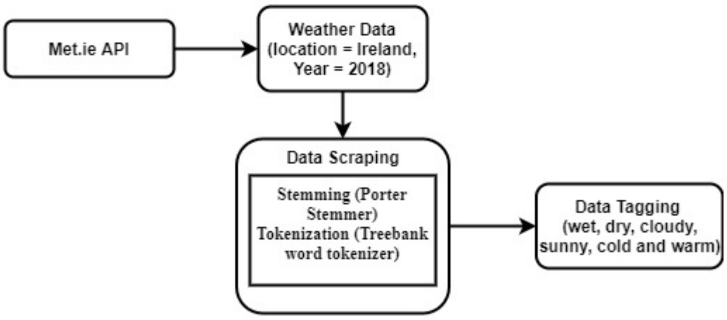
Met.ie API
Weather Data (location = Ireland, Year = 2018)
Data Scraping
Stemming (Porter Stemmer) Tokenization (Treebank word tokenizer)
Data Tagging (wet, dry, cloudy, sunny, cold and warm)

Figure 2: Weather data process flow.

3.2.3 Data Processing

The textual forecast data on its own needs to be processed before any meaningful correlations can take place. To this end, we utilised the Python NLTK libraries to perform the following:

- Stemming and tokenization of the raw text. NLTK provides pre-implemented stemmers and tokenizers for the English language. For stemming we utilised a `PorterStemmer`[15], and for tokenization, a `TreebankWordTokenizer`[16]. For example, the raw string `‘‘TODAY ... Continued mild with a few outbreaks of light rain, highs near 12 C.’’`, after stemming and tokenization, becomes a list of tokens `[ ‘today’, ‘...’, ‘continu’, ‘mild’, ‘with’, ‘a’, ‘few’, ‘outbreak’, ‘of’, ‘light’, ‘rain’, ‘,’ ,‘high’, ‘near’, ‘12’, ‘c’, ‘.’ ]`.

- Annotation of the tokenized forecast data with some predefined features. For an initial first pass, we defined a naïve approach whereby a number of binary features were defined: *wet, dry, cloudy, sunny, cold, warm.* For each feature, a number of tokens were selected to signify if the feature should be 1, and 0 otherwise. For example, the presence of the token *rain* in the forecast signifies a value of 1 for the feature *wet.* These features were then extracted for all of the collected forecast data and serialized to CSV format.

3.3 Sentiment Analysis

After preprocessing of the tweet corpus, the same is processed for carrying out the sentiment analysis. Initially, the text corpus is processed using TextBlob, which is a python library built over Natural Language Toolkit(NLTK). TextBlob provides a sufficient set of tools for performing tasks like Sentiment Extraction, Spelling Correction and Detection of Language. There are two types of sentiment analyzers, by default, the TextBlob uses PatternAnalyzer and second is NaiveBayesAnalyzer. The next steps are to evaluate the sentiment score given by the default and overridden implementation of Analyzer. Once the tweets are

[15] *nltk.stem package.* `https://www.nltk.org/api/nltk.stem.html`.

[16] *nltk.tokenize package.* `https://www.nltk.org/api/nltk.tokenize.html`.

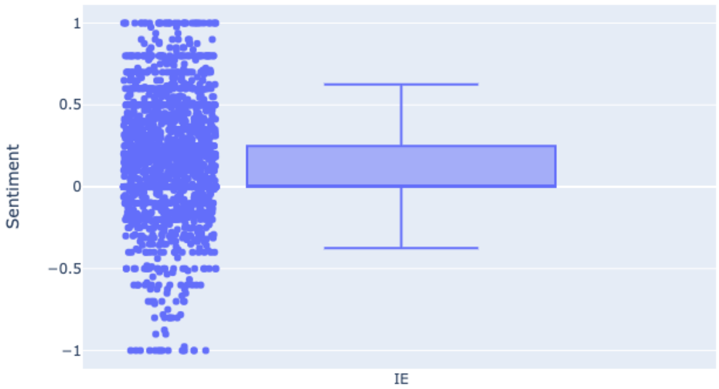

Figure 3: Box plot for the sentiment analysis of the tweets for the month of January 2018.

labelled with its sentiment score, the analysis is done to find correlations between weather data and sentiment score of tweets. The weather data is classified into categories for analysis.

4 Results

The findings of our initial analysis of sentiment in Twitter data were that more tweets showed positive sentiment than negative sentiment. This was observed in the case of the tweets analyzed for the month of January in 2018 for the tweets from Ireland. The box-plot showcasing this shown in the Figure 3. The sentiment takes the values range from [-1, 1], where -1 is a completely negative sentiment, +1 is a completely positive sentiment and 0 means neutral sentiment.

5 Future Work

Our future work would involve collecting Twitter data posted during different seasons of the year. More specifically, when the weather data collected is disparate. This data would help us in gauging the effects of weather on Twitter posts. For example, an increase in temperature from lower values to values between 15C and 20C are associated with significant increases in positive expressions. On the contrary, the number of expressions of positive sentiment decline when the temperature exceeds a value of 30C.

6 Conclusion

Large amounts of data is readily available on Twitter which can be used to analyze sentiment of the users' tweets. Also, historical and real time weather data is made available by the meteorological department. We could use weather data and analyze it together with the tweets to predict sentiment for a particular weather forecast using Machine Learning algorithms. Studies have shown that weather conditions such as temperature, precipitation, cloud cover, wind speed and humidity each significantly relate to the expression of sentiment in social media such as Twitter and Facebook. Analysis of sentiment correlated with weather data gives us some important background information of the users of the system, which could be very useful in decision making.