HPV Vaccination (Topic 2)

A PROJECT REPORT

Submitted by

Team Index #12

Team members:

(Pavan Kumar Reddy Gunnala, 81173384 pgunnala@kent.edu)

(Jagadeeshwar Reddy Kothakapu, 811180574 jkothak1@kent.edu)

(Divya Sri Karingula, 811172271 dkaringu@kent.edu)

Under the Guidance of

Qiang Guan

Associate Professor

Department of Computer Science

Kent State University

Table of Contents

1. INTRODUCTION

1.1 OVERVIEW

The aim while performing sentiment analysis on tweets is basically to classify the tweets in different sentiment classes accurately. In this field of research, various approaches have evolved, which propose methods to train a model and then test it to check its efficiency. Performing sentiment analysis is challenging on Twitter data, as we mentioned earlier. Here we define the reasons for this:

- Limited tweet size: with just 140 characters in hand, compact statements are generated, which results sparse set of features.
- Use of slang: these words are different from English words, and it can make an approach outdated because of the evolutionary use of slangs.
- Twitter features: it allows the use of hashtags, user reference and URLs. These require different processing than other words.
- User variety: the users express their opinions in a variety of ways, some using different language in between, while others using repeated words or symbols to convey an emotion

The HPV vaccination was developed to prevent HPV infection and the majority of cervical malignancies. Negative information about HPV vaccines from celebrities, health experts, and the news media has been shown to increase vaccine reluctance and refusal. However, because the HPV vaccine is still a relatively new addition to public health, it is critical to monitor social media for varied viewpoints on immunization.

1.2 Motivation

The use of social media and twitter has increasingly become popular in the pandemic period. The effect of the HPV Vaccination on the sentiments of the society, mainly due to the tweets and comments of the influential people of the society and the response to their tweets was the primary source of motivation behind this study.

Internet is amongst the most rapidly developing technologies and has become an essential part in today’s world. Data on internet varies from areas like academics, criticism or conclusion about items, remarks on social issues and so forth. Individuals regularly communicate examine and share data via web. It helps individuals to compare and settle on choice in numerous things. Large number of individuals dependably tune into other’s assessment before making any choice of the service. For instance, in a case of preference for watching a movie, fairly large number of individuals prefer to select a movie based on reviews or ratings given by other individuals.

Various organizations gather data through their websites and the data which is assembled is analysed to decide the sentiment behind it. One such example is e-commerce where before buying any item, people prefer to check the item review and ratings by other customers. The project exhibits strategies to analyse the reviews and extract their sentiment. The fundamental goal is to anticipate the sentiment of a review by performing and analysing a group of feature reduction methodologies. This project also attempts on extracting compelling features that can give better outcome.

1.3 Problem statement:

This project aims to study the community trends on HPV vaccination via Twitter. Whether it may be a positive opinion or negative opinion on the vaccination. We collected Twitter data with the searching keywords of HPV, HPV vaccine, HPV vaccination, etc. So, we are going to analyze the HPV vaccination on the community trends.

1.4 OBJECTIVE

The main objective of the project is Data Analyzing using sentiment analysis, topic analysis, and community analysis. All results are required to visualize via web interface.

- In sentiment analysis, we are going to extract the opinions based on the given text. So, mainly the text is classified as objective and subjective text.
- In community analysis, we are going to form the hard clusters which belong to one community. So, the data can be classified into positive, negative or neutral.
- Topic analysis is a machine learning technique for assigning subjects to text input automatically where topic analysis examines unstructured text such as emails and social media

2. Background

A major research field has emerged around the subject of how to extract the best and most accurate method and simultaneously categorize the customers' written reviews into negative or positive opinions. In a 2002 publication, Pang, Lee and Vaithyanathan were the first to propose sentiment classification using machine learning models on movie reviews dataset. They analysed the Naive Bayes, Max Entropy and Support Vector Machine models for sentiment analysis on unigrams and bigrams of data. In their experiment, SVM paired with unigram feature extraction produced the best results. They reported a result of 82.9% accuracy.

In a 2004 publication, Mullen and Collier performed sentiment classification on clothing, shoes, and jewellery product review datasets. They compared methods of hybrid SVM, Naive Bayes, LR, and decision tree with feature extraction methods based on Lemmas and Osgood theory. In their study, SVM produced the best results with an accuracy of 86.6%. In a 2015 publication, Lilleberg, Zhu, and Zhang performed a comparison study of TF-IDF and Word2vec feature extractions using SVM. They also compared the classification results with and without including stop words. The best result of SVM with TF-IDF and without stop words that they saw was 88% accuracy.

In recent years, the common classification techniques for document analysis include SVM and LR. In a 2017 publication, SVM and sentiment analysis were proposed by Elmurngi and Gherbi to detect fake movie reviews. They compared SVMs with Naive Bayes, decision tree, and KNN classifications performance on a corpus with stop words and a corpus without stop words. In both cases, SVM performed the best, with accuracies of 81.75% and 81.35%, respectively. In another publication, Ramadhan et al. conducted a sentiment analysis using logistic regression and TF-IDF feature extraction on a social media Twitter dataset. The classification accuracy was reported to be close to 83%. In 2018, Das and Chakraborty conducted an experiment using SVM, TF-IDF model coupled with Next Word Negation on an Amazon product review dataset and reported accuracy 88.86%.

In a publication in 2018, Bhavitha, Rodrigues, and Chiplunkar also performed a comparative study of several machine learning methods, lexicon-based methods and sentiment analysis on movie reviews. For the SentiWordNet method, they reported an accuracy of 74%, and for the SVM method, they reported an accuracy of 86.40%. In the same year, Athanasiou applied Gradient Boosting machine learning for sentiment analysis and found superior performance over SVM, Naive Bayes, and neural network for both balanced and imbalanced data sets. The Gradient Boosting machine learning performed best with an accuracy of 88.20%.

3. SYSTEM REQUIREMENTS

S.no	Hardware	Description
1	Processor	Min I 3 Processor
2	Software	Python V3 or above
3	RAM	4GB RAM
4	Hard Disk space	250GB

4. METHODOLOGY

4.1 The ML tasks and their evaluation criteria

Sentiment analysis is sometimes called opinion mining, a method to process Natural language. Natural language Processing (NLP) is identified with territory of machine-human cooperation. Sentiment analysis can be termed as an errand of recognizing the survey's opinion. The conclusion might be classified as negative, neutral, or positive extremity. Sentiment analysis can be classified into three diverse as sentence level, document level, and entity-aspect level. In a sentence level, a supposition of specific sentence is considered as a priority for sentiment prediction. Whereas document level is a more generalized feeling which considers the whole document for sentiment prediction. And if the focus is straightforwardly on the opinion itself then it can be termed as an entity-aspect level sentiment analysis.

The databases are inflating enormously because of the vast collection of data electronically. Information retrieval is the procedure of extracting important information regarding data from a larger collection of data in the databases. Naïve Bayes, Logic Regression, and Support vector machine are the most used machine learning algorithms for prediction of sentiment. The analysis of a sentiment faces a couple of arguments during its investigation.

Classification accuracy is the major issue. This gives a motivation for acquiring a good classification precision picking great feature determination, pre-processing along with order procedures. Process of Sentiment analysis is shown in figure 1. Customer's opinion is posted on websites, blogs or forums. The data format of the customer's opinions is unstructured and messy. At the first place, the unstructured data is changed over into organized frames. After that point, features are extracted from that organized frame utilizing feature selection strategy. The last step of the analysis goes with the classification algorithm for predicting the sentiment of the records.

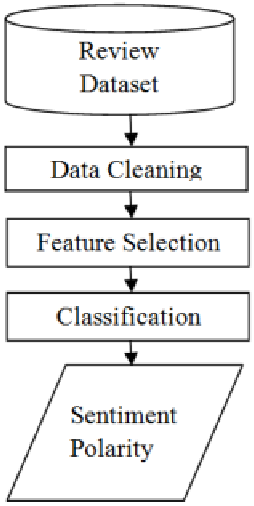


Figure 1: Sentiment Analysis Process flow

The pre-processing of the data is a very important step as it decides the efficiency of the other steps down in line. It involves syntactical correction of the tweets as desired. The steps involved should aim for making the data more machine readable to reduce ambiguity in feature extraction. Below are a few steps used for pre-processing of tweets

Data cleaning

Special character removal and stop word expulsion strategy has been performed. New lines, unwanted punctuations, and stemming is likewise executed as a piece of pre-processing strategy. Tokenization separates given text into tokens. Natural language processing tool kit (NLTK) is utilized as a part of numerous existing papers using python to pre-process the dataset. NLTK is a python library for solving various text analytics and natural language processing tasks. The following different strategies forms pieces of the pre-processing step, as a part of sentiment analysis:

- Conversion from upper case to lower case letter, expel undesirable punctuations and additional white spaces, and evacuate newline and special characters.
- Stemming, a process to reduce inflected or morphological word forms to their base form or root form i.e., word stem is used to reduce the total number of unique words in the dataset. Porter Stemming algorithm is an algorithm to expel suffixes from English words. Porter stemmer is most broadly utilized algorithm for stemming words.

4.2 Feature Extraction

A feature is a piece of information that can be used as a characteristic which can assist in solving a problem (like prediction). The quality and quantity of features is very important as they are important for the results generated by the selected model. Selection of useful words from tweets is feature extraction.

- Unigram features – one word is considered at a time and decided whether it is capable of being a feature.
- N-gram features – more than one word is considered at a time.
- External lexicon – use of list of words with predefined positive or negative sentiment.

Frequency analysis is a method to collect features with highest frequencies Further, they removed some of them due to the presence of words with similar sentiment (for example happy, joy, ecstatic etc.) and created a group of these words. Along with this affinity analysis is performed, which focuses on higher order n-grams in tweet feature representation.

4.3 Classification approaches

Machine learning and lexicon-based approaches are two the types of classification approaches. Lexicon-based approach is also classified into 2 categories,

1.Corpus-based:
The corpus-based approach classifies words by considering the bunch of words as word list. Furthermore, the Corpus based is classified as semantic and statistical approach. The semantic approach uses terms to represent in semantic space for finding relationship in terms. The approach using statistical terms identifies the sentiment by utilizing the co-occurrences of words. The approach using dictionary as a collection of words helps the sentiment to be classified using the antonym and synonym of the words from WordNet, a lexical dictionary.

2.Dictionary-based approach:
The second most popular classification approach is using Machine learning (ML) algorithms. Supervised learning and Unsupervised learning are the two main types of ML algorithms. Supervised classification algorithms are the classifiers which can be a decision tree classifier, linear classifier, probabilistic classifier, or rule-based classifier. The supervised learning works in a fashion in which it relies on labelled dataset as a training model input. This model is then used to predict the test dataset for performing classification. While the unsupervised learning algorithms works on an unlabelled training model input. The main motive is to find patterns and get inferences which help to understand the nature of the dataset. The procedure of sentiment classification with machine learning goes with the first step being extracting the compelling features and forming an input for the ML algorithm. And the second step is to apply the ML algorithm to create a model for classification and prediction.

4.4 The ML/DL methods you used and their description

Lexicon based approach follows its procedure based on semantic orientation. Semantic orientation of expressions is determined as positive on the off chance that it is more identified with "best” and is negative if it is more identified with "poor". There are certain hard coded allocations given by the algorithm to the respective words for example the negative word has been given a -1 value. In the same way, the word with a positive polarity has been given a +1 value. A neutral word is allocated a null value i.e., a numerical zero. Apart from it, weak positive and weak negative has been given +0.5 and -0.5 values each. These values are called as semantic orientation scores.

The algorithm identifies the synonyms of the words by using calculated scores and WordNet. The algorithm compares each word feature with other features of the dataset according to their scores and the features with the relative scores are clustered. There are two functions in the algorithm namely Sentence Sentiment Scoring Function (SSSF) and Sentiment Aggregation Function (ESAF). The semantic orientation score for every entity is distinguished by SSSF. And ESAF ascertains the aggregate sentiment scores for an entity.

5. TWITTER SENTIMENT ANALYSIS WITH PYTHON

5.1 Python

Python is a high level, interpreted programming language, created by Guido van Rossum. The language is very popular for its code readability and compact line of codes. It uses white space inundation to delimit blocks. Python provides a large standard library which can be used for various applications for example natural language processing, machine learning, data analysis etc. It is favoured for complex projects, because of its simplicity, diverse range of features and its dynamic nature.

5.2 Natural Language Processing (NLTK)

Natural Language toolkit (NLTK) is a library in python, which provides the base for text processing and classification. Operations such as tokenization, tagging, filtering, text manipulation can be performed with the use of NLTK. The NLTK library also embodies various trainable classifiers (example – Naïve Bayes Classifier). NLTK library is used for creating a bag-of words model, which is a type of unigram model for text. In this model, the number of occurrences of each word is counted. The data acquired can be used for training classifier models. The sentiment of the entire tweets is computed by assigning subjectivity score to each word using a sentiment lexicon.

The pre-processing in Python is easy to perform due to functions provided by the standard library.
Some of the steps are given below:

- Converting all upper-case letters to lower case.
- Removing URLs: Filtering of URLs can be done with the help of regular expression (http|https|ftp) ://[a-zA-Z0-9\\./] +.
- Removing Handles (User Reference): Handles can be removed using regular expression - @(\w+).
- Removing hashtags: Hashtags can be removed using regular expression - #(\w+).
- Removing emoticons: We can use emoticon dictionary to filter out the emoticons or to save the occurrence of them in a different file.
- Removing repeated characters.

6. The Experimental Evaluation

6.1 Dataset:

Data:

Twitter data (https://drive.google.com/file/d/10XJJWIn3-RTnvZDVf7J9TBQ8My2W1GHX/view)

Twitter began as an SMS text-based service. This limited the original Tweet length to 140 characters (which was partly driven by the 160character limit of SMS, with 20 characters reserved for commands and usernames). Over time as Twitter evolved, the maximum Tweet length grew to 280 characters - still short and brief but enabling more expression.

All Twitter APIs that return Tweets provide that data encoded using JavaScript Object Notation (JSON). JSON is based on key-value pairs, with named attributes and associated values.

7. Implementation:

In this project we will be having different installations which are described below
1. Install Python for your platform
2.Install all the libraries required for this project. We need to install all libraries present in the software requirements section.
3. Run the app.py file using your interpreter which allows to compile all your python code and store it in a file named pycache folder.
4.You would find all the front-end development from the html file located in the template folder in the plotly folder. All the frontend related work will be present in that folder.
5.For dataset we used the dataset given in the blackboard which is a JSON file which consists of tweets from January 2018 to April 2019.
6.Preprocessing the dataset before applying any transformations or analysing this involves the following things to be done which are

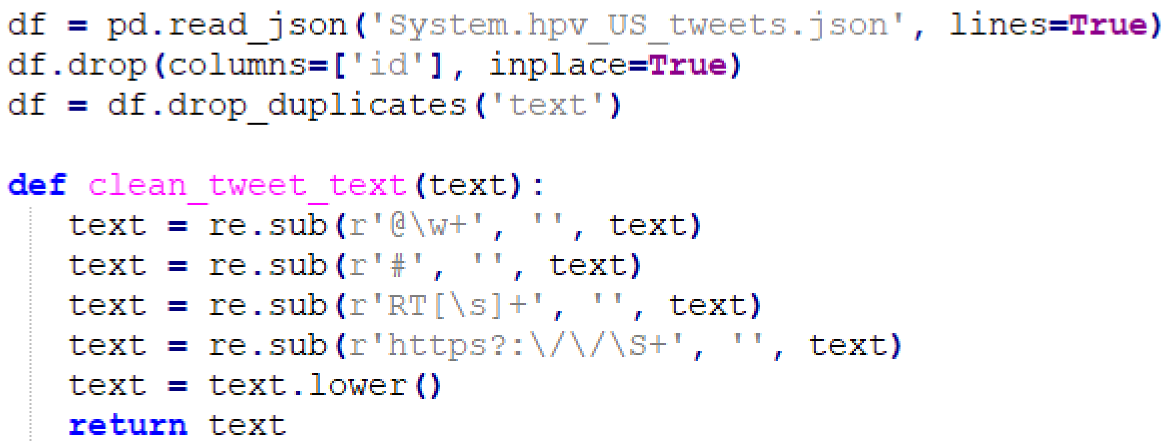
I. Importing Dataset

II. Deleting redundant data

III. Removing punctuations

IV. Sorting by dates

The code for these pre-processing can be found in the below figure



```
df = pd.read_json('System.hpv_US_tweets.json', lines=True)
df.drop(columns=['id'], inplace=True)
df = df.drop_duplicates('text')

def clean_tweet_text(text):
    text = re.sub(r'@\w+', '', text)
    text = re.sub(r'#', '', text)
    text = re.sub(r'RT[\s]+', '', text)
    text = re.sub(r'https?:\/\/\S+', '', text)
    text = text.lower()
    return text
```

7. The above code will be written in a python file and the further process in this code include as

→ Cleaning the tweeted data
→ Then we will be working on text blob library

Text Blob is **a Python (2 and 3) library for processing textual data**. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more

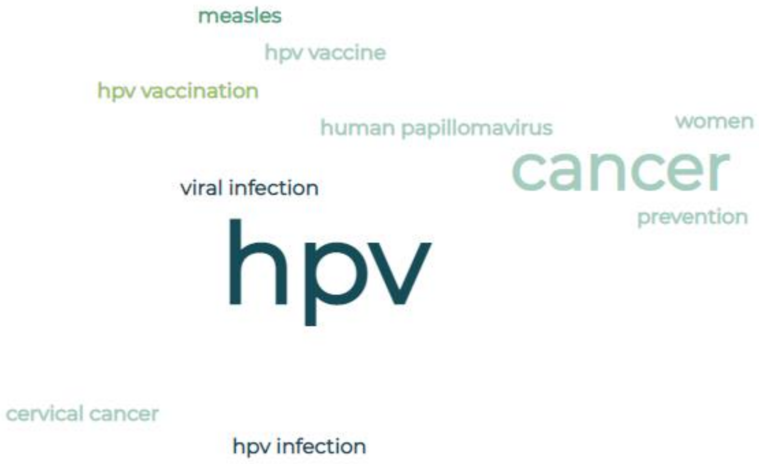
8. Now we will be working on the Natural Language tool kit where we will be using the following packages

```
Package punkt is already up to date!

[nltk_data] Downloading package averaged_perceptron_tagger to

[nltk_data]     C:\Users\91897\AppData\Roaming\nltk_data...

[nltk_data]   Package averaged_perceptron_tagger is already up-to-

[nltk_data]       date!

[nltk_data] Downloading package brown to
```

- Punkt **Sentence Tokenizer**. This tokenizer divides a text into a list of sentences by using an unsupervised algorithm to build a model for abbreviation words, collocations, and words that start sentences. It must be trained in a large collection of plaintexts in the target language before it can be used. Averaged_perceptron_tagger is **used for tagging words with their parts of speech (POS)**

- The Brown Corpus was **the first million-word electronic corpus of English**, created in 1961 at Brown University. This corpus contains text from 500 sources, and the sources have been categorized by genre, such as news, editorial, and so on.

9. We will be using word cloud library from python inorder to have reference what are all the tags repeated and what kind of words we need to use to analyse sentiments



10. Now coming to the main code we will be retrieving the tags using nltk library that is we will be printing the tags we got using that library

11. Then We will be knowing the noun phrases and word which ends with line end, so that we will be knowing what sentiment need to be analysed from different sentence.

12. Now we will be knowing the number of sentences in the dataset, and we will be retrieving the polarity and subjectivity of the sentences.

#HPV is a common virus that can lead to six types of cancers later in life.

polarity: -0.15

subjectivity: 0.25

There is no cure, but it can be prevented… https://t.co/wu5CQtKmaF

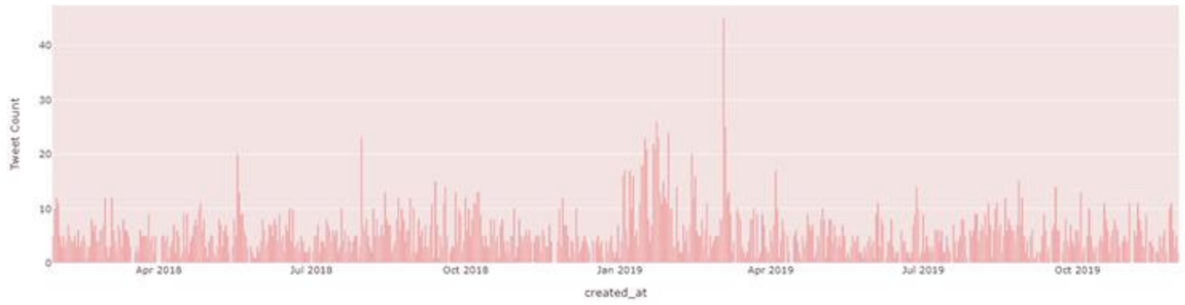
polarity: 0.0

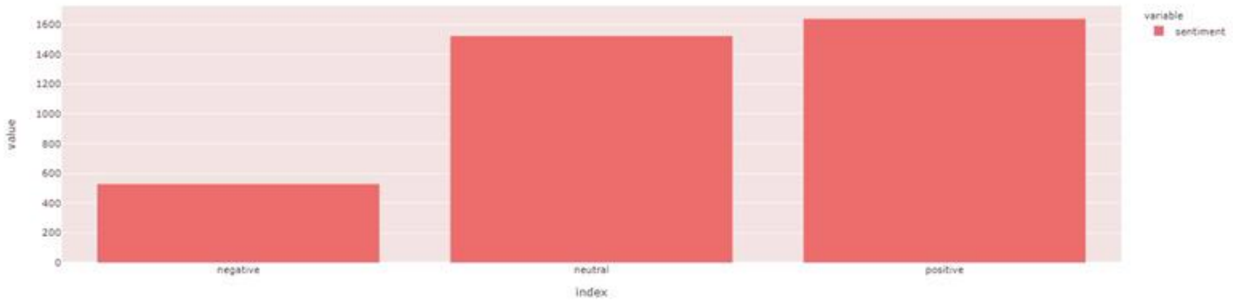
subjectivity: 0.0

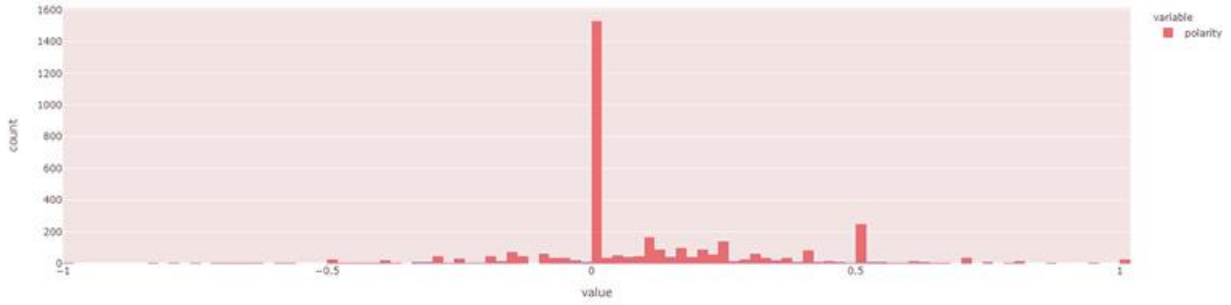
13. Now we will be plotting a histogram which tells us about the polarity of the sentences from dataset. The overall tweets in the dates are plotted in this project to know the number of tweets we got in different months of a year.

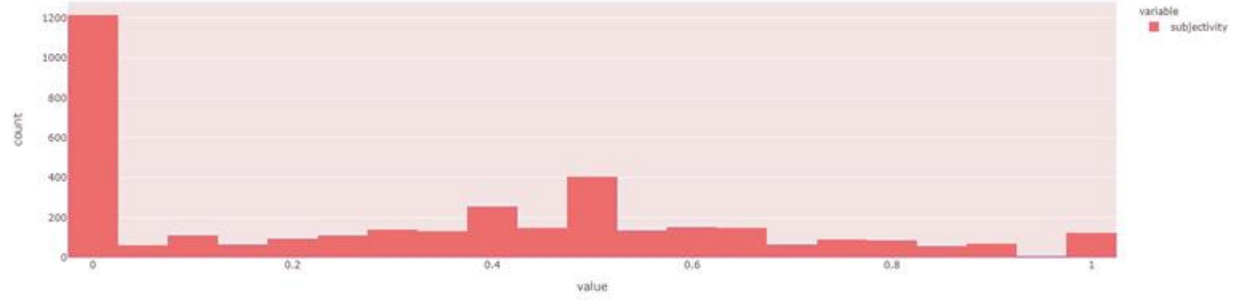
Then we retrieved the the histogram for all the tweets with sentiments as positive, negative and neutral.

Before plotting the main polarity bar graph, we analyzed the polarity and subjectivity using twi different bar graphs. The output can be viewed in the command prompt while running the code, here we are plotting out results to be visualized.

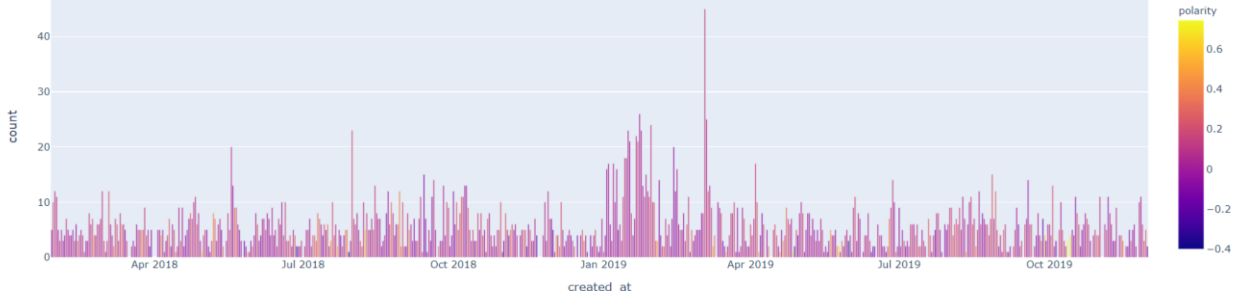


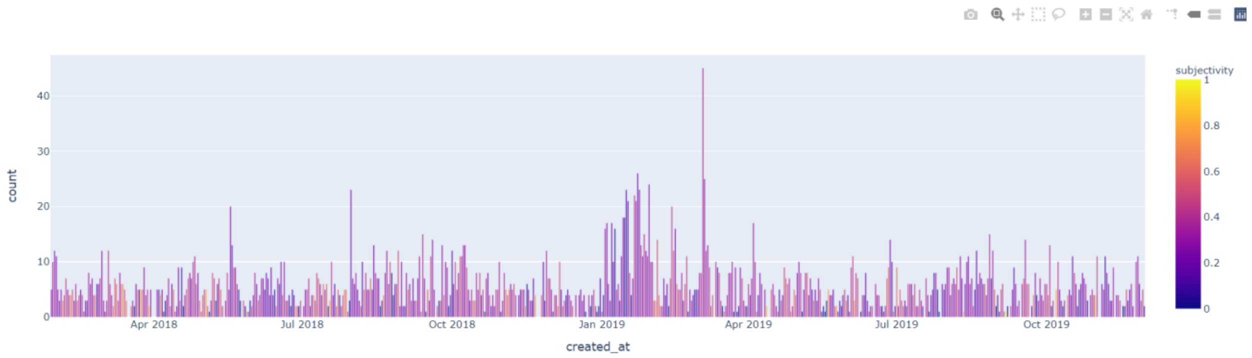






14. Then we plotted the bar graphs showing us the polarity and subjectivity of the sentiments analyzed using the sentences we retrieved using nltk library.

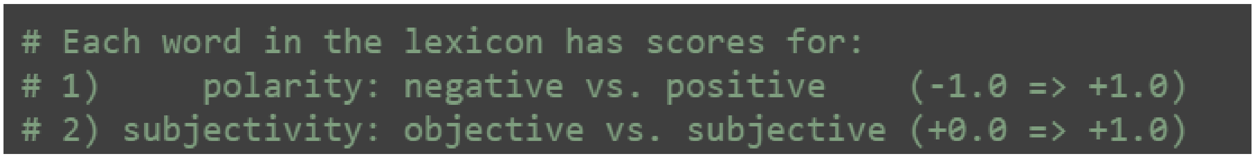




- The above plotting shows us the polarity and subjectivity of every tweets tweeted in different time zones. With the help of this heatmap we can analyse the Community detection algorithms which can be used to find sets of nodes in a graph that have a greater density of connections within their set compared to across sets. The combination of community identification and topic modeling appears to be a suitable technique to describe Twitter communities for the purposes of public health opinion surveillance.

8. Results:

As we see the polarity and subjectivity in the above two graphs now we can know the final result using the lexicon scores which can be known using the below statements. By the results we got from the above plotting we can know the different sentiments from different tweets, tweeted by different people in various accounts.
This project is involved in identifying health issues within a community, so in this the issue is the hpv vaccination reach; gathering data about the community, the target group, and the health concern, here we got the JSON data given by the professor; analyzing data; and assessing community sentiments which are retrieved and visualized using the bar graphs present in the implementation section.



```
# Each word in the lexicon has scores for:
# 1)     polarity: negative vs. positive    (-1.0 => +1.0)
# 2) subjectivity: objective vs. subjective (+0.0 => +1.0)
```

9. Conclusion and Future works:

In the conclusion we got the result of the tweets are happy by polarity we can say that as positive tweets and the least tweets we got are negative tweets.

In this project we fulfilled all the objectives given in the document, which are developing a web-based UI with sentiment and community analysis along with plotting different graphs for polarity and subjectivity.

In Future We shall be dealing with the intensity of lexicons in the future, that is, words that affect the meaning of subsequent words, as well as linguistic interpretations.

10. Reference:

1. Forman D, de Martel C, Lacey CJ, Soerjomataram I, Lortet-Tieulent J, Bruni L, et al. Global burden of human papillomavirus and related diseases. Vaccine 2012 Nov 20;30 Suppl 5:F12-F23.
2. Tabrizi SN, Brotherton JML, Kaldor JM, Skinner SR, Liu B, Bateson D, et al. Assessment of herd immunity and cross-protection after a human papillomavirus vaccination programme in Australia: a repeat cross-sectional study. Lancet Infect Dis 2014 Oct;14(10):958-966.
3. Brotherton JML, Fridman M, May CL, Chappell G, Saville AM, Gertig DM. Early effect of the HPV vaccination programme on cervical abnormalities in Victoria, Australia: an ecological study. Lancet 2011 Jun 18;377(9783):2085-2092.
4. Madden K, Nan X, Briones R, Waks L. Sorting through search results: a content analysis of HPV vaccine information online. Vaccine 2012 May 28;30(25):3741-3746.
5. Robbins SCC, Pang C, Leask J. Australian newspaper coverage of human papillomavirus vaccination, October 2006-December 2009. J Health Commun 2012;17(2):149-159.
6. Mason BW, Donnelly PD. Impact of a local newspaper campaign on the uptake of the measles mumps and rubella vaccine. J Epidemiol Community Health 2000 Jun;54(6):473-474
7. Hoffman SJ, Tan C. Following celebrities' medical advice: meta-narrative analysis. BMJ 2013 Dec 17;347(dec17 14):f7151.
8. Betsch C, Renkewitz F, Betsch T, Ulshöfer C. The influence of vaccine-critical websites on perceiving vaccination risks. J Health Psychol 2010 Apr;15(3):446-455.
9. Signorini A, Segre AM, Polgreen PM. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. PLoS One 2011;6(5):e19467

10. Paul MJ, Dredze M, Broniatowski D. Twitter improves influenza forecasting. PLoS Curr 2014;6:-
11. Mocanu D, Baronchelli A, Perra N, Gonçalves B, Zhang Q, Vespignani A. The Twitter of Babel: mapping world languages through microblogging platforms. PLoS One 2013;8(4):e61981.
12. Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM. Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter. PLoS One 2011;6(12):e26752.
13. Eichstaedt JC, Schwartz HA, Kern ML, Park G, Labarthe DR, Merchant RM, et al. Psychological language on Twitter predicts county-level heart disease mortality. Psychol Sci 2015 Feb;26(2):159-169.
14. Scanfeld D, Scanfeld V, Larson EL. Dissemination of health information through social networks: Twitter and antibiotics. Am J Infect Control 2010 Apr;38(3):182-188.
15. Salathé M, Khandelwal S. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLoS Comput Biol 2011 Oct;7(10):e1002199.
16. Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PLoS One 2010;5(11):e14118.
17. Zhou X, Coiera E, Tsafnat G, Arachi D, Ong M, Dunn AG. Using social connection information to improve opinion mining: identifying negative sentiment about HPV vaccines on Twitter. Stud Health Technol Inform 2015;216:761-765.
18. Dunn AG, Leask J, Zhou X, Mandl KD, Coiera E. Associations between exposure to and expression of negative opinions about Human Papillomavirus vaccines on social media: an observational study. J Med Internet Res 2015;17(6):e144.
19. Mahoney LM, Tang T, Ji K, Ulrich-Schad J. The digital distribution of public health news surrounding the Human Papillomavirus Vaccination: a longitudinal infodemiology study. JMIR Public Health Surveill 2015;1(1):e2.
20. Oliver JE, Wood T. Medical conspiracy theories and health behaviors in the United States. JAMA Intern Med 2014 May;174(5):817-818.
21. Weng L, Menczer F, Ahn Y. Virality prediction and community structure in social networks.

22. Girvan M, Newman MEJ. Community structure in social and biological networks. Proc Natl Acad Sci U S A 2002 Jun 11;99(12):7821-7826

23. Newman MEJ. Detecting community structure in networks. The European Physical Journal B - Condensed Matter 2004 Mar 1;38(2):321-330.