BREAST CANCER SENTIMENT ANALYSIS USING TWITTER

TEAM 8

AKSHAY LAVHAGALE
EJONA KOCIBELLI
PRIYANKABEN SHIYANI
SAGAR TANNA

Table of Contents

1. INTRODUCTION

1.1. Motivation behind the topic selection

Breast cancer is the most common cancer among women in the United Locations with a high percentage of 12% risk of developing the cancer over their lifetime course. To detect the breast tumors in an early stage making it easier to prevent it is important for women to have regular breast screenings. The more women are aware of breast cancer and follow the screening guidelines that are recommended for the breast cancer, the more the benefits are and on the other hand failing to do so can be harmful and unfortunately sometimes too late to prevent and recover. We just passed October month else known as “The Breast Cancer Awareness Month” and as a team we thought that this project will be beneficial not only to us while working on it, but also, we aim that the results will be beneficial to raise awareness in this critical topic.

1.2. The importance of social media platforms. Why Twitter out of all platforms?

Nowadays, social media is widely used to share information not only by individual but also by many healthcare stakeholders. It plays a pivotal role in communicating information. There are many social media platforms that provide information that we could extract from. *So, why Twitter?* Approximately 70% of cancer cases occur in women 45 years and older. Thus, taking this in consideration, the data we are looking for will mostly be perceived by women in this age-group. Among all the social media platforms, Twitter was the platform that had the highest percentage of 41% usage by people in the age group of 45 years and older. So, we decided to use Twitter as our data source platform.

1.3. Project scope

The aim of our project is:

- Analyze people’s attitudes (i.e., emotions) towards breast cancer.
- How do people’s attitudes towards breast cancer differ across geospatial regions?
- Can we identify latent topic trends and hashtags in breast cancer-related tweets?

To come to conclusions, we will identify and collect all the tweets that are related to breast cancer using hashtags such as #breastcancer, #womenbreastcancer, #breastcancerawareness. We will model the information gathered and then apply various sentiment analysis techniques that will help us better understand the perceptions and emotions that the Twitter users have towards breast cancer.

1.4. What is sentiment analysis?

To have a better understanding of sentiment analysis, firstly we need to understand what a sentiment is. A sentiment is an emotion or an opinion towards something. It can be a topic or something in general. This can either be positive, negative, or neutral. This is known as sentiment. Now, when we understand the sentiment of masses, with a “score” attached to it, then it is called sentiment analysis.

1.5. The technology stack

The technologies we used are Tweepy/Twitter, PySpark, AmazonS3, Amazon EMR, Python, TextBlob, Tableau.

2. DATA COLLECTION

The data we used in this study to perform sentiment analysis on breast cancer tweets is a collection of breast cancer related tweets from the last three months: September, October, November taking in consideration tweets posted within the United States only. Firstly, we applied, and we were approved for a Twitter Developer account that we used to utilize Tweepy. Tweepy is an open-source library which enabled us to gather all the tweets related to breast cancer automatically through the Twitter API.

We use a list of total 10 hashtags to cover all the possible data regarding breast cancer:

#breastcancer #breastcancersurvivor #goingflat #breastcancerawareness #mastectomy #busylivingwithmets #bccww #BCSM #breastcancerawarenessmonth #breastcancerwarrior

Datasets Summary:

- The total dataset we collected has 17887 rows and 9 columns
- Monthly datasets:
    - September dataset has 5472 rows
    - October dataset has 6884 rows
    - November dataset 5519 rows

The columns refer to:

- Tweet ID – The unique id of each tweet
- Tweet Date – Time and date that the tweet was posted
- Tweet Content – The original uncleaned content of each tweet
- Sentiment – The score given to each tweet using the Textblob library
- Language – The language of the tweet
- Retweet Count – The number of retweets that the tweet had
- Hashtags – The hashtags that were included in the tweet content
- Location – The location where the tweet was posted
- Classes – A categorical variable that classified tweet into 0,1,2 categories which corresponded to Negative, Neutral and Positive sentiment, respectively. This variable was the target variable for the machine learning model

We used this data to:

- Train the machine learning model
- Analyze the trend of opinion of masses on breast cancer
- Analyze how do people's attitudes towards breast cancer differ across geospatial regions
- Identify latent topic trends and hashtags in breast cancer related tweets

3. DATA PREPROCESSING AND CLEANING

Once we collected all the data that we needed for sentiment analysis, we preprocessed the data by eliminating the duplicates, and by eliminating all the data where the location was not in the United States and language was not in English. The initial data collected was not structured. It was raw, noisy, and needed to be cleaned before we used it for modelling. Preprocessing and cleaning the data is critical step because the higher the quality of the data, the more reliable the results are. These processes involved a series of tasks such as removing all the irrelevant information such as emojis, special character and extra blank space. We improved the format, deleted all the duplicated tweets, or tweets that were not valid i.e., shorter than three characters. Figure 1 shows the raw form of the tweet content and the cleaned tweet after the data processing.

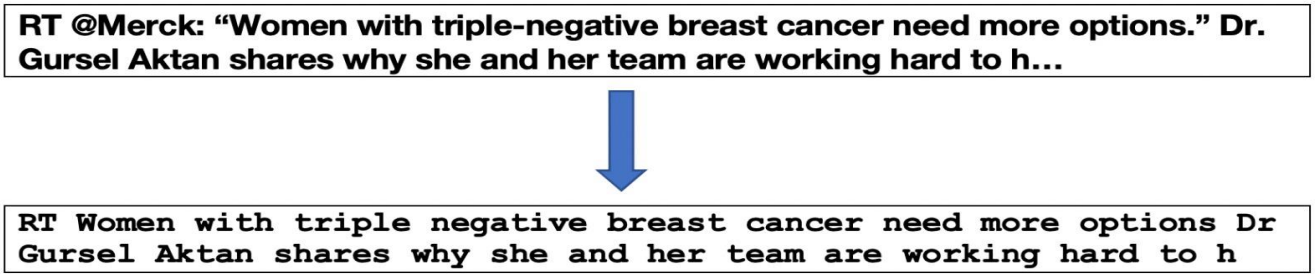

Figure 1. Tweets before and after the cleaning data processing

We utilized TFIDF Vectorizer to convert textual data of the tweet to numeric form, so that it can be further used in the modelling process. Figure 2 shows the head of the data after it was processed through TFIDF Vectorizer.

```
+--------------------+--------------------+--------------------+------------+----+--------------+-------------+-------
--------+-------------+-------+--------------------+--------------------+--------------------+-----+
|          created_at|       original_text|           polarity|subjectivity|lang|favorite_count|retweet_count|origina
l_author|        place|classes|              tokens|                  tf|            features|label|
+--------------------+--------------------+--------------------+------------+----+--------------+-------------+-------
--------+-------------+-------+--------------------+--------------------+--------------------+-----+
|Mon Sep 07 19:28:...|Thanks for sharin...|0.35714285700000004| 0.441666667|  en|             0|            0|     Ma
ssiveBio|Nashville, TN|      1|[thanks, for, sha...|(65536,[463,9639,...|(65536,[463,9639,...|  0.0|
|Mon Sep 07 19:28:...|I had a fantastic...|                 0.7|         0.9|  en|             5|            0|
DrTashaG|Nashville, TN|      1|[i, had, a, fanta...|(65536,[3975,5232...|(65536,[3975,5232...|  0.0|
|Mon Sep 07 19:28:...|New research publ...|0.21666666699999998| 0.443939394|  en|             1|            2|      W
ileyNews|Nashville, TN|      2|[new, research, p...|(65536,[8436,9863...|(65536,[8436,9863...|  1.0|
|Mon Sep 07 19:28:...|Thank you for pro...|                -0.2|         0.0|  en|             1|            0|     Ma
ssiveBio|  Seattle, WA|      0|[thank, you, for,...|(65536,[1386,9639...|(65536,[1386,9639...|  2.0|
|Mon Sep 07 19:28:...|RT Our scientist ...|                 0.0|         0.0|  en|             0|           17|
Craybo23|Nashville, TN|      1|[rt, our, scienti...|(65536,[9616,1143...|(65536,[9616,1143...|  0.0|
+--------------------+--------------------+--------------------+------------+----+--------------+-------------+-------
--------+-------------+-------+--------------------+--------------------+--------------------+-----+
```

Figure 2. Data after processed through TFIDF Vectorizer

4. DATA UNDERSTANDING

Now that the data is structured and clean, the next step is exploring the data and analyzing it for any patterns. This will help us to have a better understanding of the data. These are some of the observations that we concluded by structuring and analyzing the data:

- Out of the three months that we collected the data, October was the month with the most data, and this can be due to October being the Breast Cancer Awareness month.
- The total dataset results in Neutral sentiments dominating the Negative or Positive sentiments. (shown in Figure 3)
- The monthly datasets: (shown in Figure 4)
    - September and November – Neutral sentiments dominate by 41.89% and 56.22%
    - October – Positive sentiments dominate by 39.21%

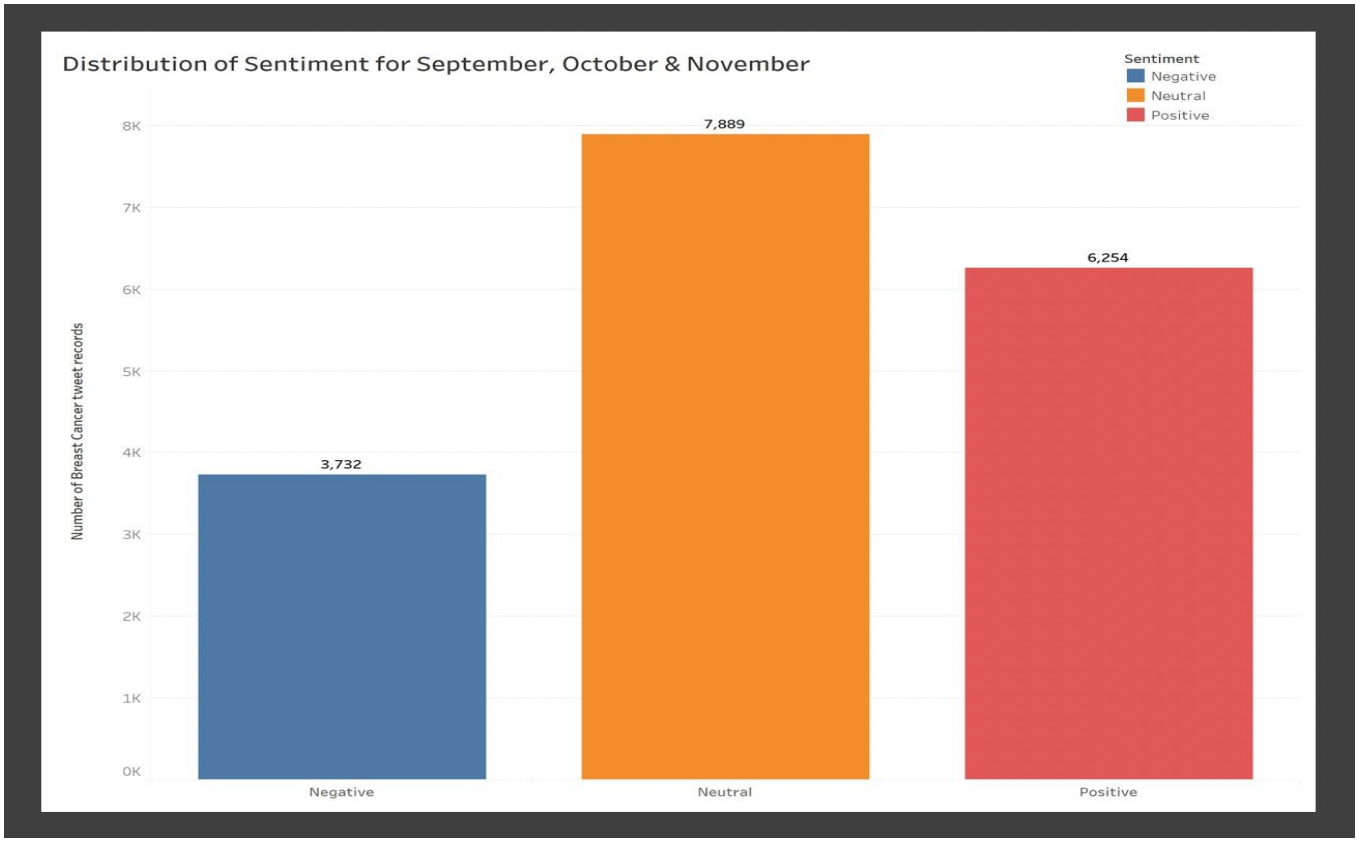

Figure 3. Total data sentimental observations

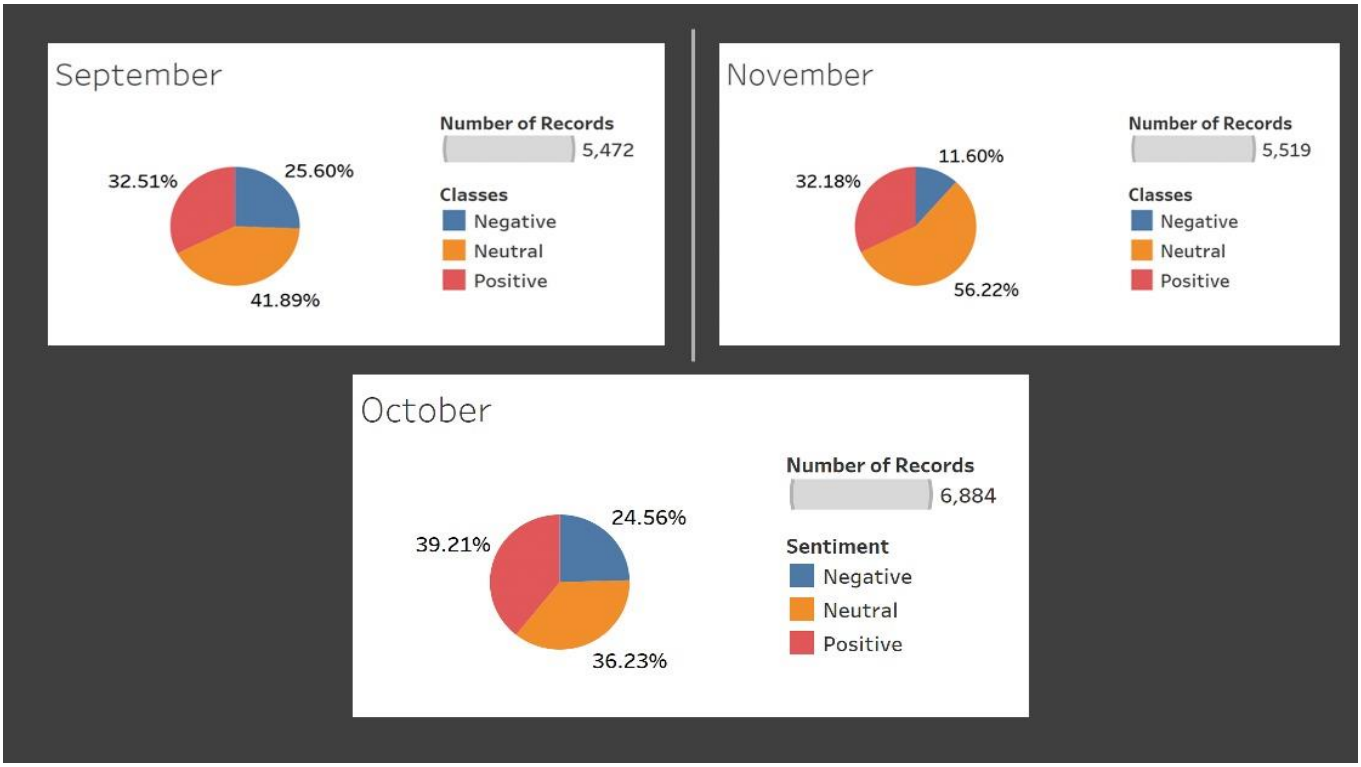

Figure 4. Monthly data sentimental observations

5. DATA ANALYSIS

In data analysis, we get into the modeling process and the process of building a classifier to predict whether the tweet sentiment is negative, neutral, or positive. We decided to run our collected data through three popular algorithms used for classification: *Logistic Regression, Decision Trees* and *Random Forests*. We also created four different AWS EMR clusters: *1 instance (no parallel processing)*, *3 instances, 5 instances* and *7 instances* to assess the performance of our models using parallel processing as compared to different clusters with different number of instances but also compared it with using a single processor (1 instance cluster). We then fetched data from the bucket we had created with Amazon S3 services. We had stored September, October and November datasets on it and uploaded these datasets into a notebook instance running on our cluster. We followed the preprocessing steps of cleaning the data as we mentioned in the previous section and once, we have TFIDF vectors, we are ready to run this through our models.

One important performance measure is to see the time vs scale and accuracy vs scale comparisons. To do this, we split our data into training and test data four times each time with a different train size. We split the data into 30%, 50%, 70% and 90% of the data and the rest we considered as test data. We imported the Pyspark modules of our three algorithms and ran our training datasets with its features (tf-idf vectors) with the labels (0,1,2 classes) on these models. We noted down the accuracies and the time taken to run these steps. We terminated the cluster (3 instances) and repeated the same steps over by creating other clusters: 1 instance cluster (no parallel processing), 5 instances cluster, 7 instances cluster. The results will be shown in the next section.

6. PERFORMANCE EVALUATIONS

6.1. Measuring the impact of scale on quality of analysis

To measure the impact of scale on quality of analysis, we created graph with accuracy vs scale for our all three algorithms. The algorithm we are using are Logistic Regression, Decision Tree and Random Forest. In scale, we split the data into 30%, 50%, 70% and 90% of the data and the rest we considered as test data. Then we got the accuracy results for each data and plotted in a graph to have a better understanding on accuracy by visualizing it. From graph, it is noticeable that Logistic Regression is greater than Decision Tree and Decision Tree is greater than Random Forest. The Reason behind that is the performance of Random Forest will be low when the number of features is less in the data set. In our breast cancer data set we only

have one feature ‘original text’. On the other hand, Logistic regression is known to perform better with lower dimensional data and also, Random forests are an ensemble of weak decision trees. But still together they perform better than just a one single decision tree. Overall, we had highest accuracy by using Logistic Regression algorithm.

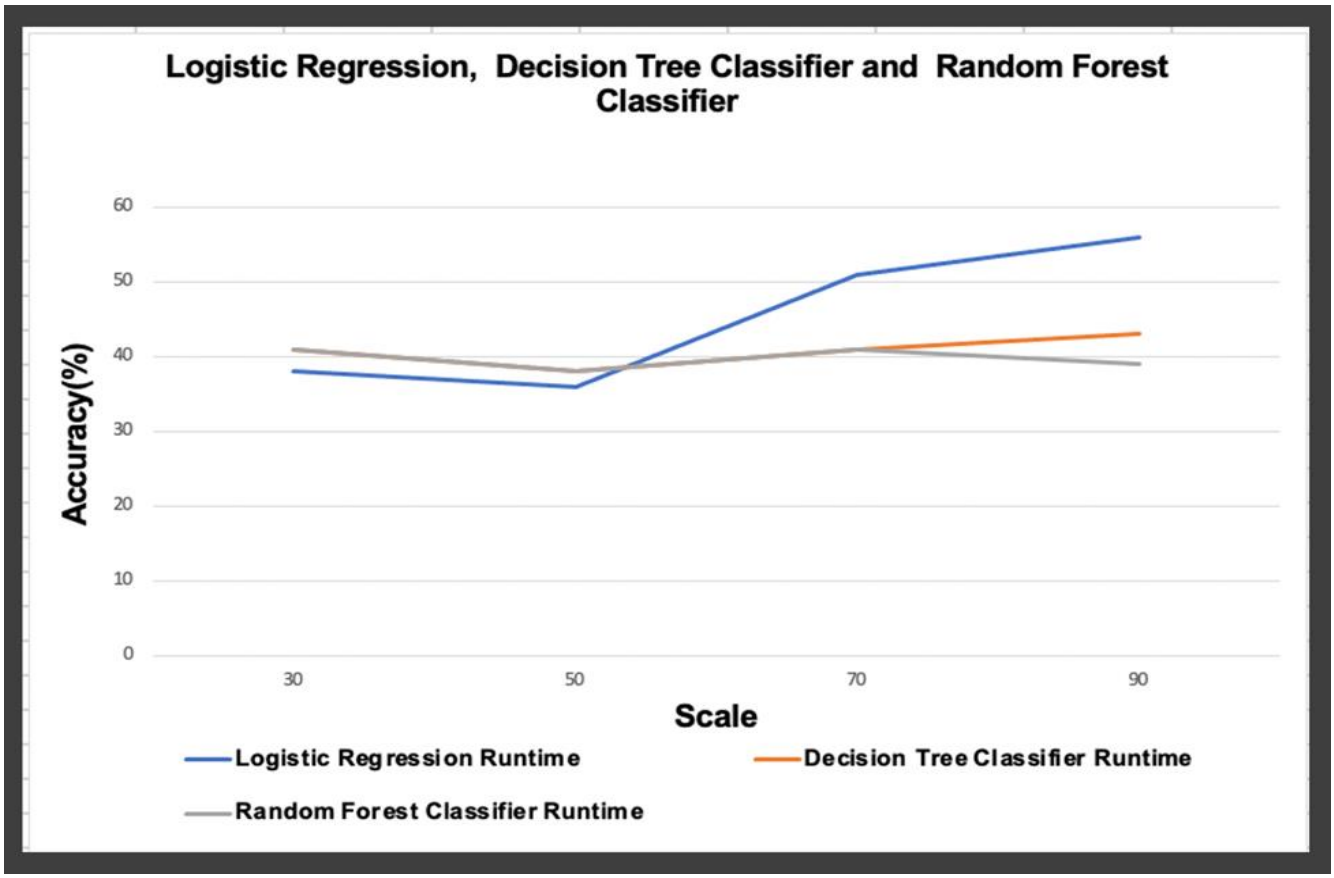

Figure 5. Accuracy graph for all the three algorithms

6.2. Measuring the impact of scale on performance

To measure the impact of scale on performance, we created graph with time vs in scale for our all three algorithms to see which algorithm performs the fastest and slowest. So, from graph we can see that Logistic regression are much faster than decision trees classifier and Random forest classifier. We split the dataset into subsets of data 30%, 50%, 70% and 90% to compare the time performance of each subset of data using the three algorithms. Then we plotted the results that we got to create the graph to help us visualize and compare the results. So, for instance, when we ingested the 30% of data into each algorithm then we noticed that time is between 10 to 15 seconds in all algorithms but after ingesting the 90% of data into each

algorithm we noticed that the time difference increased between all three algorithms. Decision trees classifier's performance is slowest in all three algorithms.

Figure 6. Running time comparisons for 30%, 50%, 70% and 90% of data for all algorithms

6.3. Measure the impact of parallel computation on performance

To measure the impact of parallel computation on performance, we used AWS EMR cluster with 1 instance cluster (no parallel processing), 3 instance cluster, 5 instance cluster, 7 instance cluster and based on result we got, we generated 4 graphs. In the graphs we can see that using cluster with 1 instance cluster (no parallel processing) and 3 instance cluster took more time than other clusters with 5 and 7 instances cluster. The cluster with 7 instances gave the better performance and it took less time compare to other three clusters.

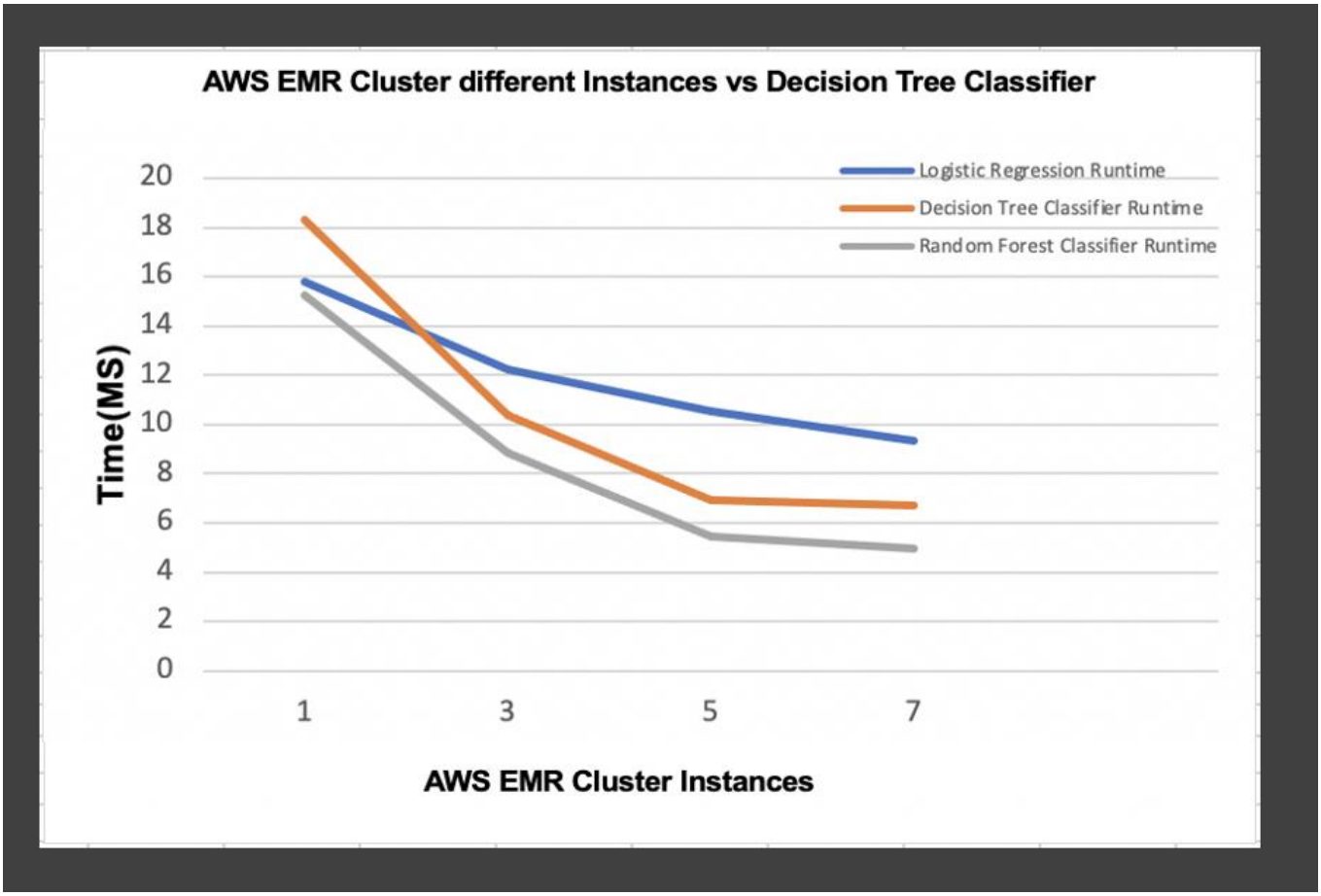

Figure 7. Running time in non-parallel cluster vs 3 instances, 5 instances and 7 instances clusters

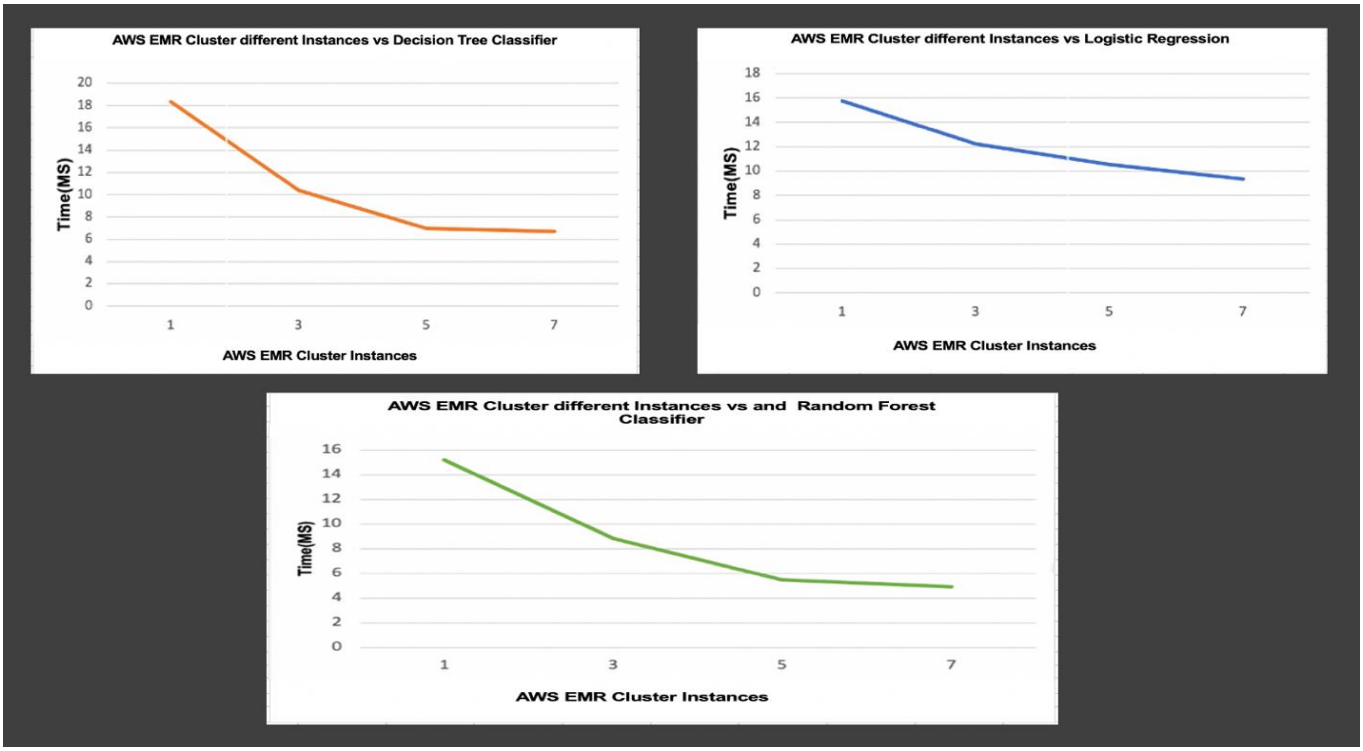

Figure 8. Running time in non-parallel cluster vs 3 instances, 5 instances and 7 instances clusters for each algorithm

7. SCOPE EVALUATIONS

7.1. Analyze people's attitudes (i.e., emotions) towards breast cancer

Based on the graph below, we can better analyze the people's attitudes towards breast cancer. We selected these three particular months for a good reason as October is the month for the Breast Cancer Awareness and we wanted to see if that had an impact in the data, and it actually did. As we notice in the Figure below, during September month Neutral sentiments were dominating while as we go during the October month Neutral sentiments decrease while Positive and Negative sentiments increase, and Positive sentiments dominate. As November month comes by Neutral Sentiments increase and dominate again while we have a noticeable decrease in Positive and Negative sentiments. The previous sections cover the sentiment analysis of people's attitudes in details.

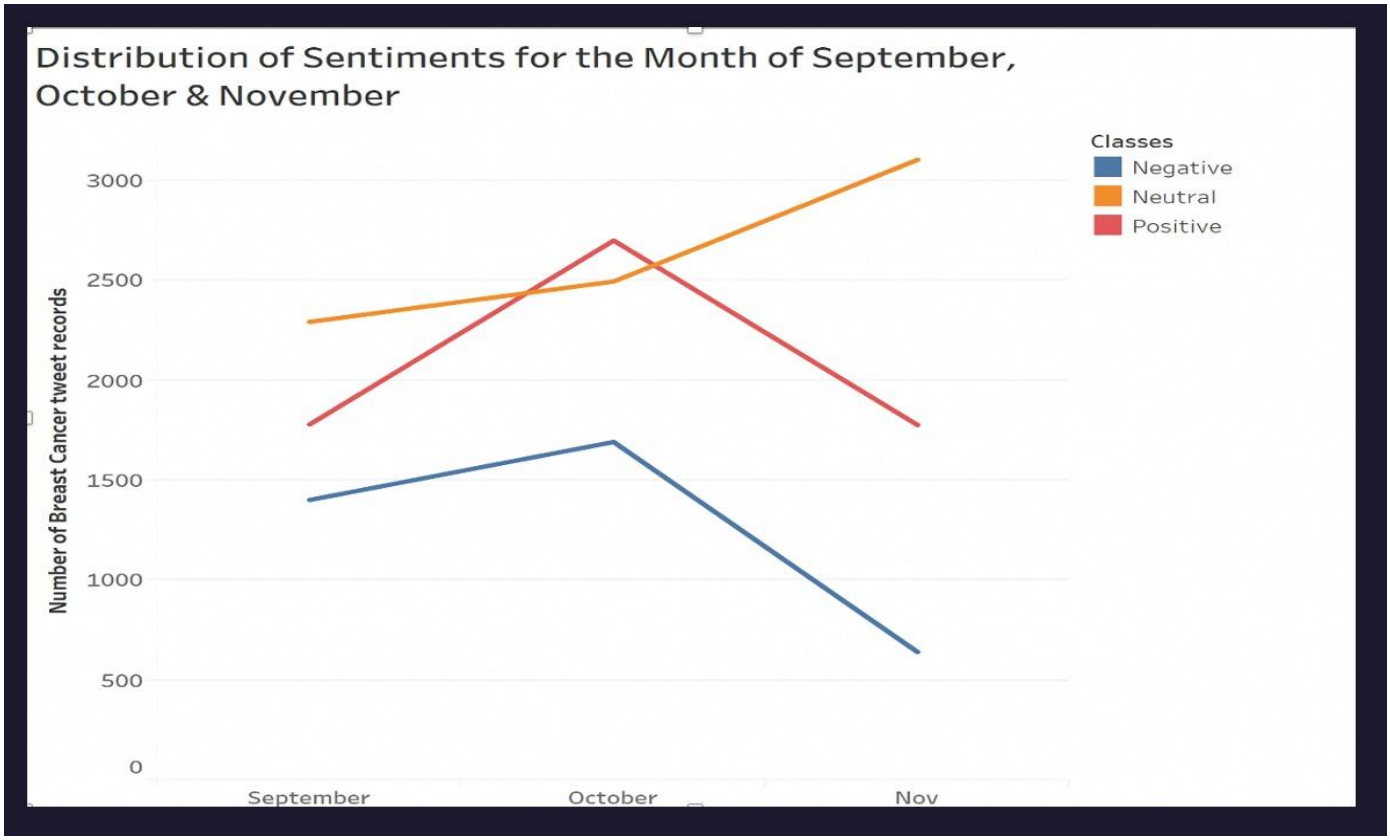

Figure 9. People's attitude graph showing Positive, Negative and Neutral Sentiments for the whole dataset

7.2. How do people's attitudes towards breast cancer differ across geospatial regions?

To be able to come to the conclusions of the graphs below, we used the location of the dataset to determine which are the top 10 locations that have the highest number of tweets and how was the people's emotions in each leading location. We firstly calculated for the whole dataset, and then repeated the same process for each month. The first graphs below provide this information:

- The top 10 locations for the whole three months where San Francisco and New York lead in the list.
- The amount of how much of the data was Negative, Neutral, and Positive for each location. We can say that in San Francisco, Positive sentiments were dominating with just a few more tweets than Neutral sentiments, while in New York it is a different scenario. The Neutral sentiments are dominating with quite a difference from the Positive and Negative ones.

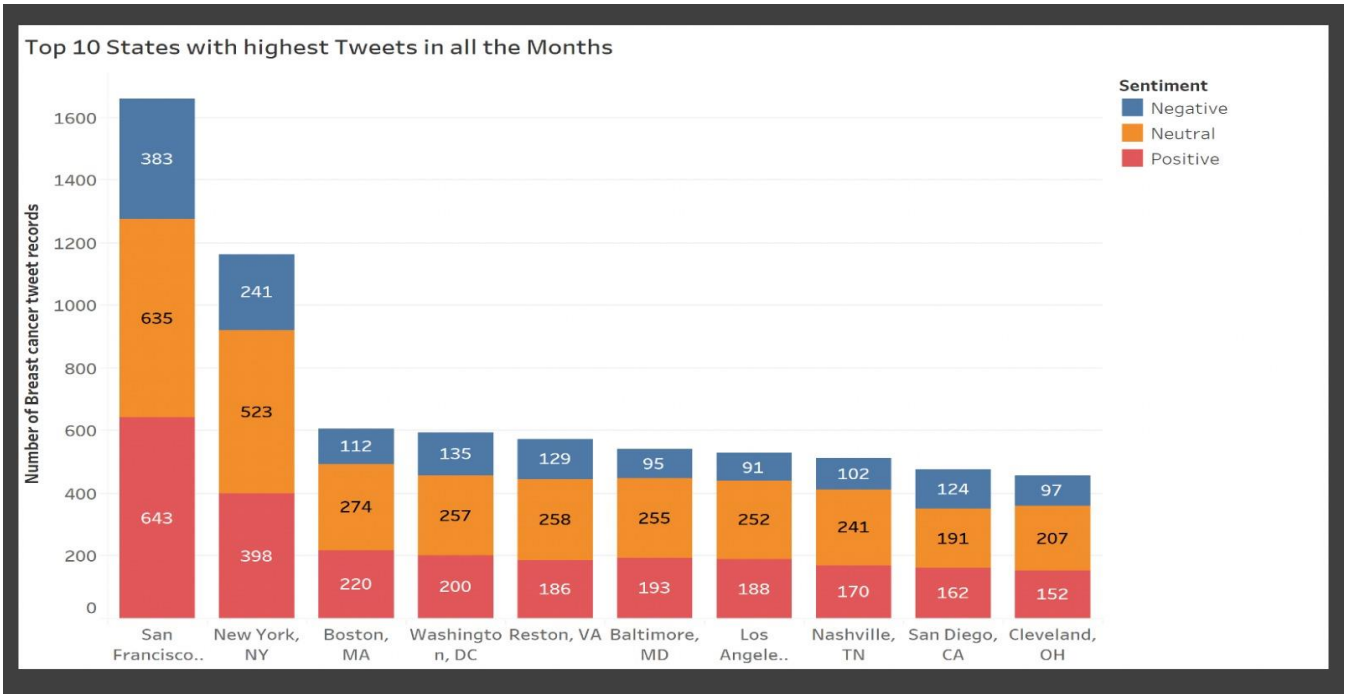

Figure 10. Different people's attitudes towards breast cancer across geospatial regions

The second graph is similar to the first graph but gives more insights for the monthly top 10 leading locations:

- September Month: New York is the first leading location with Neutral sentiments dominating, San Francisco follows with Neutral sentiment dominating. The graph gives the whole top 10 list.
- October Month: San Francisco is leading with Positive sentiments dominating, New York follows with Neutral sentiment dominating just with tweets from Positive sentiment. Noticeable is the increase of the number of tweets on this month.
- November Month: New York is the first leading location with Neutral sentiments dominating, Washington D.C. follows with Neutral sentiment dominating leaving behind San Francisco.

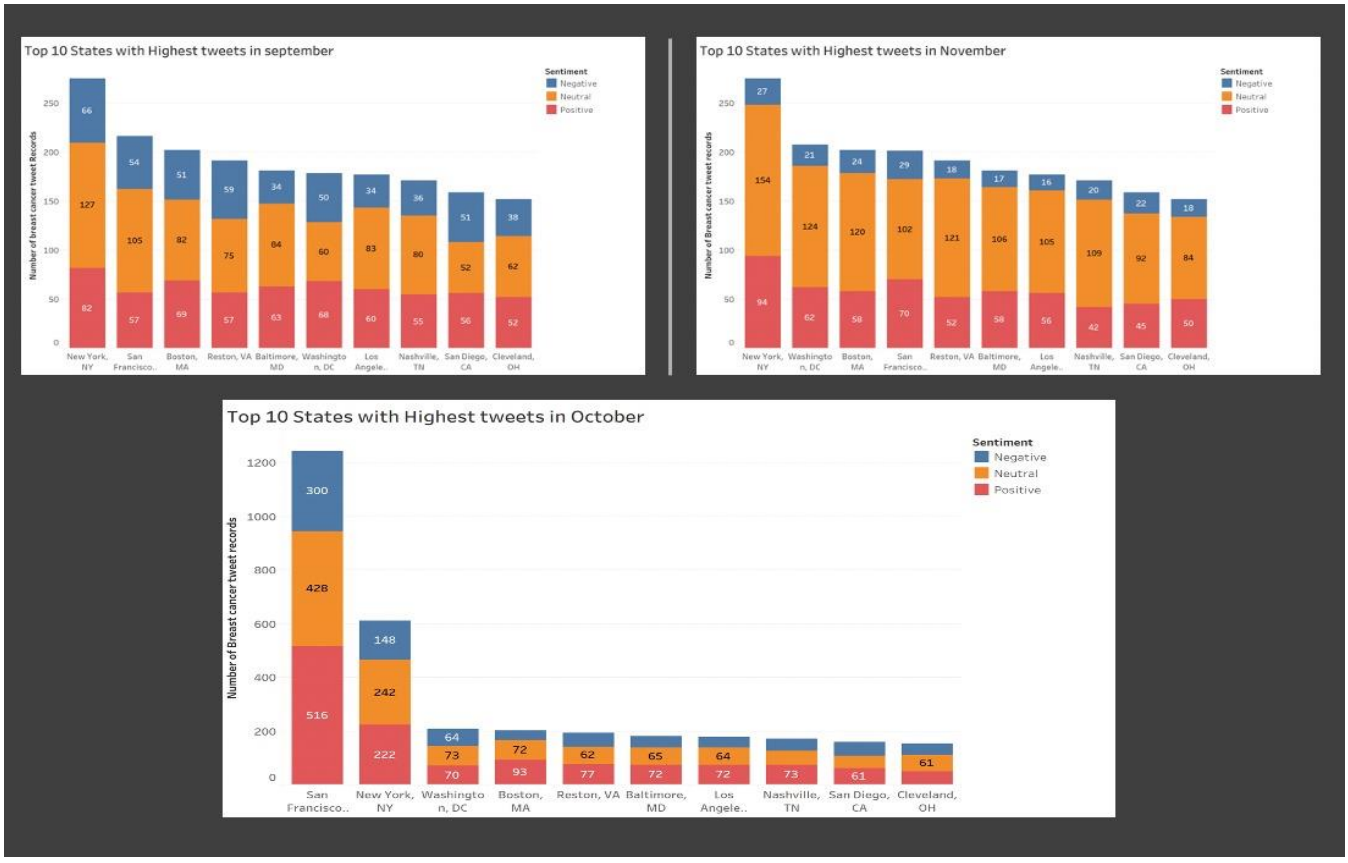

Figure 11. Top 10 leading locations and people's emotions for each month

7.3. Can we identify latent topic trends and hashtags in breast cancer-related tweets?

To come to conclusion for this question we took in consideration two pieces of data. Firstly, we considered hashtags in where we calculated the total number of each hashtag and then we sorted from the hashtag with the highest number to the lowest. Secondly, we considered the number of retweets each tweet. The higher number of retweets a tweet, the trendier we considered. Lastly, we combine these together and the results are as shown in the graphs below.

The first graph provides this information:

- The top trending hashtags with the highest number of retweets is #breastcancer, followed by #BreastCancer, #SABCS20, #bcsm and so on. The #breastcancer hashtag has the highest number of retweets and noticeable way more then other hashtags.

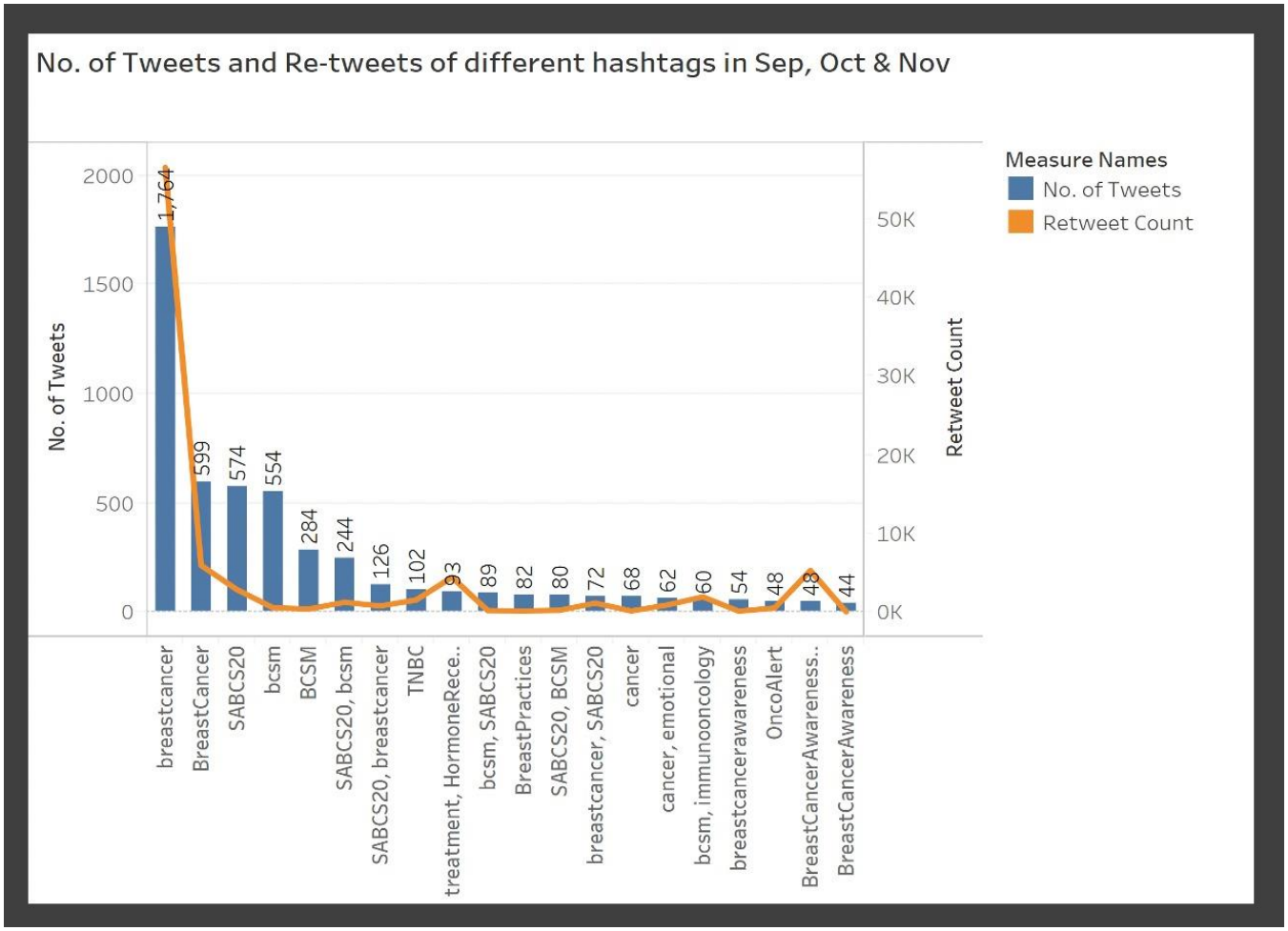

Figure 12. Leading hashtags and their correspondent retweets count for the total dataset

The graphs in the Figure 13, provide the trending hashtags for each month:

- September: Leading hashtags are: #breastcancer #BreastCancer #bcsm, #SABCS20.
- October: Leading hashtags are: #breastcancer, #bcsm, #SABCS20, #BreastCancer
- November: Leading hashtags are: #breastcancer, #SABCS20, #BreastCancer, #bcsm

Note: The graph shows the full list of trending hashtags and their correspondent number of retweets.

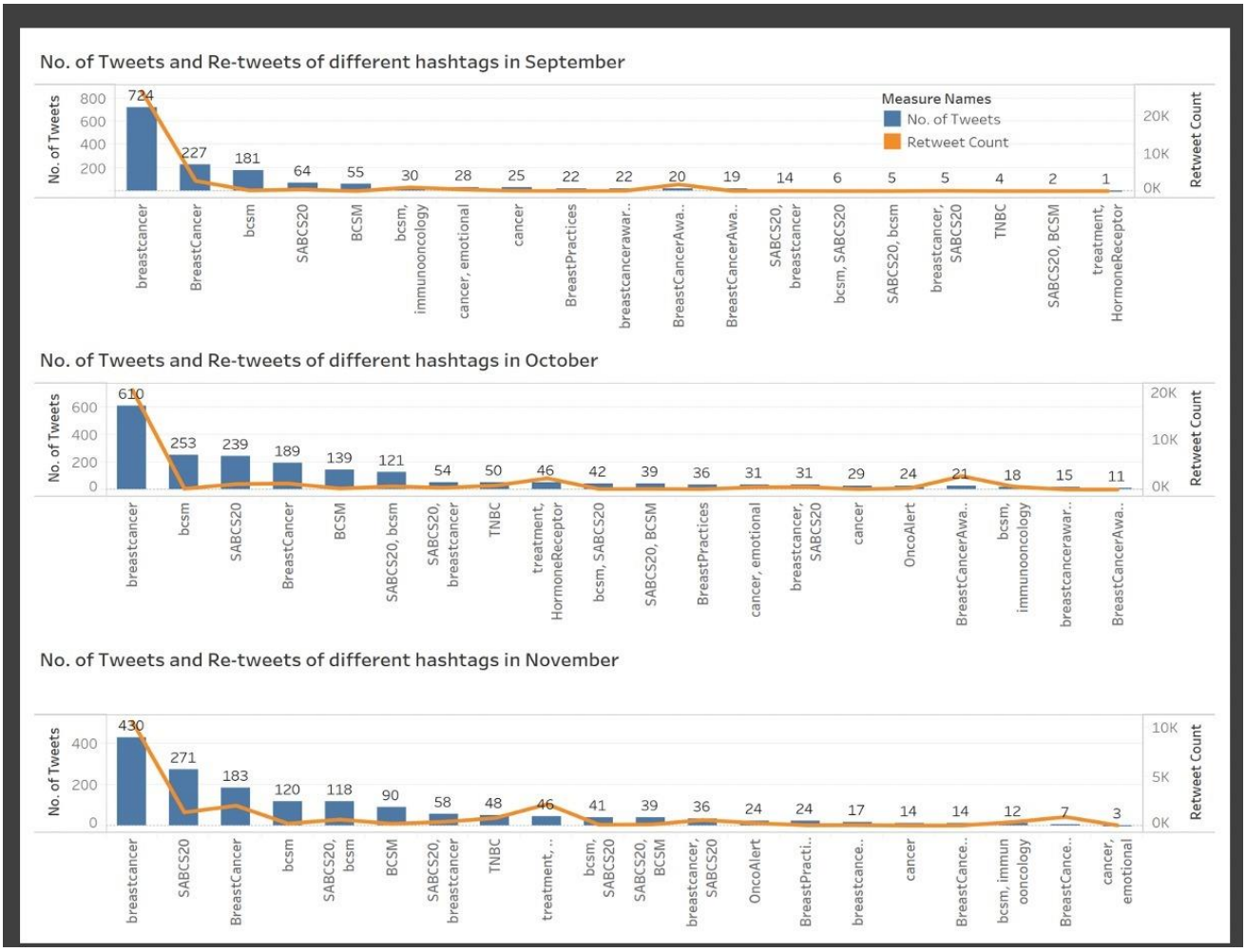

Figure 13. Leading hashtags and their correspondent retweets count for each month

7. CONCLUSION

In conclusion, we learned the following:

We identified and collected all the tweets which are related to breast cancer using hashtags. After successfully cleaning and processing the data, we modeled the information gathered and then we applied various sentiment analysis techniques which helped us to get better understanding for the perceptions and emotions which the Twitter users have towards breast cancer.

Performance wise we concluded:

- From all the three algorithm we used, Logistic Regression gave us the highest accuracy.
- The more instances in the cluster, the less time to run the data so the higher performance.
- The more data we added (30%, 50%, 70%, and 90%), the more time it took to run the data.

We successfully analyzed and came to conclusions our scope related:

- The metropolitan areas or super cities such as San Francisco and New York have the highest reactions, the highest number of tweets. While suburbs and not as big and developed cities stand behind.
- The vast majority of people concluded in Neutral emotions towards breast cancer over all the three months period, but you could sense the difference that October played in increasing the positive emotions.
- The trending hashtags for breast cancer are the classical ones and they continue to lead over the months: #breastcancer, #BreastCancer. These are the hashtags that also have the highest number of retweets.

Overall, we learned a lot while working on this project as we practiced all the technologies and information that we have covered through this course.

8. REFERENCES

[1]. How to collect tweets from the Twitter Streaming API using Python

[2]. How to Extract Tweets from Twitter in Python

[3]. Sentiment Analysis with PySpark

[4]. Machine Learning with PySpark and MLlib — Solving a Binary Classification Problem

[5]. Sentiment Analysis: Definition, Uses, Examples + Pros /Cons

[6]. Understanding Perceptions and Attitudes in Breast Cancer Discussions on Twitter