**Predictive Analysis using RapidMiner for Twitter data**

From the text of all the social network posts, the following list of tweets can be compiled which are used the most. First 20 hashtags that are used the most are listed below, along with the number of times mentioned.

|  |  |
| --- | --- |
| #mikebrownfuneral | 911 |
| #tntweeters | 728 |
| #blacktwitter | 555 |
| #tcot | 400 |
| #cir | 333 |
| #mikebrown | 280 |
| #lgbt | 241 |
| #ferguson | 192 |
| #elections2014 | 179 |
| #missing | 92 |
| #eff | 84 |
| #yahoo | 76 |
| #cnn | 75 |
| #stl | 73 |
| #blacklivesmatter | 72 |
| #trayvonmartin | 72 |
| #bbcnews | 70 |
| #bcot | 68 |
| #emmys2014 | 68 |
| #wits | 56 |

While the total data is close to 10000 rows, the hashtag information above is comparatively less, due to the way unstructured data is processed. During textmining, a majority of data, including special symbols used in tweets and junk characters in the text that were a result of data rendering are ignored. The list above is compiled only using all the text available.

Using the above hash tag information, data is mined with all the user information included, time and other details etc. This gives the accurate count of hashtags.

|  |  |
| --- | --- |
| Hashtag | Total |
| #mikebrownfuneral | 3901 |
| #tntweeters | 728 |
| #blacktwitter | 3900 |
| #tcot | 399 |
| #cir | 333 |
| #mikebrown | 5868 |
| #lgbt | 245 |
| #ferguson | 1734 |
| #elections2014 | 179 |
| #missing | 92 |
| #eff | 84 |
| #yahoo | 76 |
| #cnn | 76 |
| #stl | 92 |
| #blacklivesmatter | 183 |
| #trayvonmartin | 81 |
| #bbcnews | 70 |
| #bcot | 68 |
| #emmys2014 | 131 |

Looking at the list again, #mikebrown is used in maximum number of posts, 5868, followed closely by #mikebrownfuneral, and #blacktwitter which are mentioned 3901 and 3900 times respectively.

List of tweets segregated by date -

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **Hashtag** | **On 8/25** | **On 8/26** | **On 8/27** | **Total** | **Avg. tweets/day** | | #mikebrownfuneral | 1152 | 2611 | 138 | 3901 | 1300 | | #tntweeters | 259 | 388 | 81 | 728 | 243 | | #blacktwitter | 1242 | 1653 | 1005 | 3900 | 1300 | | #tcot | 266 | 114 | 19 | 399 | 133 | | #cir | 104 | 199 | 30 | 333 | 111 | | #mikebrown | 2894 | 2779 | 195 | 5868 | 1956 | | #lgbt | 89 | 131 | 25 | 245 | 82 | | #ferguson | 812 | 760 | 162 | 1734 | 578 | | #elections2014 | 58 | 88 | 33 | 179 | 60 | | #missing | 32 | 49 | 11 | 92 | 31 | | #eff | 44 | 40 | 0 | 84 | 28 | | #yahoo | 23 | 49 | 4 | 76 | 25 | | #cnn | 48 | 28 | 0 | 76 | 25 | | #stl | 66 | 26 | 0 | 92 | 31 | | #blacklivesmatter | 22 | 127 | 34 | 183 | 61 | | #trayvonmartin | 74 | 4 | 3 | 81 | 27 | | #bbcnews | 42 | 28 | 0 | 70 | 23 | | #bcot | 48 | 20 | 0 | 68 | 23 | | #emmys2014 | 34 | 89 | 8 | 131 | 44 | |  |  |  |
|  |  |  |  |

Looking at the hashtag counts, it can be observed that hashtags are mentioned the most on Day 2 (08/26), while a substantial number of tweets are mentioned on Day 1, the number subsided by Day 3. There is a lot of other data available in the provided document, however, data is based on account information which is inturn based on activity of the user on that day. While we can consider other values as well, like Reach and original reach, we are only considering occurrences of hashtags, as they can exist independent of a specific user. It will be a complex phenomenon to estimate if a user shows the same level of participation every day. Hence, we are only considering the occurrence of hashtags.

With the available data as above, we can see what hashtags went viral and not. Thus, the compiled data can be compiled as below –

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Hashtag** | **Day 1** | **Day 2** | **Day 3** | **Propagation** |
| #mikebrownfuneral | 1152 | 2611 | 138 | Viral |
| #tntweeters | 259 | 388 | 81 | Moderate |
| #blacktwitter | 1242 | 1653 | 1005 | Viral |
| #tcot | 266 | 114 | 19 | Low |
| #cir | 104 | 199 | 30 | Low |
| #mikebrown | 2894 | 2779 | 195 | Viral |
| #lgbt | 89 | 131 | 25 | Low |
| #ferguson | 812 | 760 | 162 | Moderate |
| #elections2014 | 58 | 88 | 33 | Low |
| #missing | 32 | 49 | 11 | Low |
| #eff | 44 | 40 | 0 | Low |
| #yahoo | 23 | 49 | 4 | Low |
| #cnn | 48 | 28 | 0 | Low |
| #stl | 66 | 26 | 0 | Low |
| #blacklivesmatter | 22 | 127 | 34 | Low |
| #trayvonmartin | 74 | 4 | 3 | Low |
| #bbcnews | 42 | 28 | 0 | Low |
| #bcot | 48 | 20 | 0 | Low |
| #emmys2014 | 34 | 89 | 8 | Low |

Hashtags that has an average of 1000 times or more per day are determined as ‘Viral’, between 200 and 1000 are determined as ‘moderate’ and less than 200 as ‘low’.

**Predictive Analysis**

Predictive analysis uses trends available in existing data and applies the same to future data to predict the outcome. These datasets are called Training and Scoring. Data produced by above process will be considered as Training dataset.

Consider a future scenario, where in, we have the number of hashtags, then we can use the existing data and predict which tweets can be viral. We can consider the following scoring dataset –

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  | | --- | --- | --- | --- | | **Hashtag** | **Day 1** | **Day 2** | **Day 3** | | #A | 2510 | 59 | 1100 | | #B | 2710 | 1200 | 90 | | #C | 1500 | 900 | 230 | | #D | 650 | 25 | 130 | | #E | 18 | 4500 | 1090 | | #F | 89 | 140 | 240 | | #G | 2100 | 145 | 190 | | #H | 130 | 56 | 84 | | #I | 170 | 210 | 125 | | #J | 340 | 349 | 90 | | #K | 4500 | 2110 | 1750 | | #L | 1200 | 1450 | 140 | | #M | 40 | 90 | 130 | | #N | 150 | 120 | 90 | |  |  |  |  |

After applying Linear Discriminant Analysis to the existing data, prediction about propagation is as follows –

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Moderate | 0.0 | 0.0 | 0.0 | #A | 2510.0 | 59.0 | 1100.0 |
| 2 | Moderate | 0.0 | 0.0 | 0.0 | #B | 2710.0 | 1200.0 | 90.0 |
| 3 | low | 0.0 | 0.0 | 0.0 | #C | 1500.0 | 900.0 | 230.0 |
| 4 | low | 0.0 | 0.0 | 0.0 | #D | 650.0 | 25.0 | 130.0 |
| 5 | Viral | 0.0 | 0.0 | 0.0 | #E | 18.0 | 4500.0 | 1090.0 |
| 6 | low | 0.0 | 0.0 | 0.0 | #F | 89.0 | 140.0 | 240.0 |
| 7 | Moderate | 0.0 | 0.0 | 0.0 | #G | 2100.0 | 145.0 | 190.0 |
| 8 | low | 0.0 | 0.0 | 0.0 | #H | 130.0 | 56.0 | 84.0 |
| 9 | low | 0.0 | 0.0 | 0.0 | #I | 170.0 | 210.0 | 125.0 |
| 10 | low | 0.0 | 0.0 | 0.0 | #J | 340.0 | 349.0 | 90.0 |
| 11 | Viral | 0.0 | 0.0 | 0.0 | #K | 4500.0 | 2110.0 | 1750.0 |
| 12 | low | 0.0 | 0.0 | 0.0 | #L | 1200.0 | 1450.0 | 140.0 |
| 13 | low | 0.0 | 0.0 | 0.0 | #M | 40.0 | 90.0 | 130.0 |
| 14 | low | 0.0 | 0.0 | 0.0 | #N | 150.0 | 120.0 | 90.0 |

Linear Discriminant Analysis groups observations together into like types of values and will predict accordingly. Predictive analysis shows us that hashtag #E and hashtag #K are both predicted to go Viral, whereas #A, #B and #G are predicted as Moderate and others as low. More accurate results can be predicted by having more data to be modeled.