

Oscars 2017 – Text Mining & Sentimental Analysis

Karthik Sripathi, Oklahoma State University, Stillwater, OK

ABSTRACT

It has always been fascinating to realize how the magnitude of award shows have been increasing year after year. It is the enormously positive response of audience that keeps the stage shows to envisage. We know that sentiments of people play a crucial role in deciding the prospects of a particular event. This paper summarizes the sentiments of individuals towards one of the most awards popular show, Oscars. It provides crucial insights on how people sentiments could determine the success or failure of a show. The paper involves text mining of people's reactions towards the 2017 Oscars in general and a sentiment analysis of regarding the best picture mix up using SAS® Sentiment Analysis Studio.

This paper aims to determine the success of an awards show based on individual sentiments before the show, during and after the show. This information will give a better picture of how to handle any unwanted circumstances during the event. We can conclude from the 2017 Oscars that the sentiments of the people were more positive or neutral indicating that the excitement about the show will over shadow any unwanted events. This analysis can be extended to build a text predictive model wherein there is a scope of predicting the sentiments towards unwanted events and will help us to set the stage better and be prepared for potential problems.

INTRODUCTION

Oscars is the most awaited event for every actor including the world audience every year as it leaves memorable moments for them to cherish. Analyzing people's reactions to such an event will give us an understanding of whether the audience enjoyed the show and whether the awards that were given away to actors fall in line with their opinions. It will also be interesting to see how people reacted to the best picture mix up that happened in Oscars 2017.

Social media has evolved as a platform where we can directly evaluate people's liking or disliking to an event. Understanding people's opinion on a social media platform will open us to an unbiased environment. There are no filters to the way people react to an event, and the information that we can tap in from such a platform gives us different perspectives. This provides us a lot of scope to improve the events in the future, and we get a sense of how people receive when something unexpected happens at grand events.

This research paper puts forward an analysis based on the sentiments of the audience during 2017 Oscars with the Best Motion Picture mix up. Analyzing people's reaction towards the Best picture mix up will help us to handle such unexpected events in a better way.

DATA INFORMATION

The social media platform chosen for this analysis is Twitter. The tweets are collected using twitter archiver in google spreadsheets. The tweets are collected in the timeframe Feb 17th,2017 to Feb 28th ,2017 and around 7605 tweets are collected in this timeframe in English language.

We can observe the volume of tweets collected each day in the period considered. The tweets collected were maximum on the day before the OSCAR 2017 event and on the date of the event as seen below.

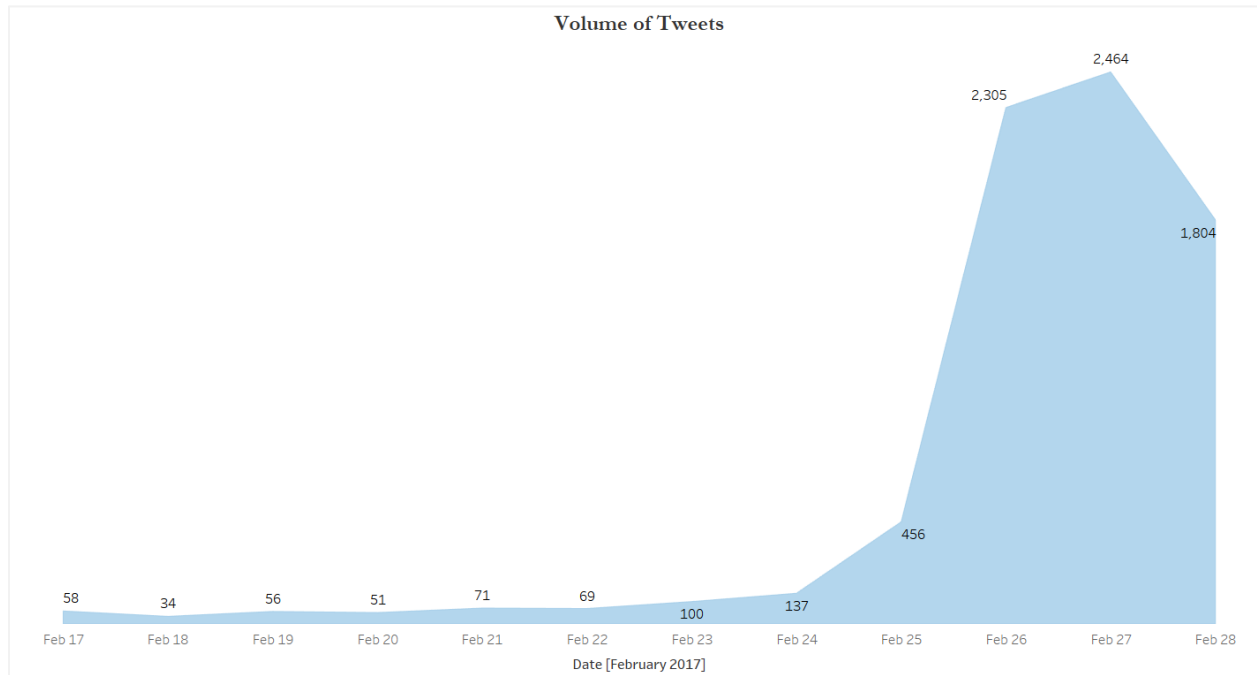


Figure 1. Volume of Tweets

One example of the database for one tweet:

| | Date | App | Followers | Follows | Retweets | Favorites | Verified | Location | Bio | Screen_Name | Full_Name |
|---|------------------|---------------------|-----------|---------|----------|-----------|----------|----------|----------------------------|---------------|-----------|
| 1 | 17FEB17:06:43:17 | Twitter for Android | 1098 | 1236 | 2 | 0 | No | | Top Rated Plus eBay Seller | @AunesDesigns | Aunie |

| Tweet_Text | Tweet_ID | User_Since | Profile_Image | Google_Maps |
|--|--------------------|--------------|---------------|-------------|
| RT @RealZoeHewitt: Did #AmyAdams deserve an #Oscar2017 nomination? Check out my #moviereviews w link in bio. #amivalmovie #femalemoviecrt... | 832601314559348736 | 19237.988113 | View | |

With techniques of text mining, we create a database where we use each word in the tweet like variables. Then using the frequent words in the database, we apply text-mining techniques.

Metadata

| Variable Name | Type | Format | Length | Description |
|---------------|------|--------|--------|------------------------------------|
| Date | Date | DATE9 | 8 | Date on which the tweet was posted |

| | | | | |
|-------------|--------|----------|-----|---|
| Screen Name | String | CHAR20 | 20 | Username of the twitter |
| Full Name | String | CHAR20 | 20 | Name of the Tweeter |
| Tweet Text | String | CHAR 200 | 200 | The actual text of the tweet |
| Tweet ID | String | CHAR20 | 20 | Unique ID of the tweeter |
| App | String | CHAR50 | 50 | Device through which the tweet was posted |
| Followers | Number | BEST12 | 8 | Number of followers for that twitter |
| Follows | Number | BEST12 | 8 | Number of fellow tweeters he/she follows |
| Retweets | Number | BEST12 | 8 | Number of times hi/her tweet has been retweeted |
| Favourite | Number | BEST12 | 8 | Total number of likes for that tweet |
| Verified | Number | Boolean | 1 | Flag of verification |
| User Since | DATE | DATE9 | 8 | Time since the twitter account is active |
| Location | String | CHAR150 | 150 | Location of tweet |
| Bio | String | CHAR150 | 150 | Biography information the user |

Table 1. Metadata of the dataset

METHODOLOGY

The modelling approach followed for the project is SEMMA (Sample, Explore, Modify, Model, Access). The data was partitioned into training, validation & score data. The training data is used to build the model. Validation data is used to test the accuracy of the model. Scoring dataset is used in the sentimental analysis to score the model & get sentimental distribution in the data.

For Sentimental Analysis, the twitted feed was classified into positive & negative categories and this sample was used to train the statistical model in the sentimental analysis studio.



Figure 2. SAS EM flow diagram

SAS TEXT MINING ANALYSIS

The figure shows the different nodes that are used for extraction & analysis of different words from the tweets database.

The ZIPF plot shows the ranking of the words based on its frequency of occurrence in each tweet.

From the below plot we can see that only two words that occurred with the highest frequency are Oscar2017 & rt. rt does not have any significance and is dropped using the text filter node.

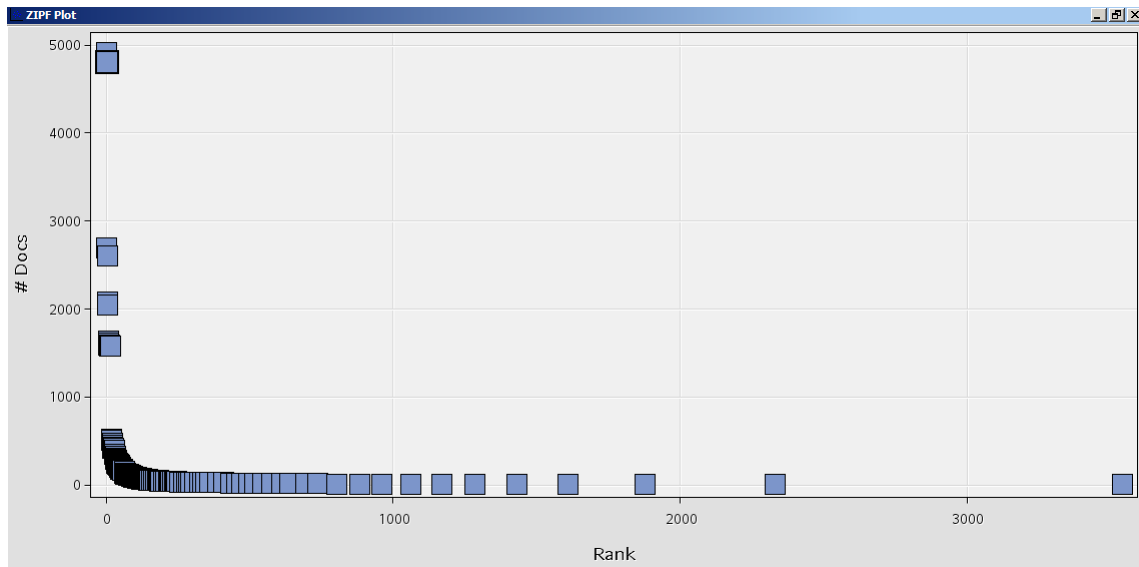


Figure 3. ZIPF plot

As we can see from the below term matrix that oscar2017 & RT are the most frequent words and RT is dropped. We can also observe that the frequency of oscar2017 has increased, it is because all the synonyms of oscar2017 have been merged together.

| Terms | | | | | | | |
|--------------------------|-------------------------|--------|--------|-------------------------------------|--------|----------------------|-----------|
| | TERM | FREQ ▼ | # DOCS | KEEP | WEIGHT | ROLE | ATTRIBUTE |
| <input type="checkbox"/> | oscar2017 | 9889 | 7546 | <input checked="" type="checkbox"/> | 0.009 | Miscellaneous Pro... | Entity |
| <input type="checkbox"/> | rt | 4932 | 4916 | <input type="checkbox"/> | 0.0 | Miscellaneous Pro... | Entity |
| <input type="checkbox"/> | s | 3655 | 2039 | <input type="checkbox"/> | 0.0 | | Alpha |
| <input type="checkbox"/> | be | 2709 | 2598 | <input type="checkbox"/> | 0.0 | | Alpha |
| | more | 1635 | 1634 | <input type="checkbox"/> | 0.0 | | Alpha |
| <input type="checkbox"/> | hour | 1608 | 1608 | <input checked="" type="checkbox"/> | 0.174 | | Alpha |
| <input type="checkbox"/> | russia | 1594 | 1594 | <input type="checkbox"/> | 0.0 | Location | Entity |
| <input type="checkbox"/> | investigation | 1589 | 1589 | <input checked="" type="checkbox"/> | 0.175 | | Alpha |
| | trump | 1589 | 1589 | <input checked="" type="checkbox"/> | 0.175 | Miscellaneous Pro... | Entity |
| <input type="checkbox"/> | screw-up | 1584 | 1584 | <input checked="" type="checkbox"/> | 0.176 | | Mixed |
| | ananavarro | 1582 | 1582 | <input checked="" type="checkbox"/> | 0.176 | | Alpha |
| <input type="checkbox"/> | tie | 1582 | 1582 | <input checked="" type="checkbox"/> | 0.176 | | Alpha |
| | 12 hours | 1581 | 1581 | <input checked="" type="checkbox"/> | 0.176 | Time Period | Entity |
| | clarification | 1581 | 1581 | <input checked="" type="checkbox"/> | 0.176 | | Alpha |
| | oscar2017 screw-up | 1581 | 1581 | <input checked="" type="checkbox"/> | 0.176 | Miscellaneous Pro... | Entity |
| <input type="checkbox"/> | win | 565 | 513 | <input checked="" type="checkbox"/> | 0.306 | | Alpha |
| <input type="checkbox"/> | best | 536 | 513 | <input checked="" type="checkbox"/> | 0.305 | | Alpha |
| <input type="checkbox"/> | not | 533 | 502 | <input type="checkbox"/> | 0.0 | | Alpha |
| <input type="checkbox"/> | lalalandmovie | 513 | 477 | <input checked="" type="checkbox"/> | 0.313 | Miscellaneous Pro... | Entity |
| | syria | 452 | 452 | <input checked="" type="checkbox"/> | 0.316 | Location | Entity |
| | oscars | 428 | 428 | <input checked="" type="checkbox"/> | 0.322 | Miscellaneous Pro... | Entity |
| <input type="checkbox"/> | statement | 413 | 413 | <input checked="" type="checkbox"/> | 0.326 | | Alpha |
| <input type="checkbox"/> | defence | 412 | 412 | <input checked="" type="checkbox"/> | 0.326 | | Alpha |
| | civil defence statement | 411 | 411 | <input type="checkbox"/> | 0.0 | Noun Group | Alpha |
| | civil | 411 | 411 | <input type="checkbox"/> | 0.0 | | Alpha |
| | 25th february 2017 | 411 | 411 | <input checked="" type="checkbox"/> | 0.327 | | Mixed |

Figure 4. Term Matrix

The following concept link diagram has been obtained using the most frequent term in the tweets oscar2017.

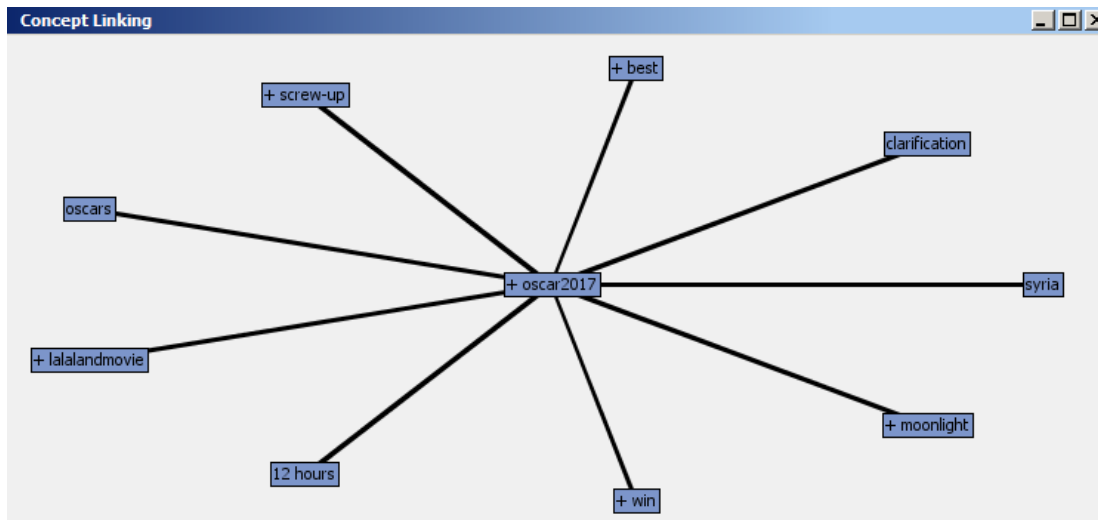


Figure 5. Concept Link Diagram – Oscar2017

The above concept link diagram shows how oscar2017 is related to words like:

Oscars – it shows that the people not only expressed their views about oscar2017 but also about the Oscars event in general.

Best, win – The word best has a strong link with oscar2017 since people have strong inclination towards knowing the best actor, best movie and other actors who won the award.

Lalalandmovie, moonlight, screw-up, clarification – These words are significant in revealing people's sentiment towards the best picture mix up that happened between Lalaland & moonlight movies. The thickness of these words linking to oscar2017 is more, which shows that lot of people, have expressed their views on the confusion that took place for the best picture award.

The following concept link diagram is based on the term 'screw-up'.

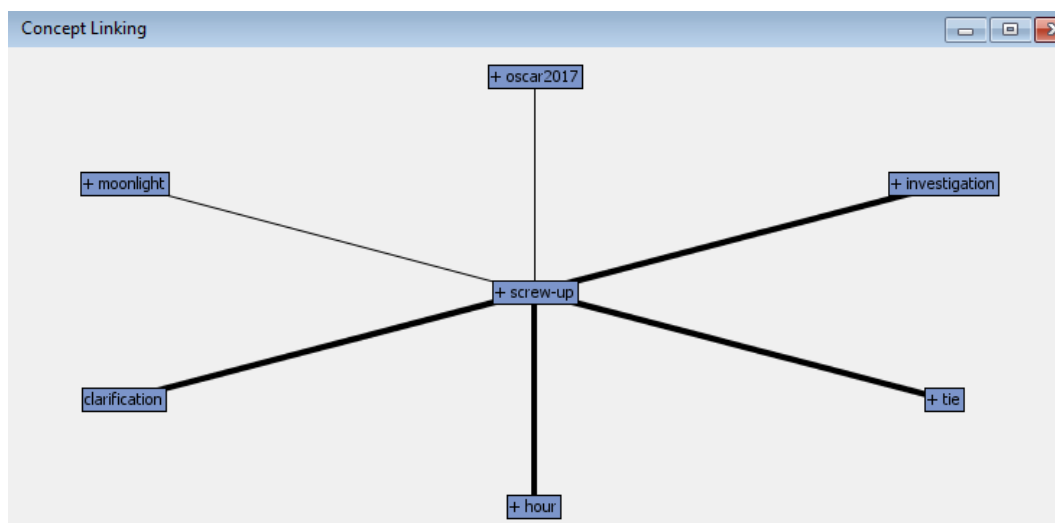


Figure 6. Concept Link Diagram – Screw-up

The above concept link diagram is taken into consideration since we wanted to analyze the sentiments of the people towards the best picture mix-up and this concept link diagram will lay a foundation for the sentimental analysis. The concept link diagram above helps us to train our statistical model better, we can identify easily the terms that are strongly associated with 'screw-up'.

The terms strongly associated with 'screw-up' are

Oscar2017 – association of this term is blatant since the mix up happened in Oscars 2017.

Investigation – An Investigation has been called upon after the best picture mix-up happened.

Clarification- This term appeared as one of the strong links since there are many people who have commented their opinions on the clarification later happened about the best picture mix-up.

The following cluster diagram was obtained using the text cluster analysis on the twitter feeds.

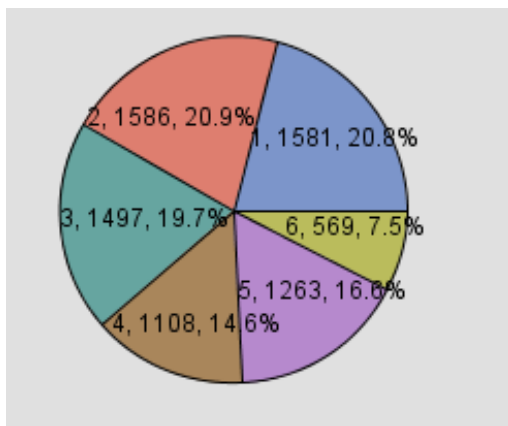


Figure 7. Cluster diagram

The method used for text clustering is Expectation Maximization method. There are 6 clusters obtained with significant difference. The clusters are separated by a significant distance & are the observations are uniformly distributed.

| Cluster ID | Descriptive Terms | Frequency | Percentage | Coordinate 1 | Coordinate 2 | Coordinate 3 | Coordinate 4 | Coordinate 5 | Coordinate 6 | Coordinate 7 |
|------------|---|-----------|------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| 1 | +hour +investigation +screw-up +tie clarification +oscar2017 ... | 1581 | 21% | 0.99999 | .0009478 | -0.0027 | 0.001576 | -0.00164 | .0009007 | 1.495E-5 |
| 2 | +moonlight +best +love land +lalaland +movie +oscars2017 +lol +oscarfail +actor +manchester +arrival +prediction hidd... | 1586 | 21% | 0.001998 | -0.01036 | 0.219762 | -0.05384 | 0.096204 | -0.05195 | 0.021783 |
| 3 | +win +best +night +winner +picture +film complete +tonight ready +pay +nominee +watch Inlta andresnavyz +list ... | 1497 | 20% | 0.001768 | -0.01127 | 0.142381 | -0.11629 | 0.234516 | -0.07164 | -0.0192 |
| 4 | +oscar2017 +statement majdihalaf1993 +news +play +remember +oscarfail +want +live denzelwashington +stop +show ... | 1108 | 15% | 0.17301 | -0.06881 | 0.196987 | -0.22838 | 0.174401 | -0.07911 | -0.02841 |
| 5 | +miss +tune +celeb +fashion hit +choice +award +year +look +theacademy +academyawards +emmastone +oscar2017 ... | 1263 | 17% | 0.005751 | -0.14622 | 0.066379 | -0.05862 | 0.098793 | -0.06336 | 0.01445 |
| 6 | +white +hollywood fake group face-lift +good red carpet +red carpet +white helmet ryan +gosling +life +wait redcarpet ... | 569 | 7% | 0.001473 | -0.00563 | 0.038329 | -0.35967 | -0.1198 | 0.012367 | -0.00463 |

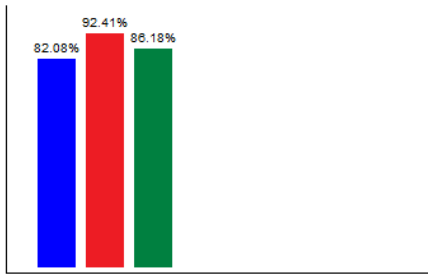
Table 2. Cluster Table

The above cluster table shows the clusters formed and are sorted based on the frequency.

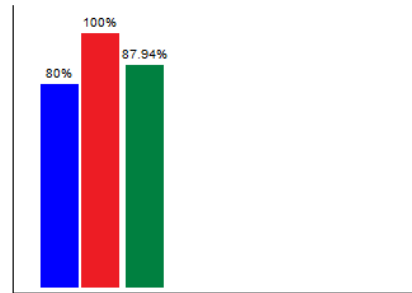
The cluster Id 1 contains the terms hour, investigation, screw-up, tie, clarification & Oscar2017. This cluster contains the data mainly about how best picture mix up has happened. The clarifications given after the show & the investigation that happened post the event. This cluster will be useful for analysis of people's sentiments towards the best picture mix up.

BEST MODEL is Smoothed Relative Frequency and Chi Square

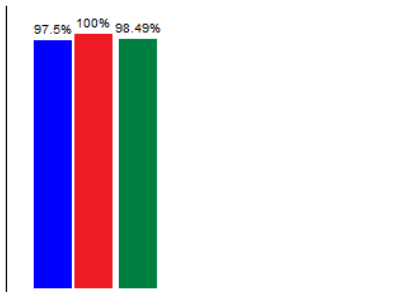
Smoothed Relative Frequency No Feature Ranking



Smoothed Relative Frequency Risk Ratio



Smoothed Relative Frequency Chi Square



Smoothed Relative Frequency Information Gain

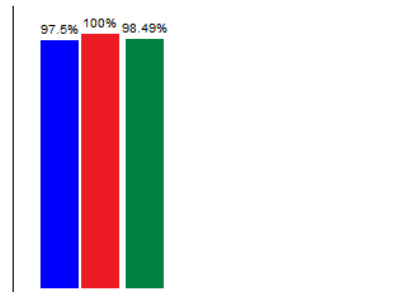


Figure 8. Statistical Model

A statistical model has been built based on twitter feed classified into positive & negative comments. The above statistical model shows the relative frequency of the tweets data towards Oscars2017 & the best picture mix-up happened combined. This statistical model helps us to identify the polarity of the comments.

Sentiment Distribution

■ Positive ■ Negative ■ Neutral



Figure 9. Sentimental Distribution of Statistical Model

The above sentiment Distribution has been obtained by scoring the statistical model using 500 distinct comments. We can observe the distribution consists of more positive & neutral comments. This is a positive sign as people who were excited about the Oscars2017 outnumbered the people who were unhappy about the event.

CONCLUSIONS AND FUTURE WORK

This paper sets a stage in order to analyze stage shows in general & people sentiments towards the shows will give us an idea of the success of the show. This paper not only deals with sentiments of the people before the show but also during & after the show and hence it gives a better picture of how to handle any unwanted circumstances during the event. This paper was started with an idea to identify the people reaction towards unwanted events during stage shows. We can conclude that taking Oscars2017 show into consideration, the sentiments of the people were more positive & neutral stating that the excitement of the people towards the show will over shadow any unwanted events. This analysis can be extended to build a text predictive model wherein there is a scope of predicting the sentiments towards unwanted events & will help us to set the stage better and prepared.

REFERENCES

1. <http://support.sas.com/publishing/pubcat/chaps/65646.pdf> - Text Mining and Analysis – Practical Methods, Examples & Case Studies using SAS®
2. <http://oscar.go.com/news/nominations/oscar-nominations-2017-view-the-complete-list-of-nominees>
3. <https://support.sas.com/edu/schedules.html?id=2889&ctry=US&locationId=sf>
4. https://en.wikipedia.org/wiki/89th_Academy_Awards

ACKNOWLEDGMENTS

I sincerely thank Dr. Goutam Chakraborty for his valuable guidance and motivation for accomplishing this paper. I also thank Dr. Miriam McGaugh for her constant support and suggestions throughout this study.

CONTACT INFORMATION

Karthik Sripathi
Master's in Business Analytics
Oklahoma State University, Stillwater
405-385-3377
Karthik.sripathi@okstate.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.