Cinema rating from sentiments perspective

Ing. Juan Eduardo Rioja, Ing. Luis Enrique Ostos, Lic. Yesenia Becerra

Abstract-- **The purpose of this analysis is to have a tool that helps to understand the current tendency of the audience with films that currently are in theaters. Another objective is to answer the question if is possible to understand which movie is the most popular by the audience and know if this tool provides the enough knowledge to select the top movie by the audience. To achieve this, we used data gathered from Twitter APIs and Stanford CoreNLP to assign a sentiment rating**.

I. Introduction

The social network Twitter, is one of the most popular service that provide to the user the creation of microblogging service where the users can create short messages as status, know them as tweets. Commonly the tweets express the opinions from the users about different topics. The team from Stanford proposed a method to extract the sentiment, which can be positive or negative from the tweets. That with the goal to allow a feedback without a manual intervention.[3]

With the growth of the use of the internet and the amount of data that can be found in it, you can take advantage of large amounts of data that are generated day by day by millions of users worldwide. Web 2.0 was used for the studio since it is a rich source of information with constant updates, the network used was Twitter since it is one of the most immediate networks and from which information can be extracted at the time of any topic of interest that the researcher needs to investigate. Twitter is a famous micro-blogging and social networking service which provides the facility to users to share, deliver and interpret 140 words' post known as tweet Twitter have 320 M monthly active user. Twitter is accessible through website interface, SMS, or mobile devices. Eighty percent users are active through mobile. In the micro-blogging services users make spelling mistakes and use emoticons for expressing their views and emotions. Natural language processing is also playing a big role and can be used according to the opinions expressed.

Ratings from critics can be found from many professional critical sites, and this paper study the possibility to take the opinion based on public opinion.

Here it was taken as a hypothesis that the analysis of feelings allows us to know the approval or disapproval of a movie, considering first that when a movie has negative feelings it is because the film did not cause a good feeling. For the investigation three movies were chosen Ad Astra, Down Town Abbie and Joker ; the decision of the three movies was for a similarity in dates in which all were on the billboard, all three are in the taste of the published, and the age range is similar for the 3 movies. The aim is to identify the most successful films and identify the areas with the best results and then predict the success of films with similar characteristics.

II. RELATED WORK

Rafael Michal Karampatsis described the twitter sentiment analysis for specifying the polarity of messages. They used the two-stage pipeline approach for analysis. Authors used the sum classifier at each stage and several features like morphological, POS tagging, lexicon etc. are identified.

Sentiment analysis has been handled as a Natural Language Processing task at many levels of granularity. It has been handled at the sentence level and more recently at the phrase level.

III. Methods

To develop this work, the methodology was based in ETL (Extraction, Transformation and Loading). To extract data was necessary to generate a java application which extracts the information and load it into a Mongo database (NO SQL). The program gets the information from twitter, so each tweet is been analyzable by NPL Stanford library.

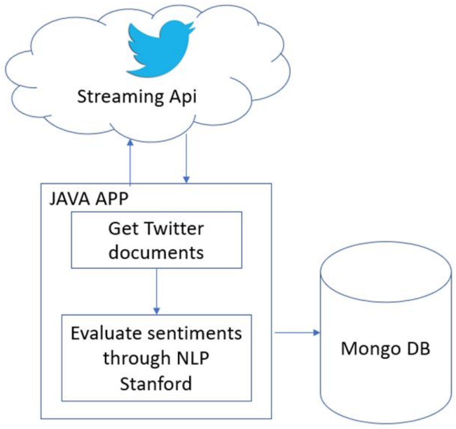

Figure 1: Process to extract streaming data

A. *Streaming API and Data Collection*

Data collection was achieved by login in twitter user account and going to the application site https://apps.twitter.com/which enable the researcher to create an application. It provides specific and unique keys: access-token, access-token-secret, consumer-key, and consumer secret. The keys used in Java program application that access Twitter Streaming API for downloading tweet data. Download operation was done on specified keywords then saving the data in JSON file.

The keywords used to this problem were three different records because we evaluate three movies

Joker keywords

String topic = "(Movie Joker) or (Joker Joaquin Phoenix) or (Joker Todd Phillips) or (Joker Film)”

Astra keywords

String topic = "(Movie,Ad Astra) OR (Ad Astra,Brad Pitt) OR (Ad Astra James Gray) OR (Ad Astra 2019) OR (Ad Astra Rated:PG-13)"

Downton Abbey

String topic = "(Movie,Abbey) OR (Film,Abbey) OR (Movie,Downton Abbey) OR (Movie,DowntonAbbey) OR (Movie DowntonAbbey) OR (Movie Downton Abbey) OR (Film,Downton Abbey) OR (Film,DowntonAbbey) OR (Film DowntonAbbey) OR (Film Downton Abbey) OR (Downton Abbey film) OR (Downton Abbey Hugh Bonneville) OR (Downton Abbey Michael Engler) OR (Downton Abbey 2019) OR (Downton Abbey Rated:PG)”

B. *Natural Language Processing*

Natural language processing, is the process of interpreting text or speech with the help of a machine learning algorithm. Natural language processing, is the machine’s approach to interpret texts and divide sentences into words, subsequent tag words according to their parts of speech, grammar checking, correction of spelling and further advanced tasks as well like sentiment analysis, context extraction etc. The common tasks perform by NLP are:

• Tokenization: A process to break a sentence into individual words or to break a paragraph into individual sentences.

• Parts of speech tagging: A process to categorize every word from a text/speech to 16 categories of parts from the speech.

• Grammar Checking: Natural language processing that can check the grammar for a sentence, by introducing regular expressions or using a free context parser. 6

• Sentiment analysis. Identify the mood or opinion from how a speaker/writer feels on a particular text/speech.

• Speech-to-text and text-to-speech conversion. Transforming voice commands into written text, and vice versa. [8]

CoreNLP

Stanfords CoreNLP API has been used to obtaining the sentiments on a real-time. It is fundamentally based on JVM annotation pipeline framework, that provides most of the popular and natural language processing steps. Three datasets have been considered, each of which contain filtered data based on the tags of each movie.

C. *Big Data*

The Big Data term is used to describe huge quantity of data in the network, by the order of gigabytes and terabytes. The data mainly came from many sensors for the Internet of Things (IoT), in a world impulse by the constant generation of information by many different technology dispositive that currently exist [4].

The Big Data technology born from the incapacity of the architectures of traditional data to process in an efficient way the exorbitants volumes of new group of data generated daily. The characteristics of Big Data that give strength to the new related technologies are:

- Volume (the size of the data);
- Variety (many repositories, domains or types for the data);
- Velocity (rate of flow since the data is generated);
- Variability (the change in other characteristics) [7].

D. *Data visualization*

The data visualization is based in a simple idea: the human brain is not prepared to handle looseness arbitrary and abstracts symbols, like numbers. The human brain is capable to interpret the meaning of huge quantity of values only by indirectly way: by example, when we represent proportionally by many different visual objects properties, as high, length, size, angle and thickness [1].

The visualization try to build a graphic group, synthetic or complementary, that stand out the main and key topics, that allow understand, group, relate or generate a statistic tendency, that minimize the entropy and facilitate the way to get the conclusions or tests, for the own interpretation [6].

E. Data Extraction

Some techniques and algorithms of data extraction are easy to understand and to use to take the decisions. The visualization can do that the data and the results of the extraction are more accessible, which allow to compare and verify the results. The visualization can be used to drive some algorithms of data extraction [5].

IV. Results and discussion

This section shows the results achieved in the work and offers an interpretation of those results. Acknowledge any limitations of the work and avoid exaggerating the importance of the results.

Ad Astra
Tweets: 56,326
First: 04/10/2019 04:46:06 am
Last: 08/10/2019 05:24:04 pm
Time: 5 days

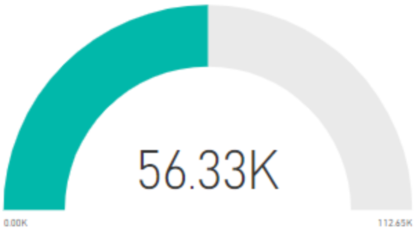

Figure 2 Count of tweets by Day, October

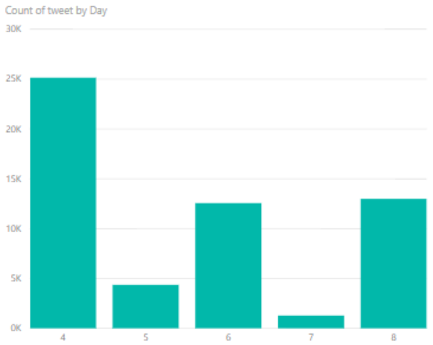

Figure 3 Records by day of AD Astra Movie

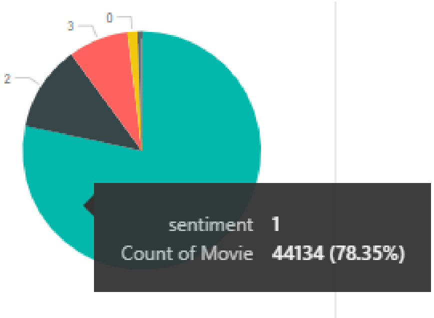

Figure 4 Sentiments generated by AD Astra Movie

The 78.35% of the tweets about Ad Astra Movie address to a Negative sentiment (value 1 that means "Negative"), however that result doesn’t reflect the preference from the audience according with the sentiment analysis, most of the tweets gets a negative score because some of their words have a negative effect. That can be reflected in the next analysis:

1 - Negative sentiment
"Ad Astra (2019) spoilers without context"

Is a good example of how the lack of expression can lead to a misunderstanding by the algorithm, the word without context, consider in a movie environment can be understood as a bad impression for the audience

4 - Positive sentiment
"Joyful tweet of the day: saw the new Downton Abbey movie and it was delightful!"

Joker
Tweets: 102,422
First: 04/10/2019 06:27:28 am
Last: 05/10/2019 09:25:46 pm
Time: 2 days

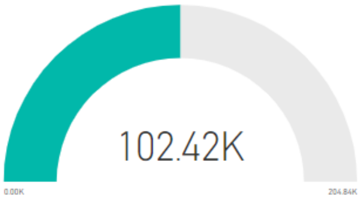

Figure 5 Count of tweets by Day, October

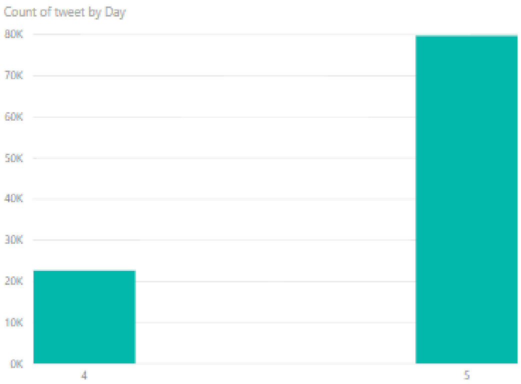

Figura 6 Records by day of Joker

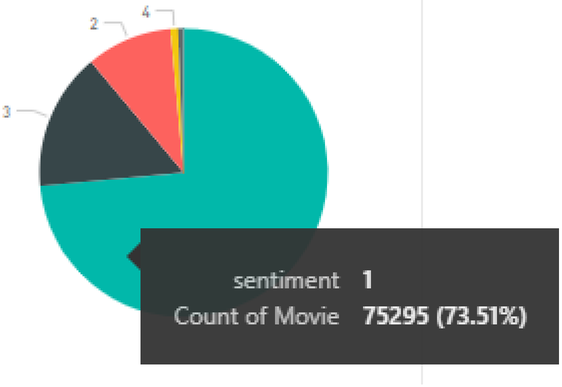

Figure 7 Sentiments generated by Joker

The 73.51% of the tweets about Joker Movie address to a Negative sentiment (value 1 that means "Negative"), however that result doesn’t reflect the preference from the audience according with the sentiment analysis, most of the tweets gets a negative score because some of their words have a negative effect. That can be reflected in the next analysis:

1 - Negative sentiment

"We love a superhero, but there's always a villain in the picture. Joaquin Phoenix plays the Joker in the 2019 thriller"

This example helps to see that even if the tweet doesn’t show if the user like or dislike the movie, using words as villain, joker and thriller can be understood as a negative sentiment

4 - Positive sentiment

"Joker is the best movie of 2019 so far."

That a is complete positive expression for the movie Joker

Downton Abbey
Tweets: 104,941
First: 04/10/2019 03:13:53 am
Last: 06/10/2019 08:13:05 pm
Time: 3 days

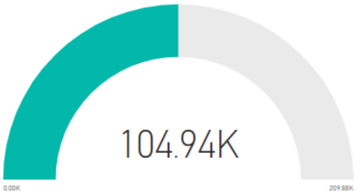

Figure 8 Count of tweets by Day, October

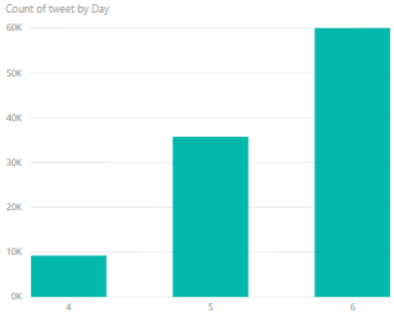

Figure 9 Records by day of Downton Abbey

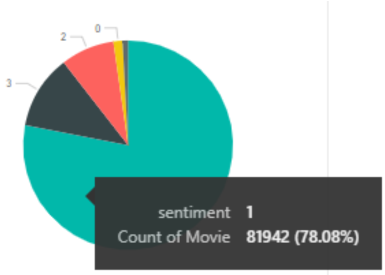

Figure 10 Sentiments generated by AD Astra Movie

The 78.08% of the tweets about Downton Abbey Movie address to a Negative sentiment (value 1 that means "Negative") however that result doesn’t reflect the preference from the audience according with the sentiment analysis, most of the tweets gets a negative score because some of their words have a negative effect. That can be reflected in the next analysis:

1 - Negative sentiment

"seriously though I'll likely see the Downton Abbey movie long before I ever watch Joker"

As we can see in these tweets the audience made the comparison versus another movie, however the word joker can be interpreted as an insult for the movie

4 - Positive sentiment

"Joyful tweet of the day: saw the new Downton Abbey movie and it was delightful!"

This example is a clear way to see that the analysis throws a positive sentiment from the user to the movie Downton Abbey

V. Conclusion

Analyzing text data using Stanford’s CoreNLP, makes the analysis of data easy and efficient. With just a few lines of code, CoreNLP allows the extraction of all kind of text properties, such as named-entity recognition or part-of-speech tagging

Most of the tweets shows some texts that gets a negative score due some words in the sentence has a negative effect. As we advance in score, we identify positive words and less appearance of negative words or sentences.

Due to irregular, short form of text (hlo, whtsgoin etc.), short length and slang text of tweets, it is challenging to predict polarity of sentiment text. In sentiment a mixture of applications are needed to study and these all demands large number of sentiments from sentiment holder. A summary of sentiment is needed, as in polarity disambiguation and analysis; a single sentiment is not adequate for decision.

In our opinion, this tool would be throwing many good information if we put the limit of how many words can describe the main idea.

Natural Language Processing (NLP) is advancing and the sentiment analysis studies are increasing in the last years, though there are challenges for NLP like language diversity, satire and emoticon detection.

VI. References

[1] G. Eason, B. Noble, and I.N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” *Phil. Trans. Roy. Soc. London,* vol. A247, pp. 529-551, April 1955.

[2] J. Clerk Maxwell, *A Treatise on Electricity and Magnetism,* 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[3] I.S. Jacobs and C.P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[4] K. Elissa, “Title of paper if known,” no puplicado.

[5] R. Nicole, “Title of paper with only first word capitalized,” *J. Name Stand. Abbrev.*, en impresión.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” *IEEE Transl. J. Magn. Japan*, vol. 2, pp. 740-741, August 1987 [*Digests 9th Annual Conf. Magnetics Japan*, p. 301, 1982].

[7] M. Young, *The Technical Writer’s Handbook.* Mill Valley, CA: University Science, 1989.