A Novel Approach to Analysis of TV Shows using Social Media, Machine Learning and Big Data

Priyank Jain
B. Tech. (Computer Engineering)
Veermata Jijabai Technological Institute (VJTI)
Mumbai, Maharashtra, India

Prateek Jakate, Akash Dhotre, Jigar Bhati
B. Tech. (Computer Engineering)
Veermata Jijabai Technological Institute
Mumbai, Maharashtra, India

Abstract— Television Rating Point (TRP) system is used to measure popularity of Television (TV) shows using people’s meter [1]. Not only does this approach take into account the opinions of a fractional percentage of the audience, but it is also not enough to understand the target audience along various dimensions. In this paper we present a system that uses a novel approach to analyze the audience of a TV show using machine learning, social media and big data technologies. This system will allow TV show producers and potential advertisers to perform sentiment analysis, gender analysis, device analysis of a TV show. Sentiment analysis uses Naïve Bayes Classification [2]. Gender analysis uses data-driven gender determination from name combined with Naïve Bayes Classification. This paper suggests ways to improve on this system to facilitate better insights into the characteristics of the audience of a TV show. This system is economically cheaper to setup, maintain and is easily scalable to accommodate multiple shows and different dimensions of analysis.

Keywords- Naïve bayes; machine learning; big data; twitter; social media analysis; gender analysis; sentiment analysis; hadoop.

I. INTRODUCTION

Television industries in countries like USA, India and China are estimated to be worth billions of dollars. TV show producers and advertisers constantly seek ways to understand their target audience along various dimensions. This allows TV show producers to understand certain aspects of the preferences of their viewers, gain valuable feedback and improve their shows. Advertisers can elaborate more specific marketing campaigns according to their audience. Currently used system for Target Audience Measurement is Television Rating Point. For calculation purpose, a device is attached to the TV set in a few thousand viewers’ houses for judging purpose [1]. These numbers are treated as sample from the overall TV owners in different geographical and demographic sectors. The device is called as People's Meter [1]. It records the time and the program that a viewer watches on a particular day. Then, the average is taken for a 30-day period which gives the viewership status for a particular channel. Recently, many flaws have been pointed out in the TRP system:

- The number of sample respondents is a minute fraction of the actual number of people viewing television in the country. For example, in India, the number of sample respondents is 45287 [3] which is approximately 0.0035% of the population of India (1.29 billion as of 2014) [4]
- It is not possible to understand the sentiment of the sample respondents and understand their buying preferences from people’s meter. Such data is useful for advertising purposes.

The number of sample respondents is such a small fraction of the actual number of TV viewers, that it would be wrong to say that the sample captures the viewing habits of the country as a whole. Moreover, if the number of sample respondents have to be increased, more households will have to install the people’s meter, which is highly costly both in terms of installation and maintenance.

The system we present is a novel approach to understand the audience of a particular TV show. It captures the sentiments of TV audience who are active on twitter (http://twitter.com). It does sentiment analysis, device analysis and gender analysis of the tweets made about a particular TV show.

The paper is organized as follows. Section II discusses the data source - twitter, it’s Application Programming Interfaces (APIs) and metadata. Section III discusses the high level working of the system. Section IV discusses sentiment analysis component. Section V discusses gender analysis component. Section VI discusses rest of the features in the system. Section VII discusses front end results. Section VIII discusses future work and the paper is concluded in section IX.

II. DATA SOURCE

We used Twitter.com as the source of data for the system. Twitter is the most popular social networking service that enables users to send and read short 140-character messages called "tweets". There are three different ways for a developer to get data from twitter:

A. Search API

Get relevant tweets from the recent past. Free access to developers, with limitation of 180 queries per 15 minute window [5].

A Novel Approach to Analysis of TV Shows using Social Media, Machine Learning and Big Data

Priyank Jain
B. Tech. (Computer Engineering)
Veermata Jijabai Technological Institute (VJTI)
Mumbai, Maharashtra, India

Prateek Jakate, Akash Dhotre, Jigar Bhati
B. Tech. (Computer Engineering)
Veermata Jijabai Technological Institute
Mumbai, Maharashtra, India

Abstract— Television Rating Point (TRP) system is used to measure popularity of Television (TV) shows using people’s meter [1]. Not only does this approach take into account the opinions of a fractional percentage of the audience, but it is also not enough to understand the target audience along various dimensions. In this paper we present a system that uses a novel approach to analyze the audience of a TV show using machine learning, social media and big data technologies. This system will allow TV show producers and potential advertisers to perform sentiment analysis, gender analysis, device analysis of a TV show. Sentiment analysis uses Naïve Bayes Classification [2]. Gender analysis uses data-driven gender determination from name combined with Naïve Bayes Classification. This paper suggests ways to improve on this system to facilitate better insights into the characteristics of the audience of a TV show. This system is economically cheaper to setup, maintain and is easily scalable to accommodate multiple shows and different dimensions of analysis.

Keywords- Naïve bayes; machine learning; big data; twitter; social media analysis; gender analysis; sentiment analysis; hadoop.

I. INTRODUCTION

Television industries in countries like USA, India and China are estimated to be worth billions of dollars. TV show producers and advertisers constantly seek ways to understand their target audience along various dimensions. This allows TV show producers to understand certain aspects of the preferences of their viewers, gain valuable feedback and improve their shows. Advertisers can elaborate more specific marketing campaigns according to their audience. Currently used system for Target Audience Measurement is Television Rating Point. For calculation purpose, a device is attached to the TV set in a few thousand viewers’ houses for judging purpose [1]. These numbers are treated as sample from the overall TV owners in different geographical and demographic sectors. The device is called as People's Meter [1]. It records the time and the program that a viewer watches on a particular day. Then, the average is taken for a 30-day period which gives the viewership status for a particular channel. Recently, many flaws have been pointed out in the TRP system:

- The number of sample respondents is a minute fraction of the actual number of people viewing television in the country. For example, in India, the number of sample respondents is 45287 [3] which is approximately 0.0035% of the population of India (1.29 billion as of 2014) [4]
- It is not possible to understand the sentiment of the sample respondents and understand their buying preferences from people’s meter. Such data is useful for advertising purposes.

The number of sample respondents is such a small fraction of the actual number of TV viewers, that it would be wrong to say that the sample captures the viewing habits of the country as a whole. Moreover, if the number of sample respondents have to be increased, more households will have to install the people’s meter, which is highly costly both in terms of installation and maintenance.

The system we present is a novel approach to understand the audience of a particular TV show. It captures the sentiments of TV audience who are active on twitter (http://twitter.com). It does sentiment analysis, device analysis and gender analysis of the tweets made about a particular TV show.

The paper is organized as follows. Section II discusses the data source - twitter, it’s Application Programming Interfaces (APIs) and metadata. Section III discusses the high level working of the system. Section IV discusses sentiment analysis component. Section V discusses gender analysis component. Section VI discusses rest of the features in the system. Section VII discusses front end results. Section VIII discusses future work and the paper is concluded in section IX.

II. DATA SOURCE

We used Twitter.com as the source of data for the system. Twitter is the most popular social networking service that enables users to send and read short 140-character messages called "tweets". There are three different ways for a developer to get data from twitter:

A. Search API

Get relevant tweets from the recent past. Free access to developers, with limitation of 180 queries per 15 minute window [5].
B. **Streaming API**  
Get a fraction of the tweets in real time that match your query. Tweets are received as they are made in real time [6].

C. **Firehose API**  
Get all tweets in real time that match your query [7]. Not free like the other two APIs.

We used the Search API. For the purpose of this paper, we focus on two American TV shows – Game of Thrones and Suits. The hashtags/keyword lists for these shows are given in Table I. This is a list of keywords, hashtags and twitter handles which will give us almost all the tweets related to that TV show from the Search API.

The data retrieved from Twitter is a Java Script Object Notification (JSON) document, containing a number of key value pairs. A sample JSON document retrieved is shown in Fig. 1. The user field of the tweet which gives information about the user who made that tweet is shown in Fig. 2.

### III. SYSTEM ARCHITECTURE

#### TABLE I. LIST OF KEYWORDS FOR GAME OF THRONES AND SUITS

<table>
<thead>
<tr>
<th>TV Show</th>
<th>Keywords list</th>
</tr>
</thead>
<tbody>
<tr>
<td>Game Of Thrones</td>
<td>game of thrones, #GoT, GameOfThrones, @GameOfThrones, @GoTNews, daenerystargaryen, tyronnansister, aryastark, honsnow, peterdinklage, @Peter_Dinklage</td>
</tr>
<tr>
<td>Suits</td>
<td>#harveyspecter, @GabrielMacht, halfadams, #Suits, #SuitsFinal, @Suits_USA</td>
</tr>
</tbody>
</table>

The different modules in the system are summarized in Fig. 3. The four modules are explained below:

A. **Fetcher Module**  

B. **Analysis Module**  
Sentiment, device and gender analysis is done on the raw tweets stored in MongoDB and the processed tweets’ data is again stored in MongoDB. More on this later.

C. **Summarizer Module**  
A pig [12] script, which runs on top of hadoop [13] is called which will summarize information for tweets processed by analysis module. This module aggregates counts like number of positive tweets, number of negative tweets, number of tweets made by males, number of tweets made by females, number of tweets, number of favorites, number of tweets made via different sources, reach and so on. These are the different metrics which give insights into the nature of the audience of the TV show and popularity of the TV show. The results of this module are again stored in mongodb. Hadoop is used at this stage to make the system scalable.

D. **Web module**  
This module contains web logic written in nodejs [14] so as to display the results to TV show producers and advertisers on a website. Other technologies used here are Cascading Style Sheets (CSS), Javascript, Jquery and HyperText Markup Language (HTML).
IV. SENTIMENT ANALYSIS COMPONENT

While a lot of work has been done in the sentiment analysis domain, having a good accuracy on tweets and tailoring it to the requirements of the application still remains a challenge. The challenge in this system is to classify the tweets as positive, negative or neutral.

We initially used a naïve Bayesian classifier approach. The training set was downloaded from http://www.nltk.org/nltk_data/ under the section “Twitter_Samples”. Using naïve bayes on the raw tweets yielded a very poor accuracy. We modified the algorithm which now consists of following steps: 1) Emoticon based classification 2) Tokenization 3) Stopwords removal 4) Removal of hashtags, twitter handles and links 5) Use the naïve bayes classifier. Emoticon based classification detects the presence of emoticons in the tweets, so a tweet containing ':)' is considered positive and a tweet containing ':(' is considered a negative tweet. The decisions are based on an exhaustive list of emoticons. If classification succeeds at this step, rest of the steps are skipped. Tokenization splits the sentence into an array of words. Stopwords are then removed.
which are most frequently occurring words like articles, prepositions, conjunctions and adverbs. Hashtags, twitter handles and links are then removed from the array of words so as to speedup the algorithm, as such words don’t contribute in the decision of a sentiment, since they correspond to entities and are not descriptive. The rest of the words are feed into the naïve Bayesian sentiment classifier, which returns the probability of the tweet being positive and probability of it being negative. The algorithm gave the best performance of 81% accuracy when a probability of the tweet being positive above 60% was considered as positive, tweets with probability of being negative above 60% were considered as negative and rest of the tweets were considered as neutral tweets. The test set consisted of 100 tweets related to Game of Thrones and 100 tweets related to Suits.

V. GENDER ANALYSIS COMPONENT

The JSON document corresponding to a tweet metadata does not contain information about the gender of the user. Gender analysis gives useful insights into the nature of the audience of the TV show. We classify a Twitter user as either male, female or androgynous based on the name and or text. The androgynous class corresponds to users whose names can be used by both male and female, twitter handles corresponding to institutions or corporations. This classification is done using a name database. We use multiple gender-name libraries in python to achieve a higher accuracy. The genderizer python library is used first to determine the gender by sending the name of the user and the corresponding twitter profile description as input. The library uses an in-built name database along with text analysis based on naïve bayesian classification to determine the gender of the entity making the tweet [15]. If this gender classifier gives the output as androgynous, genderize.io’s online api is used to determine the gender from the name. Genderize.io’s database is a very comprehensive one. At the moment, the database contains 216286 distinct names across 79 countries and 89 languages [16].

VI. OTHER FEATURES

The three other features of this system are explained below:

A. Device analysis

The source key of the JSON metadata specifies the utility used to post the Tweet [17]. This gives insights into how many tweets were made via Iphone App, Windows Phone App, Android App, Web and other applications. This information helps us understand the preferences of the audience of that TV show

B. Popularity

Popularity of a TV show is measured as the total number of tweets made for that particular TV show in a particular period. Another metric which is used to measure popularity is the favorite count of a TV show which is measured as the sum of the favorite_count key [18] in the JSON Tweet metadata of all the tweets related to that TV show. This field specifies the number of twitter users who have favorited that Tweet.

C. Reach

Reach quantifies the network strength of a particular TV show. This measure is an indicator of the maximum number of people who might be exposed to the tweet. The user sub document in the JSON Tweet metadata has a key called followers_count [19], which is the number of people who get notified when the user makes a new tweet. The reach is calculated as the sum of the followers of unique users who have made tweets related to that TV show.

VII. RESULTS

Results are displayed on a website made using nodejs. Fig. 4 shows the home page of the website. Fig. 5 shows popularity metrics, sentiment analysis and gender analysis results for Game of Thrones. Fig. 6 shows gender-wise sentiment analysis for Game of Thrones. Fig. 7 shows top 10 tweet sources for Game of Thrones and top 10 languages in which most tweets related to Game of Thrones are made. Fig. 8 shows popularity metrics, sentiment analysis and gender analysis results for Suits. Fig. 9 shows gender-wise sentiment analysis for Suits. Fig. 10 shows top 10 tweet sources and top 10 languages for Suits. Google Charts API is used for displaying the charts.

VIII. FUTURE WORK

Here are some suggestions that can be used to make the system better and gain more insights into the audience of a TV show:

A. Use real – time data

Currently this system uses REST API, which gives historical tweets. A mechanism could be made which initially uses the REST Twitter API to get historical tweets for a new TV show, then switch to the Streaming API for getting real-time tweets.

B. Gender analysis improvement

Currently, gender analysis uses the name of the user, then the description and the tweets made by the user to determine the gender. In cases where the gender is not yet determined, it would be useful to apply facial recognition on the profile image of the user to determine the gender.

C. Country wise analysis

Understanding the audience as per country would give more insights into the nature of its audience. The time_zone and utc_offset subfields of the user field and geo, place and coordinates fields [16],[17] can be used to obtain geographical information about the user and place from which tweets was done. However, most of the times these fields are empty, hence country wise analysis remains a challenge.
D. Scalable machine learning

The sentiment analysis component can be migrated to Apache Mahout [18] for scalable machine learning. Apache Mahout runs on top of Hadoop, which will make the system more scalable.

IX. CONCLUSION

In this paper, we throw light on shortcomings of the current TRP system and present a system which can address these shortcomings. The system heavily uses open source libraries and leverages machine learning, big data and social media for analysis. This system is not only a low cost way to understand the audience of a TV shows and to contrast TV shows against each other, it's a scalable one which can accommodate multiple TV shows and analysis along various dimensions. This system, if employed, will enable advertisers to do more specific audience targeting and TV show producers to fine tune their TV shows for increased revenue and increased audience coverage.

Figure 4. Home Page
Figure 5. Game of Thrones Popularity, Sentiment and Gender analysis results

Figure 6. Game of Thrones Gender-wise Sentiment analysis

<table>
<thead>
<tr>
<th>No.</th>
<th>Tweet source</th>
<th>Tweet count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Twitter for iPhone</td>
<td>13,638</td>
</tr>
<tr>
<td>2</td>
<td>Twitter for Android</td>
<td>11,695</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No.</th>
<th>Tweet language</th>
<th>Tweet count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>English</td>
<td>51,596</td>
</tr>
<tr>
<td>2</td>
<td>German</td>
<td>11,549</td>
</tr>
</tbody>
</table>
Figure 7. Game of Thrones source and language stats

Figure 8. Suits Popularity, Sentiment and Gender analysis results
Figure 9. Suits Gender-wise Sentiment analysis

Figure 10. Suits source and language stats
ACKNOWLEDGMENT

We would like to thank Mr. Abhishek Mane and Mr. Lavin Mirchandani for their most valuable guidance and support in developing an initial version of the system.

REFERENCES


