Analysis of esoteric and entertainment movies in Microblogs

Gönül AYCI
Computer Science
Özyeğin University
Email: gonul.ayci@ozu.edu.tr

Furkan ELIBOL
Computer Science
Özyeğin University
Email: furkan.elibol@ozu.edu.tr

***Abstract*—Nowadays, access to information has become easier, however the accuracy of the information was compromised. Unfortunately, we are forced to measure the reliability of the information due to the increase of social networks via the web and social media. In this sense, twitter play an important role. Most of the users trust microblogging service Twitter which is the famous one. However twitter data may mislead us. Based on this, in this work, we want to analyze the accuracy of the information especially movie data. Analyzing the accuracy of information and its users are beneficial for the (real) Twitter user. Actually, it is useful many departments such as e-markets, business, road condition, etc.**

***Keywords*—twitter, microblogging, analyze, movie.**

I. Introduction

Today twitter is one of the most popular microblogging services. It is popular with both active users and users that only use it for following. Although twitter has a large amount of users and their active movements, it is increasingly difficult to talk about secure information. For example, we read an article and by analyzing comments about it and the number of retweets and favorites we can have an opinion about it but we can not make analysis about accuracy of information. Because there are fake accounts.

Popularity of movies increases every day, they can be watched at cinema or whenever or wherever you want by using web. This popularity can be proved by the number of people who watch it and the number of comments. Moreover, most of the users watch trailer or pay attention to comments of people who watch it before decide to watch a film. Twitter is an important microblogging service which also play a role in this decision.

Most of the movies have some messages that is expected to convey to audiences. From this point of view we have selected four movies and want to collect data about them and then analyze this data.

Users who make comments, tweets, give rates sometimes not becaue they like the movie but just to be supportive. And they can make these without watch the movie or even before the movie come out just to create positive or negative perception. There are many users who do that kind of things just to create perception. Esoteric type of movies can be example of this. Esoteric movies mean that they address specific audience and this audience support that kind of movies on any ground. For example, politic films. We have chosen two entertainment and two esoteric movies to compare.

In this project, we try to analyze types of esoteric and entertainment movies. Our main goal is to distinguish these two different types of movies just using twitter data. We choose four popular cinema movies. These are Kod Ad Koz (KOD)-esoteric, Icimdeki ses(IS)-entertainment, Bana Masal Anlatma(BMA)-entertainment, Selam Bahara Yolculuk(SBY)-esoteric. We have criticized these four movies base on the number of people watch them; such as KOD-IS and BMA-SBY. By analyzing the data we have taken it into consideration and make comments.

In addition, in this project, by visualising the relation between follower-user-following can help us for analyzing . If one of his or her fallowers of fallowing is also tweeet about that film this helps us to scrutinize the relation between them. Because of this we have tried to collect data and analyze them to reach a conclusion about these movies. It is very time consuming to collect follower-following data (some users are public users who have millions fallowers, for example hashtags of MyBilet, cinemaximum and MarsGateOnline are common for four movie data.)so we have not yet reach a conlusion but we completed collecting data.

The rest of the paper is organized as follows. Section II provides related works. Section III introduces how to collect data and what are the technics for them? and Section IV, analyze of the collected movie data and interpretion of them. Section V concludes with a summary of our main contributions and Section VI draws directions for future work.

II. Background

Much work has been done in the field of movie data analysis. However data analysis of different types of movies have not been studied on much. Most of the work follows some approaches. When it comes to mention about them, we can say that the first approach is Recommender systems. There are many works and analysis on Recommender systems such as [1].

The second approach is item/content- based analysis. Like in [2]. It is also possible to work with item-based technics and reach that results. Another challenge for this field to work with content-based recommendation system [3]. In this project, by examining user profiles learning algorithms have been applied. Because of the difficulty of working with content-based; by classifying using keywords. They should be classified correctly otherwise they may be added to a wrong cathegory or they may be added to a wrong cathegory. This can change the

			Toplam:	331.929.371 TL	30.794.515
Sıra	Sıra	Dağıtım	Vizyon Tarihi	Toplam Hasılat *	Toplam Seyirci *
1	Mucize	Pin.	*01.01.15*	37.664.285 TL	**3.578.575**
2	Hızlı ve Öfkeli 7	UIP	03.04.15	31.943.552 TL	**2.942.888**
3	Kocan Kadar Konuş	UIP	20.03.15	20.725.662 TL	**1.912.701**
4	Selam: Bahara Yolculuk	Mars.	13.03.15	13.062.740 TL	**1.682.962**
5	Bana Masal Anlatma	UIP	09.01.15	16.883.440 TL	**1.576.979**
6	Aşk Sana Benzer	Mars.	23.01.15	14.519.685 TL	**1.404.858**
7	Yenilmezler: Ultron Çağı	UIP	01.05.15	12.538.428 TL	**1.022.268**
8	Yapışık Kardeşler	Mars.	30.01.15	10.385.533 TL	**1.014.417**
9	Sevimli Tehlikeli	WB	06.02.15	9.186.465 TL	**911.833**
10	Grinin Elli Tonu	UIP	13.02.15	10.462.201 TL	**875.585**
11	Son Mektup	Pin.	18.03.15	6.513.302 TL	**810.036**
12	Son Umut	Mars.	*26.12.14*	8.530.438 TL	**763.559**
13	Yusuf & Yusuf	WB	*26.12.14*	6.574.148 TL	**626.196**
14	6 Süper Kahraman	UIP	16.01.15	7.521.144 TL	**610.651**
15	Çarşı Pazar	Pin.	27.02.15	6.033.873 TL	**582.692**
16	Mandıra Filozofu: İstanbul	CF	13.03.15	5.361.585 TL	**502.607**
17	Ali Kundilli	TME	20.02.15	4.958.401 TL	**481.685**
18	Niyazi Gül Dörtnala	UIP	08.05.15	4.815.567 TL	**441.789**
19	8 Saniye	WB	27.02.15	4.611.292 TL	**421.070**
20	Köstebekgiller: Perili Orman	WB	23.01.15	4.335.751 TL	**409.382**
21	İçimdeki Ses	Pin.	30.01.15	4.509.294 TL	**403.641**
22	Hobbit: Beş Ordunun Savaşı	WB	*17.12.14*	4.626.189 TL	**365.361**
23	Sünger Bob Kare Pantolon 3D	UIP	06.02.15	4.517.272 TL	**364.853**
24	Kuralsız	TME	20.03.15	4.042.779 TL	**319.877**
25	Kod Adı: K.O.Z.	Mars.	13.02.15	3.194.978 TL	**313.361**
26	Senden Bana Kalan	Mars.	17.04.15	2.948.994 TL	**285.255**
27	Çılgın Dersane: Ada	TME	16.01.15	2.785.947 TL	**277.257**
28	Sindirella	UIP	13.03.15	3.274.797 TL	**268.068**
29	Güvercin Uçuverdi	Mars.	27.03.15	2.364.951 TL	**225.021**
30	Çakallarla Dans 3: Sıfır Sıkıntı	Mars.	*05.12.14*	2.108.426 TL	**204.715**
31	Evim	TME	27.03.15	2.479.834 TL	**199.144**
32	Jupiter Yükseliyor	WB	06.02.15	2.563.369 TL	**197.021**
33	Bir Varmış Bir Yokmuş	Mars.	06.03.15	1.959.392 TL	**177.518**
34	Kingsman: Gizli Servis	TME	13.03.15	2.246.967 TL	**173.000**
35	7 Cüceler	Mars.	30.01.15	1.719.527 TL	**156.996**
36	Enigma	Pin.	20.02.15	2.136.565 TL	**152.428**
37	Fatih'in Fedaisi: Kara Murat	Pin.	16.01.15	1.622.563 TL	**151.416**
38	Yedinci Oğul	UIP	30.01.15	1.961.355 TL	**146.870**
39	Kuzular Firarda	M3	17.04.15	1.401.977 TL	**135.936**
40	Aşk Olsun	Mars.	10.04.15	1.456.329 TL	**133.468**
41	Whiplash	M3	16.01.15	1.814.352 TL	**122.543**
42	Ayı Paddington	Mars.	*26.12.14*	1.415.055 TL	**120.749**
43	Asteriks: Roma Sitesi	Mars.	27.02.15	1.432.912 TL	**120.174**
44	Bizim Hikaye	ChF	27.03.15	1.143.130 TL	**115.704**
45	Müzede Bir Gece: Lahitteki Sır	TME	02.01.15	1.386.998 TL	**115.246**

Fig. 1. Boxofficeturkiye 2015 movie list

result or oblige to examine the parts of text for example when twitter data is tried to be interpreted the texts of tweets that are collected one by one should be examined one by one because they include string and markers. Language can pose a problem for the analysis stage.

The third one is analysis based on components(tweets, retweets, users, audiences, ratings and so on) because of that it makes analysis on real data it achieves success. We have also added the number of retweet to our analysis. By visualising the relation between follower-user-following can help us.[4] was another work about this issue and they presented a methodology for predicting retweets in Twitter. [5] is another example. At this works also they shown how social media can be utilized to forecast future outcomes, they used also Twitter and box-office data.

As we have mentioned before we worked on esoteric and entertainment types of movies and by consulting the data we collect, we could scrutinize them. That kind of studies have not mentioned about this assignation and so we have chosen it.

III. Data Collection

In this study we consider four cinema movies which two of them are dedicated to a special social community (esoteric) and the other two movies do not. We choose Kod Ad K.O.Z. (KOD), Selam Bahara Yolculuk (SBY), Icimdeki Ses (IS), Bana Masal Anlatma (BMA). KOD and SBY are esoteric movie samples, IS and BMA are non esoteric movie samples. We choose identical movies for a fair comparison which they have approximately same audiences number and all of them Turkey mentioned movies. For this selection we used boxofficeturkiye data which shows movies' release dates and total number of audiences.

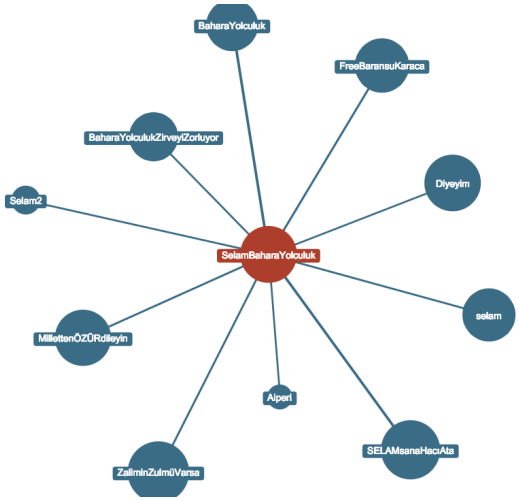

Fig. 2. Hashtagify result for Selam Bahara Yolculuk.

Movie	Date Range	# of Audiences	IMDB Rate	Hashtags
Bana Masal Anlatma	2 – 16 January	1.576.979	8.2	banamasalanlatma
Selam Bahara Yolculuk	6 – 20 March	1.682.962	9.5	selambaharayolculuk baharayolculuk selam2
İçimdeki Ses	23 Jan – 6 Feb	403.641	6.1	icimdekises
Kod Adı K.O.Z.	6 – 20 February	313.361	1.6	kodadikoz kodadikoz13subattavizyonda

Fig. 3. Summary of the movies' informations

After movie selection we started twitter data collection. We need most related hashtags and date ranges for twitter data collection. We used hashtagify.me web page for determining the all related hashtags with our selected movies. Figure.2 shows hashtagify result for SBY.

We collected tweets within the date margin one week before and after from the release date. Table.1 shows date ranges, total number of audiences, most related hashtags and IMDB rates.

After defining hashtags and date ranges we started twitter data collection. We could not use twitter API because of the API restrictions. Twitter API does not allow collecting tweets older than a month and API also restrict query number per hour. So we decided to make a web crawler for twitter search page. We use twitter advanced search web page for searching and determining url for each hashtag which will crawl later. Determined url list is added to submission file(Because some url has long text).

We wrote a python code with selenium, requests and beautifulsoup python modules for crawling search result web page html and parsing it. This python code (twitterCrawler.py) is available at the end of the report.

We run this code for every hashtag determined before. The code saves outputs as csv (comma separated values) file. The format of the csv file is as name of the user, date of the tweet, text content of the tweet, number of retweet, number of fav. Because of the twitter search web page shows only original tweets, we collect number of retweets and number of favs for finding actual tweet numbers. We also combine all hashtag results in a single csv file for each movie and we discarded

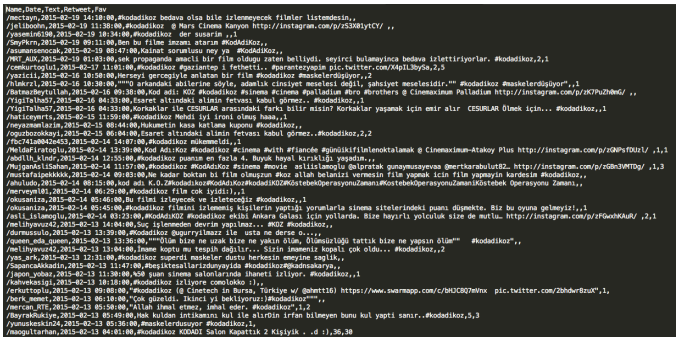

Fig. 4. Sample content of collected twitter data for KOD movie

repeated tweets, because some tweets include two or more hashtags at the same time. After combining and discarding process we get single csv file for each movie which includes all the related tweets' data within the determined date range. At the end of the first data collection part we have four csv files as banamasalanlatma.csv, icimdekises.csv, kodadikoz.csv and selambaharayolculuk.csv. Sample content of these files shown in the Figure.4.

For further analysis we decide to collect user network structure information from twitter. For this data collection we write an other python code (users.py) which makes queries to mobile.twitter.com to crawl followings and followers lists of the every single user. We filter user names from before collected twitter data and discard repeated users. The result lists used as an input for this python crawler. This python code saves outputs as two csv files for each movie which one for followings list and one for follower list. We try to use these outputs for visualizing each movies social network structure but because of the size of these files we could not visualize these network structures until now.

IV. Analysis

We analyze collected movie data under two categories. The first one name is Collected components data analysis which has tweets, users, retweets, audiences, ratings and so on. And another one is called Collected follower-followings data analysis.

A. Collected components data analysis

Figure.5 represents relation between tweets, users, and retweets. From this table, we can observe collected for four movies data. For analyzing of the data, first we seperated two categories, the first one had KOD and IS movies, and the second one had BMA and BSY movies which tempers to the number of audiences and each had both one of esoteric and entertainment movies. The second step for analyzing of the data is to seperate each movie data into the two parts which are before and after released dates. So from the table SBY-A meaned the number of Selam Bahara Yolculuk movies data which is collected in after the released date for it, and similar to this, BMA-B means the number of Bana Masal Anlatma movies data which is collected in before the released date, and the others have the same way. We collected data for two weeks, which is before and after one week from released date. In this

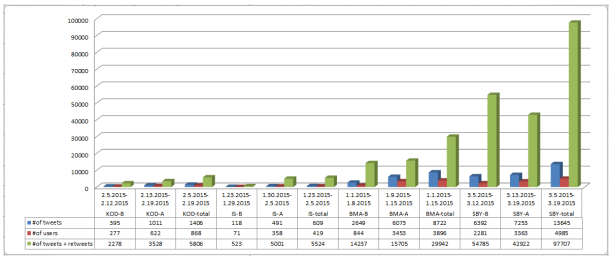

Fig. 5. Relation between tweets, users, and tweets + retweets

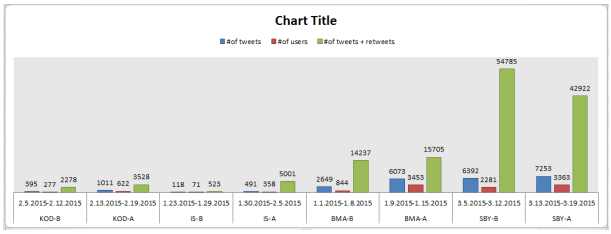

Fig. 6. Comparison between before and after

table the number of users meaned the number of different users who sent to tweets about this hashtags.

The first remarkable observation from Figure.5 is between the first category which is KOD-IS movie pairs. The ratio of the number of (tweets+retweets) for (KOD-A / KOD-B) is almost 1.5 times, nearly has the same, however for another movie from this category, the ratio is almost 10 times. Another observations about KOD-IS movie pairs have almost the same results. For instance, the fraction of the number of tweets or users for (after / before) is nearly same rate for each movies from the first category such as the ratio of number of tweets (KOD-A / KOD-B) almost 3 times, and number of tweets (IS-A / IS-B) is almost 3.5 times . From another aspect, Both KOD and IS movies have almost same number of audience but their total number of tweets, users are different(KOD has 3 times for the total number of tweets to compare with IS, and it has also 2 times for total number of users.)

The second notable observation for another category's movies which are BMA-SBY. In the same way as the first movie pairs, the ratio of the number of (tweets) for (SBY-A / SBY-B) is almost 1.5 times, nearly has the same, however for another movie from the second movie category, the ratio is almost 3 times. Another observations about SBY-BMA movie pairs have different results. For instance, the fraction of the number of (users) for BSY movie (after / before) is almost 1.5 like (tweets) observation, and for BMA ratio is nearly 4 times. From this category, the different point with the first movie pairs that caught our attention is that, it is clear to shows from next table, the ratio of the number of (tweets+retweets) for SBY-B is higher than SBY-A, in contrast for BMA movie, it has an opposite situation.

The third observation from Table I. It shows movie ratings, users who are rate the movies, and the number of audiences. Starting from this table, as we see clearly KOD is watched by 313.361 people, however the number of users from IMDB is really huge for this movie, and rating is baddish for it. But we observe other movies data form imdb and also Box-office,

Movies	IMDB-Ratings	IMDB-Users	BoxOffice-Audiences
KOD	1.6	14.924	313.361
IS	5.9	635	403.461
BMA	8.2	5.278	1.576.679
SBY	9.4	3.523	1.682.493

TABLE I. RELATION BETWEEN RATINGS, USERS AND TOTAL AUDIENCES

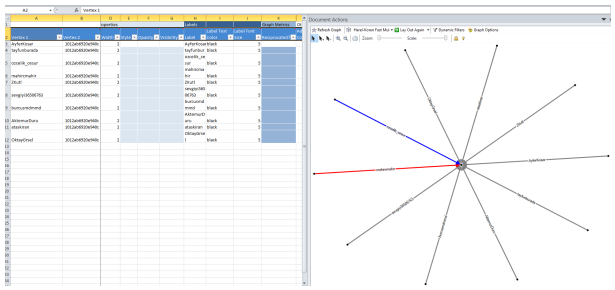

Fig. 7. NodeXL visualization of followers-users-followings relation

we obtain different result for them. For instance, SBY has the number of 1.682.493 audiences but its rating is goodish which is 9.4, and also BMA has 1.576.679 audiences, and rating is also well enough. We expect to the same or similar ratings, similar number of users from IMDB, and similar number of audiences for the same category movies. But from table, KOD is clearly difference to others.

And for another observation, in defiance of Figure.6 and Table I, we expect to relation between the number of audience and the number of total tweets, retweets. For KOD movie number of total (tweets+retweets) is 5806, and it has 313.361 audiences, however for IS (tweets+retweets) is 5524, and it has 403.461 audiences, so it is another ramerkable point, we have a contradiction or unexpected results. And for SBY number of total (tweets+retweets) is 97707, and it has 1.682.493 audiences, for BMA number of total (tweets+retweets) is 29942, and it has 1.576.679 audiences. When we compare these two, SBY movie's data are higher than BMA, it is also expected but 97707 is quite big as a BMA movie.

B. Collected follower-followings data analysis

Here is another observation with using twitter data. We did not collect data for all movies, we success to collect IS-KOD movies data this week. So whole data is not completed, yet. However, we want to show how we observe data for searching follower-user-following data. Figure.7 shows only for user1 with its followers.

We give more details in Section VI.

V. CONCLUSION

Before this study, we expected the ratio of the tweet activities before and after release dates could be enough for distinguishing, noticeable increase in the number of users after the release date for entertainment type movies, in contrast, approximately same number of users after the release date for esoteric type movies, collected twitter data could be enough for obviously distinguishing esoteric and entertainment movies.
We contributed to analyzed collected data with using Twitter, we success semi-consistent results. It is related to the number

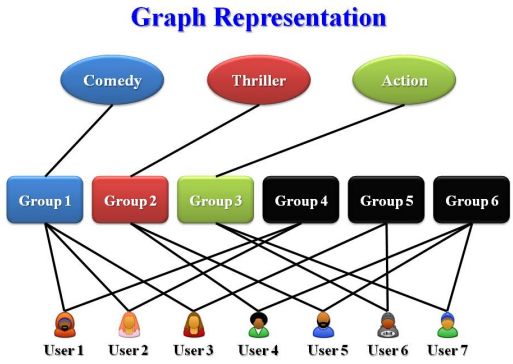

Fig. 8. Graph representation for movie types, and users

of collection data, as we mentioned in Section III, we collected only two weeks(it is dependence to time, and Twitter policies) so we want to expand time interval, such as four weeks instead of two weeks.

VI. FUTURE WORK

- Followers and followings twitter data collection is not completed yet.
- After the friend list collected, we analyze social network between users.
- For deeper analysis, we can increase the data collection date range.

In addition to these, to be inspired from related work; we want to categorize our collected/analyzed data similar to Figure.8. In our project we have only two type of movies, and we want to collect these type of movies users' . It is another representation for observation.

Another challenge for this project could be represent like bipartite graph between movies, and users. Also we can observe this one using Low-Rank Matrix Factorization.

Finally, we continue to collect followers-followings data, and we want to visualize them with users. It is also beneficial for analyzing movie data. Because we want to search and find relation between for instance we collect user1 and its followers-followings, then we observe are there any user who is in also another user's follower-following lists. It another way to learn the accuracy of movie data.

REFERENCES

[1] B. Sarwar, J. Konstan, J. Riedl and G. Karypis, *2nd ACM conference* Analysis of Recommendation Algorithms for E-Commerce .

[2] B. Sarwar, G. Karypis, J. Konstan and J. Riedl, *WWW10 May*Item-Based Collaborative Filtering Recommendation AlgorithmsHong Kong, 2001.

[3] M.J. Pazzani and D. Billsus, *Adaptive web* Content-Based Recommendation SystemsBerlin Heidelberg, 2007.

[4] T.R. Zaman, R. Herbrich, J. Gael and D. Stern, *Workshop*Predicting Information Spreading in Twitter , 2010.

[5] S. Asur and B. A. Huberman, *Agent Technology(WI-IAT), 2010 IEEE* Predicting the Future With Social Media 2010.