Finding Correlation between Content Based Features and the Popularity of a Celebrity on Twitter

Muhammad Asghar
Dept. of Computer Science & IT
The Islamia University of Bahawalpur

M. Faheem Mushtaq
Dept. of Computer Science & IT
The Islamia University of Bahawalpur

Hina Asmat
Dept. of Computer Science & IT
The Islamia University of Bahawalpur

Malik Muhammad Saad Missen
Dept. of Computer Science & IT
The Islamia University of Bahawalpur

Tareef Ali Khan
Dept. of Computer Science & IT
The Islamia University of Bahawalpur

Saleem Ullah
Dept. of Computer Science & IT
The Islamia University of Bahawalpur

***Abstract* – Popularity of a celebrity having twitter account is generally estimated by the number of twitter account holders following him on twitter. In this paper, we explore some content based features to evaluate their role for estimating the popularity of a celebrity. We try to find out the co-relation of celebrities' popularity with number of other features like frequency of tweets posted by a celebrity and the relevancy of tweets to their domains. We take support of named entity recognition for later task. For the data collection, we collected about 60K tweets of 60 different celebrities with the help of API method provided by twitter. The current study aims to reveal many unseen patterns existing in celebrities' twitter usage that includes computing per day tweet frequency and finding out how it correlates with number of followers. Also we find out the correlations of tweets with domains of the celebrities.**

Keywords* – *Short text messaging, mobile health-care, social networking, rural health

I. INTRODUCTION

Twitter has become widely debatable and most observant way into a user's attention. With an amazing total number of about 500 million existing users of Twitter it has become widely used to share all kind of information's from like latest news to various links including something as remote as pictures of one's trips. Twitter is hubs with large amount of information's in different forms like text, videos, pictures etc. Twitter is used by people belonging to different domains of life i.e. students, researchers, actors, sportsmen, politicians, etc. Twitter users belong to all age's groups who use different sources to access Twitter services. There are about 60% users that access Twitter from their cell phones, 28% retweets contain message of "kindly please retweets this tweet". Twitter has 288 million monthly active members that make it widely growing social networking site. There are around 400 million tweets post on daily basis, the average posts on twitter is 208 tweets per user's account. Figure below is showing interesting facts of Twitter in 2013.

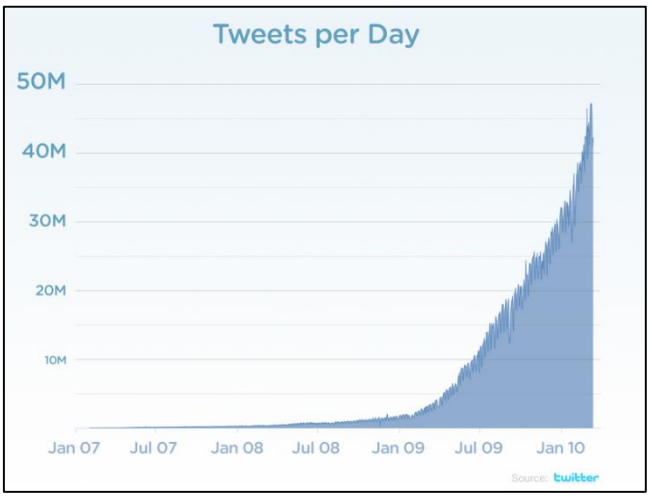

Figure 1: Total number of tweets per day over the years[1]

When big amount of populations of world keep posting information in the form message on Twitter daily bases, we effectually have important collections of latest news, user's feelings, and interesting ideas about everything in world, we have bands of favorite music, foods, what's happening in sports and immediate responses to trending topics and so on. Celebrities of different domains keep posting their tweets to let their fans know about their views and let them stay updated about latest proceedings [1]. The popularity of a celebrity can be estimated by different metrics depending upon the domain to which a celebrity belongs. For example, the admiration of a film actor can generally be valued by the total of block buster films he has appeared in while the admiration of a columnist can be valued by the aggregate of people reading his column. Twitter provides us another way of estimating this popularity where it can generally be estimated by the number of people following them on their twitter account [2]. However, there are many other features that could be explored to find out if they can be in estimating

[1] http://techcrunch.com/2010/02/22/twitter-50-million-tweets-day/

the popularity of a celebrity. In this thesis, we especially focus on finding out correlation between popularity of a celebrity and other content-based evidences.

Admiration of a celebrity on Twitter can be assessed by number of his/her followers which is a user profile feature. However, there is a need for some content based features that could be used to estimate the popularity of a celebrity on Twitter. The main objective of this thesis is to study content-based features that could play their role in estimating the popularity of a celebrity. However, following objectives have been defined:

1. To identify some content based features that could be useful in estimating the popularity of a celebrity.
2. To find correlations between these features and popularity of a celebrity.
3. To find how these features behave among celebrities of same category (i.e. sports, actors, politicians, etc.) and celebrities of different categories.
4. To find out if a celebrity himself is popular on twitter or it's his/her tweets that are making people interested in following him and to see how this process is observed in different categories.

This study may serve as a base for conducting further researches to study regarding using various types of social media on the world of internet.

A. Data Collection

Major focus of previous chapters is finding hidden patterns in twitter's follow-following process. To perform this we need considerable size of twitter data collection .To attain this goal we need to use Twitter APIs that are discussed in next section.

Twitter APIs: API is Application program Interface can be defined as instruction set which helps developers to interact with some type of technology. Twitter's API allow external developers to develop technology which depend upon Twitter's data [3].

We have used Twitter APIs (like [4] and [5]) and downloaded tweets of celebrities belonging to different categories. We decided to choose at least 10 celebrities from each category. Following categories were chosen for our data collection.

- Actors
- Anchor Person
- Businessmen
- Politician
- Sportsmen
- Writers

Top celebrities were chosen among each category by searching on Google search Engine with queries (for example, "top 10 actors of year 2013"). List of these celebrities for each category is given below in the table 1. Total 60 K tweets were crawled (i.e. 1000 tweets per celebrity) which makes a quite large number of tweets collection.

Table 1: List of celebrities for each category (with 1000 tweets for each)

Category	Celebrities
Actors	Abhishek Bchchan
	Anupam Kher
	Karina Kapoor
	Salman Khan
	Shahrukh Khan
	Angelina Jolie
	Ashton Kutcher
	Kristen Stewart
	Leonardo Dicaprio
	Tom Cruise
Anchor Person	Amy Wood
	Asma Shirazi
	Courtney Friel
	Fred Cunningham
	Hamid Mir
	Javed Chaudhry
	Kashif Abbasi
	Maggie O'Mara
	Nasim Zehra
	Sean Bailey
Businessmen	Bill Gates
	David Letterman
	Donald J. Trump
	George Lucas
	Lance Armstrong
	Martha Stewart
	Michael Bloomberg
	Micheal Bloomberg
	Oprah Winfrey
	Paul McCartney
Politicians	Abdullah Gul
	Barack Obama
	Dilma Rousseff

	Dmitry Medvedev
	Dr. Manmohan Singh
	Enrique Peña Nieto
	Juan Manuel Santos
	Maryam Nawaz Sharif
	Rania Al Abdullah
	Sh. Rasheed
Sportsmen	David Flatman
	Graeme swann
	Jack Wilshere
	Jonathan Agnew
	Kelly sotherton
	Lee westwood
	Leon knight
	Mark bright
	Rio Ferdinand
	Steve Nash
Writers	Brad Meltzer
	Chuck Palahniuk
	Desse Sarah
	Eve Mayer
	James Patterson
	Jodi Picoult
	Margaret E. Atwood
	Maureen Johnson
	Neil Caiman
	Seth Grahame-Smith

II. Correlation between Popularity and Average Number of Tweets

In this section we analyze the correlation between average per day tweets and followers category-wise. Computing per day tweet frequency and see how it correlates with number of followers (i.e. popularity). Per day tweet frequency here means that total number of sent and followed tweets in 24 hours and the topic on which it was tweeted such as hardware, mobile electronics and belongs to celebrity or not and to what extent it is followed.

Table 2: Table listing average tweets per day with number of followers (2K) for each celebrity

Category	Celebrities	Tweets/day	Followers (2K)
Actors	Abhishek Bchchan	9	1498
	Anupam Kher	6	938
	Karina Kapoor	3	3
	Salman Khan	4	2559
	Shahrukh Khan	9	2647
	Angelina Jolie	15	133
	Ashton Kutcher	2	7544
	Kristen Stewart	2	244
	Leonardo Dicaprio	1	3943
	Tom Cruise	13	18
Anchor Person	Amy Wood	42	12
	Asma Shirazi	6	64
	Courtney Friel	5	12
	Fred Cunningham	70	5
	Hamid Mir	7	161
	Javed Chaudhry	2	12
	Kashif Abbasi	19	7
	Maggie O'Mara	13	2
	Nasim Zehra	10	50
	Sean Bailey	9	0
Businessmen	Bill Gates	2	6680
	David Letterman	11	2
	Donald J. Trump	25	1195
	George Lucas	1	4
	Lance Armstrong	4	1964
	Martha Stewart	5	1431
	Michael Bloomberg	3	1
	Micheal Bloomberg	12	10743
	Oprah Winfrey	2	806
	Paul McCartney	2	806
Politicians	Abdullah Gul	5	1892
	Barack Obama	6	19053
	Dilma Rousseff	3	990
	Dmitry Medvedev	8	338
	Dr. Manmohan Singh	7	417
	Enrique Peña Nieto	4	1120
	Juan Manuel Santos	4	1111
	Maryam Nawaz Sharif	12	71
	Rania Al Abdullah	5	1472
	Sh. Rasheed	25	81
	David Flatman	13	8

Sportsmen	Graeme swann	3	288
	Jack Wilshere	5	122
	Jonathan Agnew	2	344
	Kelly Sotherton	6	20
	Lee Westwood	8	302
	Leon knight	2	246
	Mark bright	13	40
	Rio Ferdinand	7	2411
	Steve Nash	3	873
Writers	Brad Meltzer	4	17
	Chuck Palahniuk	4	241
	Desse Sarah	8	115
	Eve Mayer	9	46
	James Patterson	3	22
	Jodi Picoult	5	34
	Margaret E. Atwood	11	218
	Maureen Johnson	29	46
	Neil Caiman	22	952
	Seth Grahame-Smith	2	5

The Pearson coefficient between two variables x and y commonly is defined as the covariance of the two variables divided by the product of their respective standard deviations [6, 7]. In this section, we analyze how these two variables i.e. popularity of a celebrity and his frequency of posting tweets relate to each other according to category of domains.

$$\rho_{xy} = \frac{cov(x,y)}{\sigma x \ \sigma y}$$

Where $cov(x,y)$ is the covariance and σ is the standard deviation.

Table 3: Pearson Coefficient correlation between number of tweets per day and number of followers

Sr. No.	Category	Pearson Coefficient
1	Actors	-0.47597
2	Businessmen	0.15855
3	Anchor	-0.29428
4	Politician	-0.17159
5	Sports Person	-0.20234
6	Writers	0.468257

In the Table 3 given above, we have provided the Pearson co-relation measure between average tweets posted and popularity for celebrities of different categories. This table reveals very interesting results for different categories. It can be seen that the highest degree of correlation exist among writers while lowest degree of correlation is marked for actors. These results can be justified by the argument that actors are generally liked by everyone while writers are only liked by people who are interested in authoring i.e. number of followers for actors could be larger than number of followers for writers (as it can be seen in tables 3,6). The average number of tweets per day might vary according to the work schedule of a celebrity but it is the number of followers that might be able to affect this correlation and this is true in this case. Actors are followed by huge number of people while writers are only followed by those who are interested in their writings or only by those who have chosen the writing as their career.

Therefore, one can conclude from this table that highest degree of correlation exists between average number of tweets and popularity of a celebrity for Writers while it is very low for Actors.

III. Correlation between Popularity and Topical Nature of Tweets

In this section, we focus on our second research task i.e. finding correlation between topical categories of tweets posted by a celebrity and his/her popularity. The idea behind this research task is to know whether the followers love to follow only those celebrities who write more about their own field (i.e. sports, films, etc.) or there is no such relation between these two variables. The results of this task give us an idea about whether followers are really interested in the celebrity himself or his field. This task is really important in knowing the popularity of a celebrity regardless of his/her tweets. Therefore, topical nature score for each celebrity was computed by computing the percentage of relevant tweets out of collected 1000 tweets and later on this score was correlated with number of followers for that celebrity to find out how these two numbers correlate.

Table 4: Pearson Coefficient correlation between topical nature score of tweets of a category and number of followers

Sr. No.	Category	Pearson Coefficient
1	Actors	0.001106
2	Businessmen	-0.43995
3	Anchor	0.478717
4	Politician	0.29671
5	Sports Person	0.439368
6	Writers	-0.35122

In table 4, the correlation between popularity of a celebrity and its topical tendency towards domain of the celebrity is shown. The

major objective is to find out if it is the nature of tweets that forces people to follow the celebrities on twitter. The idea is that if we find a strong correlation between these two variables then it is evident that people following a celebrity must be interested in reading the tweets he/she posts on twitter.

On observing the numbers as given in above table, it's clear that sportsmen and anchors are two categories where we can find high correlation scores between popularity and tweet nature while lowest score for this correlation was found for businessman category. While justifying these results, we can argue that people generally are more eager to know about sportsmen and anchormen and their thoughts while caring very little about others.

IV. Conclusion

In this paper, we have analyzed a large collection of tweets of celebrities to find out the correlation between their popularity with average number of tweets per day and topical nature of their tweets. We conclude from our results that highest degree of correlation exists between average number of tweets and popularity of a celebrity for Writers while it is very low for Actors. Similarly, we used Pearson Correlation metrics for finding correlation between topical nature of tweets and popularity of a celebrity. While investigating this correlation, it was found that sportsmen and anchors are two categories where we can find high correlation scores between popularity and tweet nature while lowest score for this correlation was found for businessman category.

REFERENCES

[1] Ampofo, Lawrence, Anstead, Nick, and O 'Loughlin, Ben. (2011). Trust, confidence, credibility: Citizen Responses on Twitter to opinion polls during the 2010 UK general election. Information, Communication & Society, 14(6), 850-871

[2] Ausserhofer, Julian, and Maireder, Axel. (2013). National Politics on Twitter. Structures and topics of a networked public sphere. Information, Communication & Society, 16 (3), 291-314

[3] Van Dongen, Stijn and Enright, Anton J. , Metric Distances Derived from Cosine Similarity and Pearson and Spearman correlations, Published in CoRR, 2012

[4] Meeyoung Cha, Hamed Haddadi, Fabrício Benevenuto, P. Krishna Gummadi: Measuring User Influence in Twitter: The Million Follower Fallacy. ICWSM 2010

[5] Go, A.; Bhayani, R. & Huang, L. (2009), 'Twitter Sentiment Classification using Distant Supervision', Processing, 1--6.

[6] Pearson Product-Moment Correlation Coefficient. n.d. In Wikipedia. Retrieved July 06, 2014, from *https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient*

[7] Malik M. Saad Missen, Tareef Ali Khan, Hina Asmat, Nadeem Salamat, Nadeem Akhtar: Mobile SMS based Self-Medication Support for Health Care, Vol. 10 Issue 3, International Journal of Computer Science Issues 2013, May 2013