Ranking of Tweets for Digital Marketing

Denzil Pereira
St. Francis Institute of Technology, Borivali (west), Mumbai, India.

Jason Misquitta
St. Francis Institute of Technology, Borivali (west), Mumbai, India.

Priyank Cerejo
St. Francis Institute of Technology, Borivali (west), Mumbai, India

ABSTRACT
With the evolution of online social networking and micro-blogging media, two major changes have occurred in the landscape of the Internet usage. Primarily, the Internet is replacing traditional media like television and print media as a source for obtaining news and information about current events. Secondly, the Internet has provided a platform for common people to share information and express their opinions. Quick response time and high connectivity speed have fueled the propagation and dissemination of information, by users on online social media services like Facebook, Twitter and YouTube. Twitter is a micro-blogging service, which has gained popularity as a major news source and information dissemination agent over last few years. Users on Twitter, create their public / private profile and post messages (also referred as tweets or statuses) via the profile. The maximum length of the tweet can be 140 characters. Each post on Twitter is characterized by two main components: the tweet (content and associated metadata) and the user (source) who posted the tweet. Studies have explored and highlighted the role of Twitter as a news media and a platform to gauge public sentiments.

General Terms
Analysis, Algorithm.

Keywords
Application Program Interface, Sentiment Analysis, BM-25 Algorithm.

1. INTRODUCTION
With the rise of new technologies in the field of the internet and social media, the popularity and importance of numerous social media platforms have risen to new levels, as more people spend more time online and companies follow their potential customers because of its ease of use, speed and reach. Social media is fast changing the public discourse in society and setting trends and agendas in topics that range from the environment and politics to technology and the entertainment industry. One such social media platform that has seen an explosive rise in popularity is Twitter.

Approximately 1 billion tweets are generated by Twitter users every five days. With so many people tweeting about their various opinions about subjects ranging from toothpaste to the latest Apple products, Twitter is a rich source of real-time information regarding current societal trends and opinions.

The project aims at investigating Twitter's power at predicting real-world outcomes. The chatter of a community can be used to make quantitative predictions that outperform those of artificial markets. Using our project, various firms can get reviews about their newly launched products. Thus, this project can be used to increase digital marketing accuracy.

2. LITERATURE SURVEY
2.1 Sentiment analysis and Opinion Mining
Raisa Varghese and Jayasree M [5] describe Sentiment Analysis and opinion mining as subfields of machine learning. They are very important in the current scenario because, lots of user opinionated texts are available in the web now. This is a hard problem to be solved because natural language is highly unstructured in nature. The interpretation of the meaning of a particular sentence by a machine is tiresome. But the usefulness of the sentiment analysis is increasing day by day. Machines must be made reliable and efficient in its ability to interpret and understand human emotions and feelings. Sentiment analysis and opinion mining are approaches to implement the same.

2.2 Credibility of Content
Aditi Gupta, Ponnurangam Kumaraguru [2] has described the increasing popularity of social system and its credibility. Twitter has recently merged as a popular social system where users share and discuss about everything, including news about events. Twitter is increasingly becoming a system for obtaining real time information and source for news and latest trends. Twitter emerged as an excellent means to disseminate information to a large user community in the shortest time. On the contrary, this very open uncontrolled nature of twitter service makes micro blogging vulnerable to false information from malicious users. Consequently, it is important to formulate sophisticated methods for analysis of relevance and trustworthiness for ranking tweets. Ranking that considers content based features, place the most credible and popular tweets in the top slots.

3. PROPOSED WORK
3.1 System description
Previous research explored the credibility on Twitter with respect to trending topics. Credibility of a topic on Twitter may not be a good indicator to judge the credibility of the content of the tweet. Thus, assessment techniques are required at the atomic level of the information on twitter, i.e. at a tweet level.

In this project, credibility is assessed at tweet level by considering content and context specific based features that are used to rank the tweets according to the credible information contained in tweets. After the tweets are extracted using the Twitter API (Application Program Interface), the content of each tweet is compared to pre-defined dictionary of keywords. Using BM-25 (Best Match-25) algorithm, each tweet is allotted a respective BM score which will be a key indicator in ranking of these tweets. The tweets are then ranked and displayed on the browser GUI in the decreasing order of their BM score.

To implement a system that extracts real time tweets from Twitter, analyse the sentiment of the tweets and display the result on browser GUI for better decision making.

1. The system must involve the most important step, i.e. pre-processing. This step must eliminate any irrelevant content like links, URLs and Tweet mentions.
2. The classifier must employ an appropriate feature selection model.
3. An appropriate trade-off between performance and accuracy must be achieved.
4. Analysis must be real-time.

3.2 Working Flow

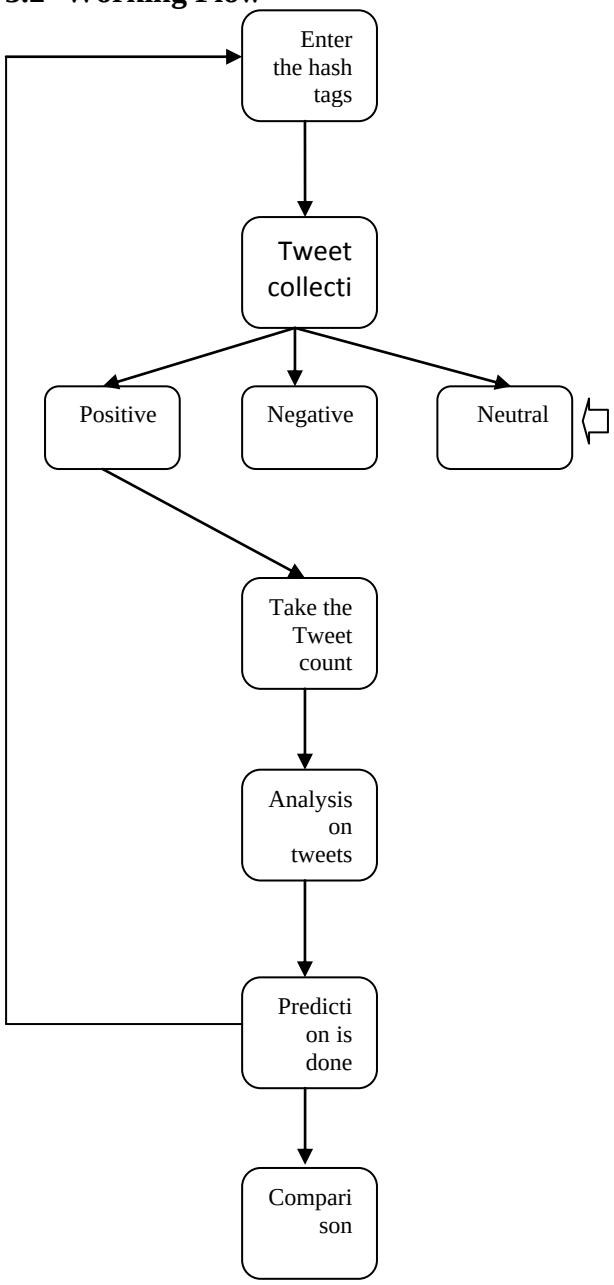

Figure 1: Work Flow

3.3 BM-25 Algorithm

In information retrieval, Okapi BM25 (BM stands for Best Matching) is a ranking function used by search engines to rank matching documents according to their relevance to a given search query. It is based on the probabilistic retrieval framework developed in the 1970s and 1980s by Stephen E. Robertson, Karen Spärck Jones, and others. [1]

The name of the actual ranking function is BM25. To set the right context, however, it usually referred to as "Okapi BM25", since the Okapi information retrieval system, implemented at London's City University in the 1980s and 1990s, was the first system to implement this function.[1]

BM25 is a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document (e.g., their relative proximity). It is not a single function, but actually a whole family of scoring functions, with slightly different components and parameters. One of the most prominent instantiations of the function is as follows. [1]

Given a query Q, containing keywords $q_1, \ldots, q_n$, the BM25 score of a document D is:

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot\left(1-b+b\cdot\frac{|D|}{\text{avgdl}}\right)},$$

where $f(q_i, D)$ is q_i's term frequency in the document D, $|D|$ is the length of the document D in words, and $avgdl$ is the average document length in the text collection from which documents are drawn. k_1 and b are free parameters, usually chosen, in absence of an advanced optimization, as $k_1 \in [1.2, 2.0]$ and $b = 0.75$. $\text{IDF}(q_i)$ is the IDF (inverse document frequency) weight of the query term q_i. It is usually computed as:

$$\text{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5},$$

where N is the total number of documents in the collection, and $n(q_i)$ is the number of documents containing q_i.[1]

4. IMPLEMENTATION

The initial stage in the project is to collect tweets from twitter website which is https://twitter.com. The tweets will be collected using a code with the help of an API (Application Program Interface) which the Twitter website provides for the developers.

The search API has a fairly rich set of operators that can filter results based on attributes like location of sender, language, and various popularity measurements. The Search API allows users to query for Twitter content. This includes finding tweets for a set of keywords, users or location posted in the past. Once the tweets are collected, they are stored in the database and sentiment analysis is performed on them. For Simplicity and easy operations we have used the WAMP server for storing and retrieving data along with eclipse IDE

for the compilation of API code. Sentiment analysis is nothing but distinguishing the tweets into positive, negative and neutral one. After the tweets were extracted using the Twitter API, the content of each tweet was compared to pre-defined dictionary of positive and negative keywords. Using the BM-25 algorithm, each tweet was allotted a respective BM score which was the key indicator in ranking of those tweets. The tweets were then ranked and displayed on the browser GUI in the decreasing order of their BM score as higher BM score indicated a more positive tweet.

Thus, the impact can be concluded from the event, subject or a product on various users that are connected to twitter by processing the displayed ranked tweets.

5. RESULTS

Table 1. Result Analysis

Test Case	Expected Result	Actual Result
Searching Top Ten Trending Topics	Top Ten Trending topics should be displayed in real time.	Top Ten Trending Topics were displayed in real time.
Search for a hashtag	The Specified Hashtag should be Searched and result should be displayed.	The Specified Hashtag is Searched and result is displayed.
Ranked Tweets	The tweets should be ranked according to BM score and displayed.	The Tweets are ranked according to BM score and displayed.

6. CONCLUSION AND FUTURE SCOPE

For a successful application of the proposed solution, tweets of high impact events must be ranked according to the credible information contained in them. In this project, tweets of users, posted on event were collected to perform sentiment analysis to obtain credible tweets. After obtaining credible tweets of an event, message features of those tweets were extracted and ranked according to BM-score

The proposed solution supports the administrator to provide the tweets of an event that contains credible information to the users who require credible information of an event. In this project, credibility was assessed at tweet level by considering content and context specific based features that were used to rank the tweets. The work, presented in this project introduces possible direction for further work. Finally, integration of enterprise hardware servers and large databases would be effective to rank considerably large volume of tweets of high impact events. The impact of the event or subject on various users that are connected to twitter can be concluded by processing the displayed ranked tweets. Data Mining can be performed on produced output and resulting analysis can be sold to various service providing firms. This project is also capable of generating revenue by selling the annotated data to various firms if implemented on large scale, thus providing Marketing Value

6. ACKNOWLEDGMENTS

We wish to offer our sincere thanks to each and every person who has helped us either directly or indirectly during the course of this paper.

Above all we wish to thank our Project Guide, Mr. Pramod Shanbhag for his valuable assistance and advice. Special thanks to the faculty member of the Information Technology Department of St. Francis Institute of Technology for their approval and guidance.

We also wish to express our deepest gratitude to our fellow students and friends who have contributed in some way to our project.

8. REFERENCES

[1] OkapiBM25algorithm.https://en.wikipedia.org/wiki/Okapi_BM25.

[2] Aditi Gupta, Ponnurangam Kumaraguru, "Credibility Ranking of Tweets during High Impact Events" July 2012 precog.iiitd.edu.in pp. 4-5

[3] Bo Zhang, Jinchuan Wang, Lei Zhang, "A Tweet-Centric Algorithm for News Ranking" 8-11 July 2013 IEEE pp. 190 - 195

[4] Cappelletti, R. , Sastry, N. , "IARank: Ranking Users on Twitter in Near Real-Time, Based on Their Information Amplification Potential" 14-16 Dec. 2012 IEEE pp. 70 - 77

[5] Raisa Varghese , Jayasree M, "A Survey on Sentiment Analysis and Opinion mining", IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308

[6] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report, Stanford.

[7] Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of LREC.

[8] Luciano Barbosa and Junlan Feng, 2010. Robust sentiment detection on twitter from biased and noisy data. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 36–44.

[9] Sysomos (2010) Exploring the use of Twitter Around the World. Available: http://www.webcitation.org/5tSpJ4fI4.

9. APPENDIX

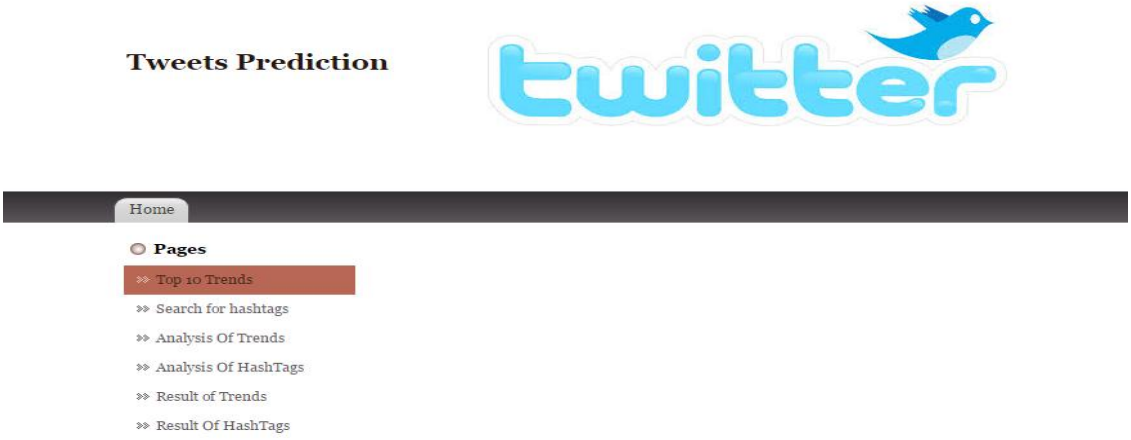

Figure 2: User Inteface

+ Options

	id	Trend	ScreenName	Created_At	Tweetid	Tweet	positiveCount	negativeCount	BMCount
Edit Copy Delete	1	#AAPKaSting	gk_2000_2007	2015-03-29 16:02:45	582128252400992256	@narendramodi Sir you are incompetent and you have...	0	0	0
Edit Copy Delete	2	#AAPKaSting	sunilrastogi84	2015-03-29 16:01:39	582127976961089536	#AAPKaSting stung by sting...	0	0	0
Edit Copy Delete	3	#AAPKaSting	swatwal75	2015-03-29 15:59:56	582127545241325570	#AAP MP Mann stayed at Ashoka Hotel during session...	0	0	0
Edit Copy Delete	4	#AAPKaSting	manik_manocha	2015-03-29 15:59:43	582127488727281664	#AAPtards will be very happy as they get #CWC15Fin...	0	0	0
Edit Copy Delete	5	#AAPKaSting	GanghorC	2015-03-29 15:59:04	582127324604198912	AAP has been one wreck of a party #AAPWar #AAPKaSt...	0	0	0
Edit Copy Delete	6	#AAPKaSting	gk_2000_2007	2015-03-29 15:58:13	582127113521606656	#AAPWar #AAPKaSting Has Modi done anything to imp...	0	0	0
Edit Copy Delete	7	#AAPKaSting	gk_2000_2007	2015-03-29 15:56:59	582126802706939904	#AAPWar #AAPKaSting Modi is doing everything anti...	0	0	0
Edit Copy Delete	8	#AAPKaSting	tushar121212	2015-03-29 15:56:44	582126739351957505	AAP leader Kumar Vishwas Caught sleeping with a AA...	0	0	0
Edit Copy Delete	9	#AAPKaSting	gk_2000_2007	2015-03-29 15:56:25	582126658158612480	#AAPWar #AAPKaSting People have understood that M...	0	0	0
Edit Copy Delete	10	#AAPKaSting	ls2008	2015-03-29 15:56:01	582126558791503872	Y AAP volunteers who die on JanLokpal silent on AA...	0	0	0
Edit Copy Delete	11	#AAPKaSting	AnuragDubey841	2015-03-29 15:55:28	582126419762802689	When @ArvindKejriwal was born Doctors said Mubarak...	0	0	0
Edit Copy Delete	12	#AAPKaSting	ibirujha	2015-03-29 15:55:23	582126398774460416	Waiting for kavi maharaj sting #AAPKaSting	0	0	0

Figure 3: Database for Collected Tweets

Tweet	BM Score
If NZ win today then mission #WeWontGiveItBack is a success. Discuss.	9.51175
#Respect u(Indian Team) Boyz 4R best performance 2 try 2 keep d Pledge #WeWontGiveItBack bt Oz Took it way nvr mind #WillFightBack @msdhoni	8.22966
#WeWontGiveItBack #wewilltakeit only suit to Aussies...they have always bn the best side #cwc15 #CWC15Final #AUSvNZ	8.22966
Missed on the dream of running his last run as winning run to take team home.. #WeWontGiveItBack	8.10309
From #WeWontGiveItBack to #WeTookItBack the journey is very exciting and hilarious #CWC15	8.10309
Smith hits d winning runs... N some at the start of the tourney said #weWontGiveitBack... Aussies didn't took it back.. Jus won it back..!!!	8.10309
Now if Australia winning 5th Worldcup claims #WeWontGiveItBack to baat banti b hey, india aiween uchal raha ta. Congrats Aussies. #AUSvNZ	8.10309
Now if Australia winning 5th Worldcup claims #WeWontGiveItBack to baat banti b hey, india aiween uchal raha ta. Congrats Aussies. #AUSvNZ	8.10309
Actually #WeWontGiveItBack suits better for #Aus I mean they won it 99-07 only we took it in 2011 and they have snatched it back ??	8.01211
There goes the #WeWontGiveItBack in trash!! Aussies Congratulations!! You deserved this win!! #AUSvNZ	7.99161
Irony.. Our WC campaign was #WeWontGiveItBack and Srini is going to be the one handing the trophy to the winners #CWC15Final #AUSvsNZ	7.99161
Aussies win the WC.. Starsports wud be saying #WeWontGiveItBack dum h toh le k dikha.. India se.. *Hides cup under the bed*#AUSvNZ	7.99161
@BLACKCAPS Don't worry, India had won the 1983 world cup making 183 in the first innings. You'll also win! ????? #WeWontGiveItBack #CWC15Final	7.99161
Skill, No Emotions will win you the #cwc15 ! #WeWontGiveItBack #StarSports @bcci @icc #India	7.99161

Figure 4: Ranked Tweets