HashJacker- Detection and Analysis of Hashtag Hijacking on Twitter

Nikita Jain
IIIT Delhi

Pooja Agarwal
IIIT Delhi

Juhi Pruthi
IIIT Delhi

ABSTRACT

Social media has become a multi purpose platform where users tend to discuss their common topics of interest as well as utilise it for endorsement and campaigns.Twitter being one of the most widely used social media platforms,provides a mechanism known as Hashtag which enables diversified online users, having coinciding interests to connect with each other. Hashtagging a tweet provides the most effective way of promoting a content as it steadily acquires the reader's attention because of it's symbolic denotation.With increasing hits on a particular Hashtag, it comes under one of the ”trending topics”.Soon this trending topic becomes a line of sight for people using Twitter. Users tend to hijack these popular Hashtags for distributing un-related content, spam, negative sentiments to tarnish the intended motive of Hashtag making its presence counter-productive. We propose a tool: HashJacker which detects and analyses hijacking of Hashtag tweets.Also, we have discussed best practises to circumvent wrecked Hashtag tweets.

Keywords

Hashtag, tweet, tf-idf, trend, hijacking.

1. INTRODUCTION

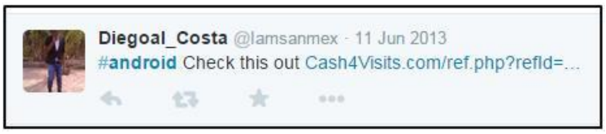
A Twitter hashtag is simply a keyword phrase, spelled out without spaces, with a pound sign (#) in front of it.It ties the conversations of different users into one stream, which one can find by searching the hastag in Twitter Search[1].For example, #android. We define hashtag hijacking as misuse of a hashtag for the purpose it is not intended to.Hijacking hashtag can happen through following ways :

- Attaching an abusive link
- Attaching an unrelated link
- Discussing unrelated content
- As megaphone to start random conversation

Hashtag reflects an industry or branded keyword that is interesting to the community.This in turn inclines the user to check out the rest of the conversation happening around that hashtag regardless of the fact whether it is in alignment with the hashtag or not.This instigates the brand competitors and spammers to hijack the trending hashtag for business gain and defamation.

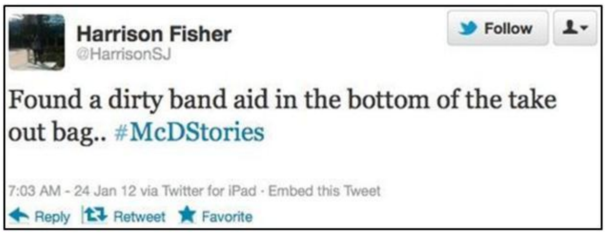
In addition to the above stated following are also the reasons why hashtag tweets are targeted :

- To seek attention and make once junk popular
- To attack the business of a particular popular brand where the detractors express their sentiments in a sarcastic or snarky way
- Posting abusive and contaminated content on social forums via popular hashtags
- Posting unwanted URL's through trending hashtags spamming the social media to a great extent



Diegoal_Costa @Iamsanmex · 11 Jun 2013
#android Check this out Cash4Visits.com/ref.php?refId=...

Fig 1: Hijacked tweet with #Android



Harrison Fisher
@HarrisonSJ

Found a dirty band aid in the bottom of the take out bag.. #McDStories

7:03 AM - 24 Jan 12 via Twitter for iPad · Embed this Tweet

Fig 2: Hijacked tweet with #McDstories

Figure 1 presents one of the hashtagged tweet with #Android a famous brand in the market of Smartphone's being hijacked by posting a unwanted URL along with it.The URL shown in the tweet has been recognised as spam and thus is now blocked by twitter.

1.1 Case Study: #McDStories

We provide an insight into the most popular hashjacking tale of corporate social media #McDStories [2].Twitter campaign using the hashtag #McDStories was launched hoping that the hashtag would inspire heart-warming stories about Happy Meals. Instead, it attracted snarky tweets and McDonalds detractors who turned it into a #bashtag to share their #McDHorrorStories. Figure 2 presents one of the tweet which accounted as hijacked tweet with respect to the hashtag #McDStories.

The paper describes an algorithm based tool to detect and analyze hijacked #tweets.For detection perform partitioning of the dataset into: training and test sample.We have categorised our hashtags under various popular divisions such as : Technology, Entertainment, Politics, Popular Brands and Others.Using this categorisation we not only detect as well as analyse the percent of hijacked tweets per hashtag but,also project which particular category accounts for the maximum hijacking.Our tool thus trained for such multiple categories is capable of analysing hijacking of any given hashtag.

We achieve this functionality by implementing a statistical approach tf-idf short for term frequency -inverse document frequency .It is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.It is often used as a weighting

factor in information retrieval and text mining. [3].The details of the tools are discussed in the following sections of the paper.This pa-per is structured as follows.Section 2 describes the dataset used.Section 3 describes the methodology and the algorithm used for the tool.Section 4 presents the analysis and results over the dataset done by the tool.Further section 5 discusses the possible ways to avoid #hijacking in live-tweeting followed by the conclusion in Section 6.

2. DATA SET

Initially, we acquire a dataset and partition it into training and test data. Each partition consists of tweets collected for around 20 popular trending hashtags using publicly available twitter API which filters data from the twitter based on a keyword i.e. hashtag.As the twitter streaming API is rate limited ,there are around 1000 collected tweets for every hashatg which accounts to a total of 20,000 tweets from various hashtags over a period of 30 days.This served as the training data for the tool designed for detecting hijacking on hashtagged tweets.

For analysis, data collection is spread across two steps.

2.1 Categorical Hashtag Repository

Top trending hashtags are collected everyday over a period of 30 days.These hashtags are then identified into one of the categories:Technology, Entertainment, Politics and Popular Brands.The set of hashtags collecetd everyday are given to a group(20) of manual annotators.Each annotator is asked to place the given hashtag in a category.Every hashtag placed in a category by more than 12 annotators is assigned the selected category.One with less than a score of 12 were similary placed in Others category.Table I presents a snapshot of few categorized hashatgs.For both the partitions: test and training sets are collected for around 40 Hashtag names.

Table 1. #Hashtag names per Category

Category	#Hashtag Name
Technology	#Android,#Apple,#Smartphone, #ios,#dell
Entertainment	#CSKvsMI,#Filmfare, #MissWorld,#Maroon5,#Justin
Politics	#namo,#congress,#AAP,#BJP,#namobirthday
Brands	#puma,#adidas,#Samsung,#Lakme
Others	#happy,#Birthday,#Rain,#Sunny,#KillMe

2.2 Tweet-Hashtag Mapping

A keyword based search query is performed using twitter API having keyword as one of the hashtag names mentioned in Table 1.A total of 40,000 tweets were collected for training testing our tool.A preprocessing of complete data set is done to repudiate special characters and slang words if any.As social media brings in a diversity of people and so does the language in tweets,it is difficult to prune slang words from a large dataset.Therefore a slang word dictionary is maintained obtained from domain information to replace such slangs with their meanings.

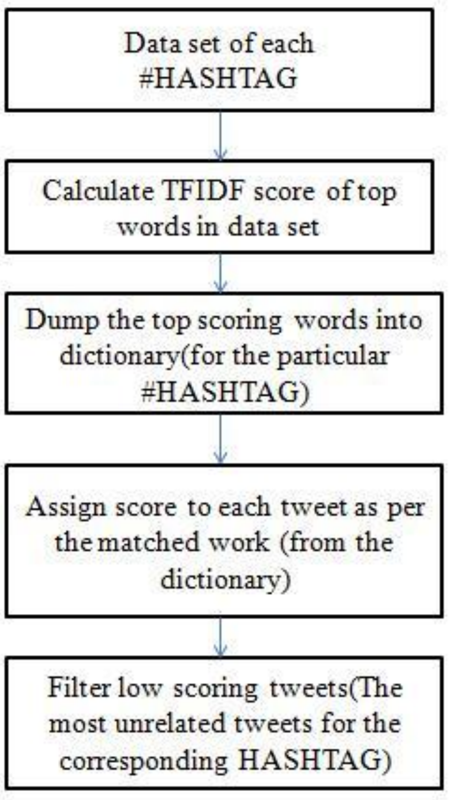

Data set of each #HASHTAG
Calculate TFIDF score of top words in data set
Dump the top scoring words into dictionary(for the particular #HASHTAG)
Assign score to each tweet as per the matched work (from the dictionary)
Filter low scoring tweets(The most unrelated tweets for the corresponding HASHTAG)

Fig 3: HashJacker-Flow diagram

3. HASHJACKER: METHODOLOGY

This section provides descriptive insight into tool developed to detect hijacking over a particular hashtag of a given category.The tool performs a dictionary based analysis for all the tweets of a particular hashtag.A tweet that contains the maximum matching words from the dictionary is considered to be the most relevant tweet.As it follows the general trend a hashtag is being talked for on the social media platform.A tweet that follows the trend or is synchronous with other similar tweets of a partiular hahstag can be thus considered as non hijacked tweet.The most ir-relevant tweet has the high probablity of being a malicious,spam containing or sarcastic one as it does not follow the talked trend for a given hashtag.The idea behind tweeting such text is misusing the hashtag for acquiring the user attention connected to that hashtag.Such tweets which has a single trending hashtag name attached to it along with all the junk a user wants to distribute. Figure 3 presents the flow diagram for the tool developed to detect hashtag hijacking.

3.1 Tf-idf Score Calculation

For each hashtag data,the tool shows the top occuring words by calculating the tf-idf score for all words in the set of 1000 tweets.The tf-idf value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general [4].Using high tf-idf score,obtain the most frequently occuring words as compared to others in a given set of data.

3.2 Per Category Dictionary Creation

Next, the tool creates a dictionary file of the top scoring words obtained from the previous step for every hashtag .This will give us multiple dictionaries for the first 20,000 tweets of given 20 hashtags falling in different categories.As there are hashtags belonging to different categories,therefore it is evident for the tool to maintain a common dictionary for every category considered which combines the top scoring words of all hashtags in a category.These top scoring words are the most relevant and most talked about words with respect to a particular hashtag of a category.As this dictionary is built over a particular category,it can be reused for upcoming hashtagged tweets belonging to the same category.

Table 2. Top scoring words per hashtag

Hashtag	Top scoring words
#Android	App,like,awesome,coins,great,launch,slow,lag
#CSKvsMI	Sachin,Win,Champion,Lost,captain,victory,won,beaten
#Elections	voter,voting,modi,namo,congress,bjp
#puma	discount,new,stock,black,white

3.3 Score Hashtagged Tweet using Dictionary

Further,a score is flagged to every hashtagged tweet based on the number of matches each tweet has with the words in the dictionary.For each match the score assigned is 1.So,we obtain a cumulative score for every tweet .This score presents the total matches a tweet has for all the words it contains with the dictionary itself.As twitter limits the number of characters in a tweet to be 140 which accounts for an average 30-35 words(considering letters per word),each tweet can get a maximum score of the same range.

3.4 Filter Low Scoring Tweets

Minimum score a tweet can get is 0 or 1. Such tweets are identified to be the most malicious or irrelevant or probably junk containing.The tool finally filters the low scoring tweets in a separate file from the hashtagged tweets of each hashtag.

4. ANALYSIS AND RESULTS

Tool is trained for tweet data from around 20 hashtags.This gives a dictionary for the corresponding categories of the considered hashtags.As the tool outputs the high scoring tweets for a given hashtag,it signifies the most relevant and trend following tweet with respect to hashtag. A low scoring tweet is considered to be a misleading one which uses hashtag name for just gaining any other user's attention who is following a hashtag conversation by his choice.

Lastly the tool calculates the percentage of these low scoring tweets over the total number of tweets present for a hashtag. We formulate a new metric as : Percentage of hijacked tweets for a particular hashtag = Number of low score tweets/Total number of tweets *100 %

Table 2 presents some of the low scoring and high scoring hashtagged tweets for a category.

After training,further testing of tool is performed on the second dataset partition.Interestingly,tool need not create another new dictionary for every incoming new hashtag.Assign it into any of the existing category and reuse the existing dictionary for that category.However it can add any non existing top occuring word from the new set of data to the existing category specific dictionary.Using this dictionary again flag every specific hashtagged tweet with a score.The hijack percentage for few hashtags tested by the tool with respect to the category it belongs to is presented in Figure 4 to 7. All the low score tweets were again manually analaysed by a group of 20 annotators.This included manually checking the URL's in these tweets,analysing the unwanted or unrelated data and abusive content distributed.

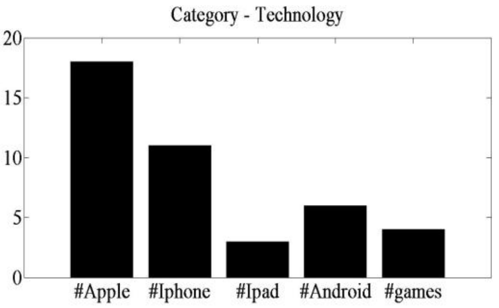


Fig 4: Percentage of Hijacked tweet for Category-technology

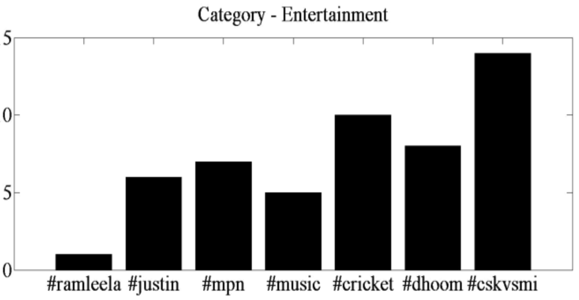


Fig 5: Percentage of Hijacked tweet for Category-politics

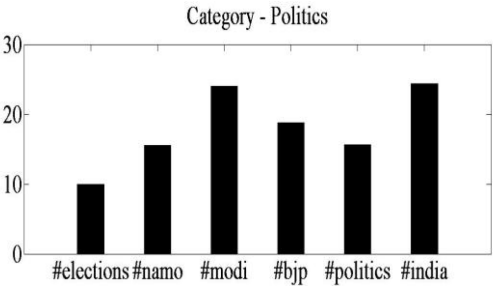


Fig 6: Percentage of Hijacked tweet for Category-Entertainment

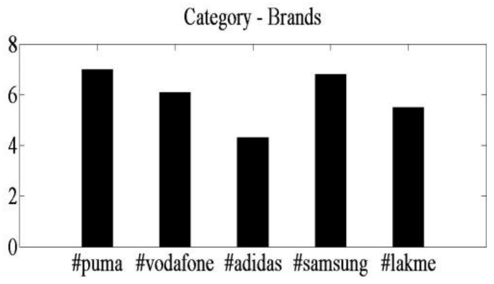


Fig 7: Percentage of Hijacked tweet for Category-brands

Following are the inferences obtained from the analysis by the tool as well as the annotators:

Table 3: #Hashtag names per Category

Category	tweet	Score
Entertainment	IPL: Final. 5.6: WICKET! DJ Bravo (15) is out, c Mitchell Johnson b Rishi Dhawan, 35/4 #PepsiIPL http://iplt2.com/match/76#cskvsmi	9
Entertainment	KEEP CALM AND TRUST BOOKIES pic.twitter.com/IJnmZbldnK #cskvsmi	0
Politics	BJP: Celebrate Narendra Modi's birthday as #GBD Global BJP Day. #ModiDiwas #NaMo #bjp	10
Politics	Watch RamLeela #NightofChampions http://t.co/9xDyu5JUgw	0
Technology .	WOW! Now #Apple fans can finally say they have a #GoogleEdition phone Welcome to the future guys.#iPhone5S #iOS7	10
Technology	#funny #******beactingup #comedy #siri #iphone #apple http://t.co/fllzMcrC3I	1

Politics and Entertainment show a higher hijacking percentage as compared to other categories.

As low score tweets were manually studied on twitter,We inferred that majority of these hijacked tweets came from a similar reccuring group of user accounts.These accounts are only created to tarnish or defame a particular brand image on social media.Many commercial website URL's are also distributed via attaching the most trending hashtag name to itself.A group of recurring accounts also existed in case of specific brand hashtags which specifically targetted to hijack a brand hashtag by posting abusive and sarcastic content on the social media.

5. AVOID HASHTAG HIJACKING: BEST PRACTICES

Hashtag creation is event driven. Majority of the users do not know about the best practices to be used while tweeting using a hashtag. Result is that various novice usually become a target for hijackers,spammers and anti brand campaigners. Some of the possible best practices which should be followed while hashtag tweeting:

- Hashtag design should align by the interest of the prospective audience as well as the platform[5].
- It should be engaging and interest provoking in relation to the brand it promotes.
- Informative and concise hashtags are less vulnerable to being hijacked[6].
- Pay close attention to the combination of words.Related and meaningful words should be used while tweeting with a hashtag[7]

6. CONCLUSION

Hashtags are about shared conversation.Today is the world of hashtag branding where the most talked or trending topic of interest itself become a brand.Many executives are using it as a marketing strategy while others use it to vilify their competitors. Hijacking a hashtag has become a serious concern for atleast people who invest and harness their brand value via social media.Apart from corporate,social media has always been a forum where novice users are being spammed and abused.Using a hashtag for a similar purpose is now a tool for many professional spammers to distribute junk and spam content in easiest and fastest way possible.With this tool we provide a feasible way to detect hijacking for any hashtag belonging to any of the popular category. This tool can be easily used to detect spam and junk distributed through popular trending hastags.We plan to devise a plug in which can be installed in various browsers as a detecter to mark such suspicious or irrelevant tweets. Similarly,Tracking these tweets back to users can help in identifying potential spammers and junk distributors who traget these popular hashtags in making their fake brand popular.Once detected various means as well as awareness can be used to curb this to an extent possible.

7. REFERENCES

[1] How to Use Hashtags on Twitter:A simple guide for Marketers,May2012,http://www.scoop.it/t/social-media-and-curation?page=4

[2] K. Hill;#McDStories: When A Hashtag Becomes A Bashtag ,Jan2012,http://www.forbes.com/sites/kashmirhill/2012/01/24/mcdstories-when-a-hashtag-becomes-a-bashtag/

[3] H. Wu and R. Luk and K. Wong and K. Kwok. "Interpreting TF-IDF term weights as making relevance decisions". ACM Transactions on Information Systems, 26 (3). 2008.

[4] Wikipedia, the free encyclopedia,http://en.wikipedia.org/wiki/Tf

[5] D. Berkowitz,Hashtag Marketing:9 ways to avert disaster,Feb 2014,http://mashable.com/2012/02/14/hashtag-marketing-disaster-tips/

[6] PR Agency Advice: How to Avoid Hashtag Hijaking,http://reimaginepr.com/pr-agency-advice-how-to-avoid-hashtag-hijaking/

[7] Social Media Best Practices:Selecting a Hashtag and Livetweet-ing,University of Michigan,2014