Hashtag the Tweets: Experimental Evaluation of Semantic Relatedness Measures

Muhammad Asif
Dept. of Computer Science and IT
The Islamia Univ. of Bahawalpur, Pakistan

Malik Muhammad Saad Missen
Dept. of Computer Science and IT
The Islamia Univ. of Bahawalpur, Pakistan

Nadeem Akhtar
Dept. of Computer Science and IT
The Islamia Univ. of Bahawalpur, Pakistan

Hina Asmat
Dept. of Computer Science
Govt. S.E. College, Bahawalpur

Mujtaba Husnain
Dept. of Computer Science and IT
The Islamia Univ. of Bahawalpur, Pakistan

Muhammad Asghar
Dept. of Computer Science
Govt. S.E. College, Bahawalpur

Abstract—**On Twitter, hashtags are used to summarize topics of the tweet content and to help search tweets. However, hashtags are created in a free style and thus heterogeneous, increasing difficulty of their usage. Therefore, it is important to evaluate that if they really represent the content they are attached with? In this work, we perform detailed experiments to find answer for this question. In addition to this, we compare different semantic relatedness measures to find this similarity between hashtags and tweets. Experiments are performed using ten different measures and *Adapted Lesk* is found to be the best.**

Keywords—component; formatting; style; styling; insert (key words)

I. INTRODUCTION

According to Danah M. Boyd & Nicole B. Ellison [1], a social network is a community, place or platform where people gather for similar cause or interest. Online social network users can interact with their social contacts as well as use number of services provided. For example, they can create profile, share posts, pictures, videos, status and use messaging [2, 3]. The huge numbers of online social network users share their feelings, thoughts, activities, knowledge, emotions and information on online network. Among popular online social networks, Facebook [4] and Twitter [5] are top. Facebook has leading position with registered user approximately 1 billion plus, 968 million active users per day, 844 million user's login on Facebook through mobile every day, and 4.75 billion posts are posted per day[1].

Twitter, another rapidly growing social networking site having registered users over 600 million, 316 million active users, 500 million tweets posted by users and 80% users use twitter's services on mobile phones [2]. The messages posted by a registered user on twitter are called tweets that can have maximum of 140 characters. User can post a lot of tweets at daily bases by using web interface, SMS or smart phones app. Unregistered users can only view and read tweets posted by different people [5]. The popularity of the twitter can be estimated by the fact that just after two years of its launching, company gained rapidly growth with 100 million tweets per quarter posted in 2008. By 2010, this amount reached to fifty million tweets per day [6] with company having 7000 registered applications [7]. This amount reached to phenomenal number of 600 million by year 2015[3].

A tweet is a short length message maximum 140 characters may consist of text, photos, web links and videos [15]. A Hashtag (like a label or metadata tag) is used by users on different social networking media and on other micro-blogging web sites that helps users to find messages, topics with a specific contents or theme. Users can create or use Hashtags by putting a "#" sign before a word or phrase without blanks and can be used in anywhere in tweet. Internationally, the hashtag became a practice of writing style for Twitter posts during the 2009–2010 Iranian election protests, as both English- and Persian-language hashtags became useful for Twitter users inside and outside Iran [36]. Beginning July 2, 2009, Twitter began to hyperlink all hashtags in tweets to Twitter search results for the hashtagged word. In 2010, Twitter introduced "Trending Topics" on the Twitter front page, displaying hashtags that are rapidly becoming popular[4]. Trending topics, the most discussed topics on Twitter at a given point in time, have been seen as an opportunity to generate traffic and revenue. Spammers post tweets containing typical words of a trending topic to attract clicks. This kind of spam can contribute to devalue real time search services [37].

---

[1] http://newsroom.fb.com/company-info/
[2] https://about.twitter.com/company

[3] http://www.internetlivestats.com/twitter-statistics/
[4] https://en.wikipedia.org/wiki/Hashtag#cite_note-13

Hashtags help searching relevant tweets from twitter. Spammers also use hashtags for their own purposes by attaching famous hashtags with spam information. In such cases tweet and their hashtags are both not related. This is the reason knowing relevancy of hashtags and tweets is not only important to search relevant information but also important to determine authenticity of the information. The purpose of this work is to find the relatedness of hashtags with tweets they are tagged to. There are many semantic relatedness measures proposed in the literature. The effectiveness of these measures for finding relatedness between hashtags and tweets is questionable. Hence, we plan to compare findings of these measures on our data collection.

To overcome such kind of spam tweets, there is a need to find if hashtags attached with a tweet are really relevant to it i.e. if the hashtags are describing actual content of tweet or an irrelevant hashtags has been attached to get some commercial benefits or because of some bad intentions. In this paper, we perform detailed experiments to find relatedness between a tweet and hashtags attached to it. We compare findings of several relatedness measures (already proposed in literature) for computing this relatedness between tweets and their hashtags. This task is challenged by many complex sub-tasks like segmentation of compound hashtags. For this particular sub-task, we use three methods and compare their results.

A. Research Questions

We focus on finding answers to following three research questions in this work:

- Do hashtags represent their tweets or not?
- Which is the best method for segregating compound hashtags?
- Which relatedness measure is good for estimating relatedness between tweets and hashtags?

I. CONTRIBUTIONS

- We develop a web application that helps to extract tweets on the bases of screen names (users) from twitter with the collaboration of twitter Rest API and save tweets into MySQL database.
- This application is capable to separate hashtags from tweets and save into selected database.
- This application uses three different methods to segment hashtags words like Regular expression, Google search engine and lexicon methods.
- We integrate a PHP library named PHP POS Tagging that helps us to get parts of speech from tweets like verbs, adverbs, nouns, adjectives and their sub types and save POS into database.
- This application uses web parsing technique to finds the similarity / semantics score b/w hashtag and tweets contents using different semantic relatedness / similarity algorithms like Adapted Extended Lesk, Lch, Jcn, Res etc. This application finds score using ten different algorithms.

- This application can help to find which types of POS are used in majority of tweets.
- This application can find mean (arithmetic) and also finds standard deviation from calculated result, from where we can predict about hashtags.

II. RELATED WORK

Hashtags have been used for several purposes in the work related to twitter. However, we discuss some important works in this section.

Piyush Bansal et al. [25] proposed a system that can analyze and segments the hashtags and return pages from Wikipedia for corresponding context of hashtag and its tweet text input to the system. Their proposed system has three components, segmentation seeder that generates a possible segmentation list by using variable length window technique. 2nd component deals with two tasks. First is the feature extraction from segmentation and 2nd is entity linking on segmentations. This component also finds the different scores like bigram score for accurate context matching, context score to find maximum contextual similarity with tweet's contents, capitalization score to find the information from hashtags written in capital words and relatedness score to find the relatedness between tweet context and hashtag segmentations. The 3rd component responsible for ranking of segmentation, produce a ranked list of segmentation along entity linking [25].

Ilknur Celik et al. [26] investigate the semantic relationships b/w entities from posts published by users on Twitter and develop a relation discovery framework that detects relations among entities of post. Their proposed system has two parts, first one is entity extraction and semantic enrichments, in which they detect entities and their semantic in reference of post, new or topic etc. the result in the form of graph that can be represented via triple G: = (R, E, Y), R & E are the resources and entities and   is a relation among resources and entities. The second step is relation discovery, in which detection of pairs of entities from graph that have a certain type of relationship in specific period of time. Relation b/w two entities e1, e2 can be defined via a tuple Rel (e1, e2, tstart, tend, w), in which tstart and tend specify the relation's validity and w is a weight of relation that's belongs to [0..1]. Highest score mean stronger relationship and lower score mean weaker relationship b/w entities.

Given the across the board of interpersonal organizations, research efforts to recover data utilizing tagging from informal communities correspondences have expanded. Specifically, in Twitter informal organization, hashtags are broadly used to define a common connection for occasions or subjects. While this is a typical practice frequently the hashtags openly presented by the user turn out to be effectively one-sided. Costa et al. [27] proposed to manage this inclination clustering so as to define semantic meta-hashtags comparable messages to enhance the classification. They utilized the client defined hashtags as the message class labels of Twitter and applied the meta-hashtag way to deal with help the execution of the message classification. The

meta-hashtag methodology is tried in a Twitter-based dataset built by asking for open tweets to the Twitter API. The test results yielded by looking at a baseline model taking into account user's defined hashtags with the grouped meta-hashtag methodology demonstrate that the general classification is moved forward. It is presumed that by joining semantics in the meta-hashtag model can have sway in different applications, e.g. proposal frameworks, occasion identification or crowdsourcing. Their proposed system has two models, first baseline model that deals with user defined hashtags. In 2nd model (meta-hashtag model), they defined meta hashtags. They get a dataset using Twitter API and use support vector machine method for classification, and they use van Rijsbergen measure for classification. For evaluation their method they use Support Vector Machine to discover ideal optimal hyperplane b/w positive and negative cases [27]. Another related work is done by Planck et al. [28] and is worth reading.

Another interesting work on twitter relates to a very important task i.e. sentiment analysis. Sentiment analysis is a fast and efficient way to get knowledge about user's interest, felling towards brands, business etc. and there are many methods training sentiment classifier for Twitter dataset. Saif et al. [29] present a novel approach of adding semantics as extra feature into the preparation set for sentiment analyses, For each concept (e.g. iPhone) from tweets, they adds their semantic concept (e.g. "Apple item") as an extra component, and measure the relationship of the agent idea with negative/positive opinion. They apply this method on Twitter datasets for prediction of sentiment analysis. They apply this approach on three datasets, Stanford Twitter Sentiment Corpus (STS), Health Care Reform (HCR) and Obama-McCain Debate (OMD). They use open source APIs to extract entities from online textual data, then they use three different methods to incorporate semantic feature into Naïve Bayes classifier training. After this they replace all entities with their corresponding semantic concepts. After this they use Unigram, Part of Speech and Sentiment topics methods for features extraction. Their proposed sentient topic method has better results as compared to Unigram and Part of speech methods, for sentiment analyses.

Kywe et al. [30] research on hashtags by analysis a Twitter dataset having more than 150,000 users. They proposed a method based on filtering, the proposed method considers both client inclinations and tweet content in selecting hashtags to be suggested. They use collaborative filtering approach that referred as a "user to user" U2U filtering approach to rating targets both items and users to assign an item to a target user, another approach "item to item" based on correlation to assign an item to target user. They also recommend another approach based on measurements of similarity b/w items by comparing their features [30].

In paper [31], they proposed a novel method that recommends hashtags for Tweets written in English Language. They use a skip-gram model for distributed word representation that uses a log-linear classifier to predict words in a range. Feed forwarded neural network (FFNNs) model is used for language modeling and natural languages tasks (NLP). They created a neural network that uses a non-linear function on each of its layers. They create a network having different computational unit ranges from 300 to 1000 and dimension of each input & output layer is fixed by 300. A component name feature vector generation in which, They get a 300-dimensional component vector for each word of the tweet, Hashtags are also included, after getting dimensions they perform an average operations on different tweets word feature vectors to make a solitary tweet feature vector that's used as an input in FFNN. Dimension of this tweet feature vector is same as the dimension of a word feature vector. By averaging tweet feature vector is close to the tweet in semantically concept. They used Batch Gradient Decent algorithm and a Mean Squared Error MSE as objective function.

We have observed that many researchers have used the potential of hashtags on twitter. However, we have hardly seen any attempt (to the best of our knowledge) that focuses on their similarity with the actual content of tweets i.e. how much the hashtags represent the original tweet content? In next section, we describe our experiments to find answer for this question.

III. EXPERIMENTS

A. Data Collection

We integrate Twitter Rest API into our web application for extraction tweets from twitter web site. It allows getting recent 200 tweets & retweets for each user in single request and we can make 180 requests in a quarter of hour. It means we can extract 36000 tweets per quarter and 144000 tweets per hour. The Twitter Rest APIs[5] also gives facility of extract tweets on the base of user (screen names). There are following steps to extracts tweets.

- A user account is required at twitter web site that's initially step towards tweets extraction, here we have already registered user of twitter,
- we create an application on twitter to get user authentications information that helps a user to authorized himself, these are consumer key, secret key, access token and access secret token. Any user gets this confidential information by registering and creating an application on twitter,
- OAuth, an open standard that authorized a user by using above mentioned confidential information,
- Next step is to write a code to make connection with twitter using Rest API, for this purpose we use PHP web programming language and adobe Dreamweaver CS6 as editor. This application is host on a temporary domain for extracting tweets.

---

[5] https://dev.twitter.com/rest/public

We extract around above 15000 tweets as a sample data set. We extracts the General Tweets posted in different domains, instead of specific domain or community i.e. Politicians or businessman. These are mixed tweets that contain Hashtags having English words, different terminologies, local language words, symbols and many more.

B. Data Processing and Storage

Once the data collection has been prepared, we do following steps:

- Remove duplicate tweets,
- Remove digits from hashtags,
- Remove tweets having words other than English language,
- Remove tweets with hashtags including abbreviations or acronyms,

After applying this processing, we are left with total 8001 tweets to work on. We use MySQL database management system for storage of these processed tweets. We store tweets, separated hashtags, segmented hashtags words. The database schema is shown in figure 1.
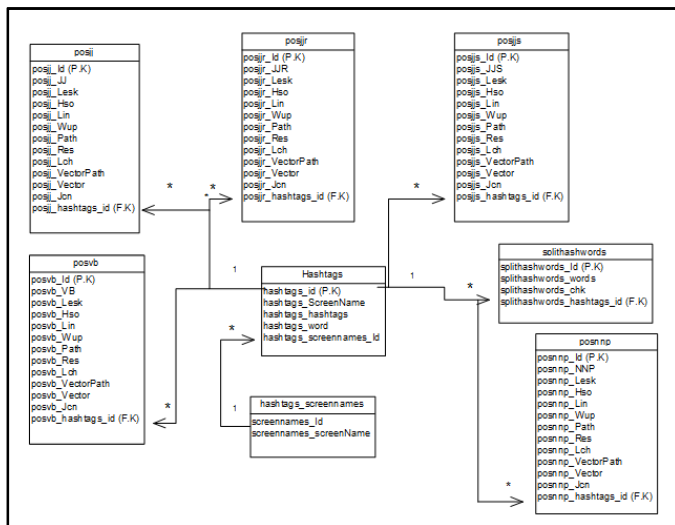

Figure 1: Database Schema

C. Segregating Compound Hashtags

As discussed earlier, the task of computing relatedness between tweets and their hashtags involves some sub-tasks. One of the sub-tasks is to extract individual words from a compound hashtag. For example, if we find a hashtag #savetheworld then we need to identify all three individual words present in this hashtag i.e. 'save', 'the', and 'world'. It is very important to identify correct words from hashtags because performance of major task i.e. finding relatedness between hashtags and tweets depends on this correct identification.

IV. SEGREGATING COMPUND HASHTAGS

This sub-task of extracting individual words from compound hashtags is performed in two steps. First step in this regards is to extract hashtags from tweets themselves. For this purpose, we write a simple code which separates hashtags from tweets with 100 percent accuracy. The second step is separation of compound Hashtags. A careful analysis of hashtags reveals that most of the people use capitalized and well defined words to make hashtags. However, some use hashtags containing words in lower case letters while many others combine lowercase and uppercase letters for this purpose. We categorized Hashtags into the followings:

- Uppercase words
- Lowercase words
- Capitalization words
- Mixed words

It is very easy to identify words in a hashtag written in capitalization form (i.e. #SaveTheWorld) while it becomes difficult for rest of the forms. Therefore, we decide to experiment with four different methods to identify individual words in hashtags and extract them. Each method is described in detail in following sub-sections.

A. Extracting Individual Words Using Regular Expression

This method based on regular expression that uses along coding and this is the most successful method because majority of people used capitalization case to make compound Hashtags. This method can segment hashtags that contains the following type of words

- English words
- English names
- English movies names
- Roman Urdu names
- Lowercase single words
- Uppercase words
- Capitalization words
- Others languages words

We use the pooling method for finding accuracy of this method by using a pool of 1000 hashtags. The accuracy of this method is $911/1000 = 91.1\%$.

B. Extracting Individual Words Using Google Search Engine

In this we method, we use Google search engine to separate Hashtags. We use this method based on query technique, in which we make a unique query for each Hashtags to Google search engine. This method can segment hashtags that contains the following type of words

- English words
- English names

- English movies names
- Roman Urdu names
- Lowercase single words
- Uppercase words
- Capitalization words
- Others languages words

We find accuracy of this method by testing and analyzing the sample of first 1000 hashtag that is 925 / 1000 = 92.5%.

C. Extracting Individual Words Using combination of Regular Expression &Google Search Engine

It is a hybrid method which combines regular expression method as well as Google search engine method. First we use Regular Expression. While segmenting hashtags it checks every segmented word from lexicon WordNet database that's we already stored in our local schema, if match found that word is stored in another "splithashwords" table with flag value 0, if match not found then it also store relevant word into same table but with a flag value 2. After finishing compilation of regular expression, second method Google Search Engine start working, in which it select all words from table by querying and try to segment words with flag value 2. Successful segmented words again save into same table with flag value 0 and compound word with flag value 2 is deleted at the last of searching completed otherwise not. Its accuracy rate is much higher as compared to first two methods. We find accuracy of this method by testing and analyzing the sample of first 1000 hashtag that is 978 / 1000 = 97.8%.

D. Extracting Individual Words Using Lexicon Method

This method is based on coding as well as lexicon database to segment Hashtags. It is effective in segmentation of compound hashtags written either in uppercase, capitalization, mixed and specially in lowercase letters but it is not easy to implement because it makes a lot of queries to database to find possible matches and processing, after finding possible matches then it makes permutations of all retrieved matches, after permutations it again find the accurate permutation that will takes a lot of time and involve a lot of processing, time and processing cost of this method is depend on hardware. It takes much time to segment a Hashtag consisting of two words like **#blindlycartouche,** it is consisting of two words blindly and cartouche. It successfully segment the above mentioned hashtag but take a lot of time. The compound hashtags may consists of different no. of words and each word has different length of characters. This method fails if hashtags contain words are not present in lexicon WordNet database. We are unable to implement this method due to a lot of time consumption, here's some calculations that describes how's this method is costly and not possible to implement. We choose random hashtags containing two to five words and each word having length from two to five characters. We assume that all hashtags belong to lexicon database and have length from two to four words, the cost of lexicon method cost shown below in

different tables (table 5 onwards). We find accuracy of this method by testing and analyzing the sample of first 1000 hashtag that is 512 / 1000 = 51.2%.

E. Extraction of Parts of Speech (POS) from Tweets

At this step we extract part of speech (POS) from Tweets for the purpose of finding semantic relatedness between segmented hashtags and contents of tweets that consists of (nouns, adverbs, adjectives, verbs and their sub forms). To perform this task we used a PHP library named "PHP Tagger", which is used to extract POS from any given sentence. We use it for extraction of POS from tweets by passing tweets to Tagger function. We extract nouns, verbs, adverbs, adjectives and respective sub forms. We extract POS and save into MySQL database. Following types of POS tags are considered for match with hashtags.

Table 1: Part of Speech

Sr. #	POS	Type
1	NN	Noun, singular or mass base form
2	NNS	Noun, plural
3	NNP	Proper noun, singular
4	NNPS	Proper noun, plural
5	RB	Adverb base form
6	RBR	Adverb, comparative
7	RBS	Adverb, superlative
8	JJ	Adjective base form
9	JJR	Adjective, comparative
10	JJS	Adjective, superlative
11	VB	Verb, base form
12	VBD	Verb, past tense
13	VBG	Verb, gerund or present participle
14	VBN	Verb, past participle
15	VBP	Verb, non-3rd person singular present
16	VBZ	Verb, 3rd person singular present

Following table describes the processed data collection statistics.

Table 2: Data Size after Processing

Screen Names	196
Hashtags	8001

Segmented Words (hashtags)	15019
Adjectives	7121
Nouns	61972
Verbs	14489

Sr. #	Algorithm	Results
1	Adapted Extended Lesk	0.701
2	Hirst & St-Onge (HSO)	0.338
3	Lin	0.089
4	Wu & Palmer (wup)	0.232
5	Path Length	0.085
6	Resnik (Res)	0.392
7	Leacock & Chodorow (Lch)	0.456
8	Gloss Vectors (pair-wise)	0.038
9	Gloss Vectors	0.117
10	Jiang & Conrath (Jcn)	0.324

V. MATCHING TWEETS AND THEIR HASHTAGS

Now our raw material is ready to calculate similarity / relatedness among segmented Hashtags and POS of tweets. For this purpose we use a web interface based on WordNet similarity package that is powered by Ted Pedersen and Jason Michelizzi. This web interface provides a way to find semantic relatedness / similarity score b/w two words. This web portal calculates semantic score by using ten different algorithms. We find the score of all segmented Hashtags and their relevant POS by using all algorithms and save it into database for further calculation / use. The list of algorithm that calculate semantic relatedness is given below

- I.   Adapted Extended Lesk
- II.   Hirst & St-Onge (HSO)
- III. Lin
- IV.   Wu & Palmer (wup)
- V.   Path Length
- VI.   Resnik (Res)
- VII. Leacock &Chodorow (Lch)
- VIII. Gloss Vectors (pair-wise)
- IX.   Gloss Vectors (Vector)
- X.   Jiang & Conrath (Jcn)

A. Normalization

Algorithms used to match tweets with their hashtags use different scoring mechanisms and hence results computed by using these algorithms cannot be compared unless brought on same scale. We use following formula for normalizing results:

Normalized values $W_i = X_i – minimum(X) / (maximum(X) – minimum(X))$

Here,

- Wi is the normalized value.
- Xi is a value belongs to variable X.
- Minimum(X) is the smallest value from variable X.
- Maximum(X) is the extreme value from variable X.

As a result of normalization, results of all algorithms are brought on 0 and 1 scale and hence are comparable. These normalized values are then averaged (arithmetic mean) for each algorithm.

B. Final Results

After normalization of semantic relatedness scores and averaging them, the final results for different algorithms are given below:

Table 3: Algorithm wise results

VI. GROUND TRUTH PREPARATIONS

One of the objectives of this work is to evaluate and compare different semantic relatedness measures. For this purpose, we need to have gold standard so that results of these relatedness measures can be compared with it. The objective is to find some overall coherence between results of automatic matching using algorithms and manual match. It will confirm the findings of matching algorithms. For this purpose we choose four educated persons from different fields of life having good experience of using social networking sites and very well know about twitter and tweets. They evaluate our dataset (Tweets & Hashtags) carefully and answered in from of two options yes or no for each tweet and hashtag to tell the relevancy b/w hashtags and tweets. They evaluate 11000 plus tweets and answered with very responsibility. To find the agreements ration between annotators, we decide to use Fleiss' kappa. Fleiss' kappa (named after Joseph L. Fleiss) is a statistical measure for assessing the reliability of agreement between a fixed numbers of raters when assigning categorical ratings to a number of items or classifying items. This contrasts with other kappa such as Cohen's kappa, which only work when assessing the agreement between two raters. The measure calculates the degree of agreement in classification over that which would be expected by chance. There is no generally agreed-upon measure of significance, although guidelines have been given. Fleiss' kappa can be used only with binary or nominal-scale ratings. No version is available for ordered-categorical ratings [35]. At next step we use Fleiss' kappa to find the agreement among four evaluations. The results are given below in table 4.

Table 4: Fleiss Kappa Results

FILENAME	Tweets.csv
File size	89616 bytes
n-coders	4
n-cases	9956
n-decisions	39824

Average pair-wise percent agreement	78.22%
FK observed agreement	0.78

Another major reason why we need to evaluate our data by users is that numerous users don't follow the netiquette. For example, they tag the irrelevant word with Hash (#) which creates the problem to specify hashtag for any Generic Field. Due to lack of standards for inputting hashtag peoples add any kind of hashtags to represent tweets. Even English people (our concerned data set) create hashtags that have no meanings in any sense but they are written followed by # sign i.e.

- #TümTwitterBuTaglaMutlu
- #aÅŸk
- #Hamburg!!!ðŸ™ˆðŸ˜ ðŸ˜ŽðŸ‡©ðŸ‡ªâ ¤ï
- #politicalprisoners
- #ARR
- #HBDPM
- #SOUTHCOM

Above mentioned hashtags and many more are written by users upon their desires, they are not following any standard or rules. Peoples use special characters, pure in upper or lower case and mixed case, digits, roman words etc. Our web application cannot fulfill judgment on all types of hashtags, so we decide to perform user evaluation.

VII. CONCLUSIONS AND FUTURE WORK

On the basis of our experimental results and user evaluation, we conclude that hashtags mostly represent the tweets they are attached with. As far as correctness of semantic algorithms is concerned, our findings are given below:

- Adapted Extended Lesk is found to be the best at finding the relatedness between hashtags and tweets, the measurements of this algorithm is **0.701,**
- Lch measurement is **0.456**
- Resnik (Res) measurement is **0.392**
- Hirst & St-Onge (HSO) measurement is **0.338**
- Jiang & Conrath (Jcn)measurement is **0.324**
- Wu & Palmer (wup) measurement is **0.232**

We keep following points as part of our future work:

- Integrate other twitter APIs specially search API to get a large no. of screen names as well as tweets.
- Write a code to search and save screen names into database on the bases of any given keywords
- Write a code to deal automatically with data rate limit restriction by twitter to save time and get better results.
- Upgrade web application to deal with every kind of compound hashtags written in lowercase and uppercase.

- Upgrade web application to deal with roman hashtags as well as hashtags written in other than English language.
- Improve lexicon method to get better results in every aspect like speed, accuracy and time.
- Improve regular expression method that also deals hashtags other than written in capitalization.
- Improve Google Search Engine Method to finds context / meaning of abbreviation and acronyms.
- We want to restructure our database schema for better performance.
- Improve our segmentation methods of hashtags and our web application to handle
  - Hashtags including abbreviations, acronyms.
  - Hashtags having words or symbols other than English language.
  - Hashtags having pure lowers case i.e. not using capitalization.
  - Hashtags having mixed abbreviations, acronyms & English Words.
  - Hashtags having mixed English words and Roman Urdu words.
  - Hashtags with roman Urdu (Pakistanis & Indians writes their Urdu & Hindi in English)

REFERENCES

[1] Boyd, danah; Ellison, Nicole (2008). "Social Network Sites: Definition, History, and Scholarship". Journal of Computer-Mediated Communication 13: 210–230.doi:10.1111/j.1083-6101.2007.00393.x

[2] Journal of Computer-Mediated Communication Volume 13, Issue 1, pages 210–230, October 2007

[3] https://en.wikipedia.org/wiki/Social_networking_service

[4] http://newsroom.fb.com/company-info/

[5] https://about.twitter.com/company

[6] Beaumont, Claudine (February 23, 2010). "Twitter Users Send 50 Million Tweets Per Day – Almost 600 Tweets Are Sent Every Second Through the Microblogging Site, According to Its Own Metrics". The Daily Telegraph (London). Retrieved February 7,2011

[7] Staff writer (March 4, 2010). "Twitter Registers 1,500 Per Cent Growth in Users". New Statesman. Retrieved February 7, 2011.

[8] Garrett, Sean (June 18, 2010). "Big Goals, Big Game, Big Records". Twitter Blog (blogof Twitter). Retrieved February 7, 2011.

[9] "Twitter Blog: #numbers". Blog.twitter.com. March 14, 2011. Retrieved January 20,2012.

[10] Kazeniac, Andy (February 9, 2009). "Social Networks: Facebook Takes Over Top Spot, Twitter Climbs". Compete Pulse (blog of compete.com). Retrieved February 7, 2011.

[11] Weil, Kevin. (VP of Product for Revenue and former big data engineer, Twitter Inc.). "Measuring Tweets." Twitter Official Blog. February 22, 2010.

[12] Krikorian, Raffi. (VP, Platform Engineering, Twitter Inc.). "New Tweets per second record, and how!" Twitter Official Blog. August 16, 2013.

[13] Official Blog. June 30, 2011.

[14] "Twitter turns six." Twitter Official Blog. March 21, 2012.social-media-from-trend-to-obsession-infographic/142323

[15] https://support.twitter.com/articles/13920

[16] Using Twitter with Your Phone". Twitter Support. Retrieved June 1, 2010. We currently support 2-way (sending and receiving) Twitter SMS via short codes and one-way (sending only) via long codes.

[17] Logical Argument on CrystalTower, showing that accessing Twitter through SMS may incur phone service provider fees

[18] https://en.wikipedia.org/wiki/Hashtag

[19] Michael Schneider (Apr 21, 2011). "New to Your TV Screen: Twitter Hashtags". TV Guide.

[20] Todd Wasserman (Dec 3, 2012). "McDonald's Releases First TV Ad With Twitter Hashtag

[21] Heather, Kelly (12 February 2013). "Twitter and Amex let you pay with a hashtag".CNN. Retrieved 2013-11-25.

[22] Akwagyiram, Alexis (17 May 2012). "Are Twitter and Facebook changing the way we complain?". BBC News. Retrieved 2012-06-12.

[23] Maynard (2014). "Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis". Proceedings of the Conference on Language Resources and Evaluation.

[24] Anna MehlerPaperny (Apr 13, 2011). "Jack Layton's debatable 'hashtag' #fail". The Globe and Mail.

[25] Towards Deep Semantic Analysis of Hashtags by Piyush Bansal, RomilBansal and Vasudeva Varma International Institute of Information Technology Hyderabad, Telangana, India (http://link.springer.com/chapter/10.1007%2F978-3-319-16354-3_50)

[26] Learning Semantic Relationships between Entities in Twitter by IlknurCelik, Fabian Abel, and Geert-Jan Houben (http://dl.acm.org/citation.cfm?id=2027790)

[27] Joana Costa, Catarina Silva, MárioAntunes and Bernardete Ribeiro, Defining Semantic Meta-hashtags for Twitter Classification (http://link.springer.com/chapter/10.1007%2F978-3-642-37213-1_24)

[28] Max Planck, Isis LymanPollard,Charles Brock, Alex George , Initial Indicators of Topic Success in Twitter Using topology entropy to predict the success of twitter hashtags (http://www.computer.org/csdl/proceedings/nsw/2013/0436/00/06609214.pdf)

[29] Hassan Saif, Yulan He, and Harith Alani, Semantic Sentiment Analysis of Twitter by (http://iswc2012.semanticweb.org/sites/default/files/76490497.pdf)

[30] Su Mon Kywe, Tuan-Anh Hoang, Ee-Peng Lim and Feida Zhu, On Recommending Hashtags in Twitter Networks (http://link.springer.com/chapter/10.1007%2F978-3-642-35386-4_25#page-1)

[31] Tomar, Aparna, et al. "Towards Twitter hashtag recommendation using distributed word representations and a deep feed forward neural network."*Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on*. IEEE, 2014.

[32] https://en.wikipedia.org/wiki/Data_collection

[33] https://en.wikipedia.org/wiki/Arithmetic_mean

[34] Law, E., Roto, V., Hassenzahl, M., Vermeeren, A., Kort, J.: Understanding, Scoping and Defining User Experience: A Survey Approach. In Proceedings of Human Factors in Computing Systems conference, CHI'09. 4–9 April 2009, Boston, MA, USA (2009)

[35] https://en.wikipedia.org/wiki/Fleiss%27_kappa

[36] "The story of the hashtag began with Iranians". Deutsche Welle Persian. 2009. Retrieved March 12, 2015

[37] Benevenuto, Fabricio, et al. "Detecting spammers on twitter." Collaboration, electronic messaging, anti-abuse and spam conference (CEAS). Vol. 6. 2010

Table 5: Cost using two words in Hashtags

Sr. #	Words	Each word's Length	Total Queries	Possible matches	Total permutations	Time in Milliseconds	Successful / unsuccessful
1	BkBw	2	147478	6	21	92.384	Yes
2	DocDoe	3	147479	8	36	91.738	Yes
3	HereHero	4	147480	10	55	96.476	Yes
4	AbaseWidth	5	147481	22	253	94.584	Yes
Total	4	14	589912	46	365	375.182	
Average	1	3.5	147478	11.5	91.25	93.795	

Table 6: Cost using three words in hashtags

Sr. #	Words	Each word's Length	Total Queries	Possible matches	Total permutations	Time in Milliseconds	Successful / unsuccessful
1	BkBwKm	2	147478	8	128	90.499	Yes
2	DocDoeDie	3	147479	13	468	88.253	Yes
3	HereHeroBraw	4	147480	20	1760	87.773	Yes
4	AbaseWidthStrap	5	147481	29	4524	88.226	Yes
Total	4	14	589912	70	6950	354.751	
Average	1	3.5	147478	14	1737.5	88.687	

Table 7: Cost using four words in hashtags

Sr. #	Words	Each word's Length	Total Queries	Possible matches	Total permutations	Time in Milliseconds	Successful / unsuccessful
1	BkBwKmZu	2	147478	11	858	87.901	Yes

2	DocDoeDiePud	3	147479	17	4199	89.533	Yes
3	HereHeroBrawGlee	4	147480	26	21203	90.912	Yes
4	AbaseWidthStrapWorth	5	147481	35	67340	93.754	Yes
Total	4	14	589912	89	93600	362.1	
Average	1	3.5	147478	22.25	23400	90.525	

Table 8: Average Cost of lexicon method

Sr. #	Words in Hashtags	Total Queries	Possible matches	Total permutations	Time in Milliseconds
1	Two	147478	11.5	91.25	93.795
2	Three	147479	14	1737.5	88.687
3	Four	147480	22.25	23400	90.525
Total	3	589912	47.75	25228.75	273.007
Average	1	147478	15.916	8409.583	91.0

Table 9: Cost of overall dataset using lexicon method

Sr. #	Words in Hashtags	Total Queries	Possible matches	Total permutations	Time in Milliseconds
1	1	147478	15.916	8409.583	91.0
	8001	8001*147478	8001*15.916	8001*8409.583	8001*91
	Average Total Cost	**1179971478**	**127343.916**	**67285073.583**	**728091**