International Journal of Advanced Research in Engineering and Technology (IJARET)
Volume 8, Issue 6, November - December 2017, pp. 15–28, Article ID: IJARET_08_06_003
Available online at http://www.iaeme.com/IJARET/issues.asp?JType=IJARET&VType=8&IType=6
ISSN Print: 0976-6480 and ISSN Online: 0976-6499

TREND WEIGHT – TRENDINESS DISTANCE - A NOVEL APPROACH TO IDENTIFY TRENDS IN MICROBLOGS

Pradyumansinh U. Jadeja
Computer Engineering Department, Nirma University, Ahmedabad, India.

Dr. Ketan Kotecha
Parul University, Waghodia, Vadodara, India.

ABSTRACT

Social media applications are widely used by billions of users who create tons of data. During last decade, this medium has established itself as a strong medium of communication which represents the voice of society. People tend to post almost everything like small event or happening of life through web-based networking media, especially microblogs are famous among these. This phenomenon generates noticeable important data with rich content of ideas, information, habits of mass, happenings and many more. Microblog is a reliable "meter" to identify the voice & flow of society as it reflects the Society but when it comes to analysis, it becomes much more difficult to identify the exact trend from discussions and also time-consuming for anyone as the nature of microblogs like quantum of data available and generated, data not having a well-defined model or structure & limited number of words to express the idea. In this paper, we have proposed a novel approach "Term Weight & Trendiness Distance Implementation to Identify Trendy Terms from Microblogs"

Key words: Trendy Terms, Trendiness Distance, Trend Identification, Microblogs.

Cite this Article: Pradyumansinh U. Jadeja and Dr. Ketan Kotecha, Trend Weight – Trendiness Distance - A Novel Approach to Identify Trends in Microblogs. *International Journal of Advanced Research in Engineering and Technology,* 8(6), 2017, pp 15–28.
http://www.iaeme.com/IJARET/issues.asp?JType=IJARET&VType=8&IType=6

1. INTRODUCTION

In the contemporary era, Society and Media are indispensable to each other and hence Social Media's tremendous popularity attracts the special scientific research & analytical study. From Primitive age to Palaeolithic age to Bronze age to current homo sapien sapien age, human civilization has inculcated different but sound methods of communication. We have never afforded to ignore the importance of "Effective Communication" especially when it comes to advancement.

At present (especially during last half of century), communication between any two persons residing at distant places has become so involving and easy that whole new "Social Media" technology and thus generation has taken birth. Giants like Google, Facebook, Tweeter and many more would NOT have been into EXISTENCE without this transformation. Major stack holders of Society have developed irresistible "trend" to communicate with each other through *"social media posts"*. They share their regular ideas, thoughts, likings, disliking, experiences, views and much more through this *"tunnel"*. Social media is regularly used by all to update for regular work done, to share solutions for problems faced, to inform others about sales and purchase of something, to inform about tours enjoyed, to share happy/sad moments, to express emotions, to update about current status. It would not be hyperbole to say that it looks each and every action of people is getting posted on Social media. *In nutshell*, people tend to post almost every small events / happenings of life on social media. To stay "updated" has become "obvious" in society. This phenomenon generates very important data with rich content of ideas, information, habits of mass, happenings and many more. Applying *"analytical eye"*, Social media do have a different kind of data like *information, emotions, actions, criticisms, happiness, praises, honours and sadness* of society. Social media is a true *"meter" to* identify the voice & flow of society as it reflects the *Society*. So, if we wish to know about society, we can't avoid *the noticeable importance of data generated through social media.*

Social media applications, for say, Facebook, LinkedIn, Telegram, Twitter are used widely by billions of users who create tons of data in the form of text, image, multimedia & animations. Among all these tools microblog is a kind of tool which allows one to share within specified-limited number of texts requiring user to write precise and up to-the-mark messages which reflects real sense of author/person. Moreover, it is very easy to write, manage & read from microblogs than other kind of Social media applications. Most celebrities and entities like actors, politicians, authors, institutions, organizations, Government bodies do have their *official* microblog accounts in applications like Twitter. They officially use Twitter account to convey messages in the form of any announcement, comment on hot discussion topics, share idea / words with followers from their authenticated account. These all stuffs are read, followed, promoted, opposed by millions of followers and these followers add their own contributions and eventually it becomes a *strong medium* which represents the voice of society.

The issue with this data is quantum of data available (produced) & also this data does not have a well-defined model or structure, so by nature it is fully unstructured. Anyone can produce data as there is no restriction on who produces data, no flow of writing as one is free to write anything, anytime in any manner. One needs to express within specified amount of characters limiting the data size. There is no dependency of data; tweets are, by nature, fully *independent*. We all have heard the famous English proverb *"Necessity is the mother of Invention"*. This proverb proves to be true from research point of view as when it comes to analysis, it becomes much-much difficult and time consuming for anyone to get exact voice from *huge bunch* of independent tweets (Tweet is a post of microblog application Tweeter) and this *logically* leads to the area of research – What is trending in Society by analyzing data of microblogs.

There is no trace of doubt about its usage as anyone can get *satellite-cum-helicopter-cum-street view* about, "What people are thinking?" "what people are discussing?", "what is happening in society?". Answer or hint to above questions helps to identify problems of society, hot topics in discussion, honest feedbacks of any product or service, useful for political parties to *'catch'* the taste of society at the crucial time of election. This is what we

can say "Trend in microblogs"**.** It is a herculean task to identify *trends* specifically from microblogs as in microblogs, one is given limited number of characters to express ideas and emotions. Anything may become trend if same kind of emotions, words written by large number of people residing in same or different geographical region within stipulated time period. If this happens, it is possible to identify hot topics, hot discussions carried out by society which can be local or global discussion.

In this paper, a novel approach *"Term Weight & Trendiness Distance Implementation to Identify Trendy Terms from Microblogs"* is proposed. The rest of the paper is organized as follows. In section 2 we summarized related work done work with unstructured data produced by microblogs, like anomaly detection, identify consumer behaviour, discovering patterns in transactional attribute-value data and many more. We present different adapted approaches for keyword identification and discussed "Trendiness Distance" in depth in Section 3. Section 4 describes Flow Diagram, Architecture, Proposed Algorithms, Data Set, Experimental Setup and Results. Section 5 concluded work done.

2. RELATED WORK

Many techniques derived from several disciplines such as data mining, text mining, natural language processing, machine learning, statistics & information theory have been adopted to work on unstructured data produced by microblogs. Different approaches are also involved like classification, clustering, statistical & information theoretic approach.

Alexandra et al. [1] presented an overview of trend & anomaly detection on unstructured data. The authors present detailed analysis of nature of network data produced. Davil Alfred [3] has presented method to identify trend through semantic filtering. He begun by keyword based filtering on topics, filtering related terms, apply fisher classification to identify consumer behaviour. Based on analysis of different research papers, Clude et al. [8] have developed different definitions on social media analysis, they have also given detail on different challenges of social media. Hila et al. [4] explored different approaches to analyse stream of Twitter messages to distinguish real world event and non-event messages using aggregate statistics & topically similar messages clustering approach. Su Gon Cho and Seoung Bum Kim [6] have worked to identify the pattern of keywords frequently appears on different research papers using low dimensional embedding methods, clustering analysis and association rule. Ceren Budak et al. [2] have worked on Online Social Networks and prepared a research on Structural Trend Analysis. Two ideas have been put forward by them: The first is coordinated trend and the other is uncoordinated trend. This study identifies various topics which are communicated among concentrated and distributed users utilising friendship information. Their carried out results imply that structural trends notably varies from traditional trends and convey new ways on how masses share information on social media. Luis González & Iván Pino [5] have discussed different issues faced by communication professionals that is the growing difficulty for managing corporate reputation crises. According that discussion the social media are mainly liable for this change, given that they have exponentially increased the visibility of an organization's risk, while amplifying the extent of any incident, and can thus make a corporate reputation that it has taken years to build wobble in a matter of minutes. They have discussed all these stuffs in detail with reference to microblog application Twitter. Gediminas Adomavicius and Jesse Bockstedt [7] have presented C-TREND, Cluster-based Temporal Representation of Event Data, a method for discovering temporal patterns in transactional attribute-value data. They have used calculation of a dendrogram solution using agglomerative hierarchical clustering.

3. PROPOSED WORK

3.1. Approaches of Keyword Identification

While calculating the weight of any term in Document (Tweet-Bin – Collection of Tweets) one can use any authenticate methodology like TF (Term Frequency), TF-IDF (Term Frequency Inverse Document Frequency), and Frequency of Term in entire corps, number of Tweets which contains given Term in Tweet-Bin or combination of all or any of these approaches.

If one is interested to find one's Trend from microblogs, he must have set of keywords for different time duration. Keywords are important words in the corpus. It is very important analysis of microblogs to decide "what a post (tweet) is about?" Can we come to conclusion by looking at different terms that makes that post / tweet? One measure of how important a term in tweet is its term frequency. One can argue that if term appears number of times in corps it should be considered as important term of that corps, means term gains importance proportionally to the frequency of that term in corps.

TF: Term Frequency, which measures of how frequently a term occurs in a document. Term Frequency $\boldsymbol{tf_{t,d}}$ of term $\boldsymbol{t}$ in document $\boldsymbol{d}$ is defined as the number of times that $\boldsymbol{t}$ occurs in $\boldsymbol{d}$ [9][10].

Normalized Term Frequency: The character limit of Tweet is fixed. It is logical to notice that the end users tend to post different size of tweets not necessarily occupying the maximum size and this results into various size of Tweet-Bins (documents). While comparing, chances of occurrence of a word/term in longer document would surpass chances of occurrence of the same word/term in shorter document. This disparity would lead a researcher to divide term frequency by number of words in document. This process is formally termed as 'normalization.' The count is normalized with a view to overcome bias towards longer documents [9].

Following formula can be taken into consideration to show the weightage of the term $\boldsymbol{ti}$ within the document $\boldsymbol{d_j}$.

$$\text{tf}_{\text{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

where $n_{i,j}$ is the number of occurrences of the considered term (ti) in Tweet-Bin dj, and the denominator is the sum of number of occurrences of all terms in Tweet-Bin dj, that is, the size of the Tweet-Bin|dj|.

Inverse Document Frequency: But many terms may be found in tweet that appears many time but not having that much of importance; e.g in English words like 'of', 'is', 'the', 'in' and many more, in Hindi के, का, एक, में, की, है and many more. Technique for adding these terms to a list off stop words and removing these stop words before analysis of Tweets may be adapted, but it is possible that few of these terms might be important in some tweets than other terms.

To consider only list of stop words is not a practical approach. Essence or intrinsic nature of another approach must be added that is term's inverse document frequency (idf) [10], which reduces the importance of commonly used words and emphasize words that are not used very much in a corps.

$$\text{idf}_{\text{i}} = \log \frac{|D|}{\{j : t_i \in d_j\}}$$

Where D is total number of Tweet-Bins in corps, $\{j: t_i \in d_j\}$ is number of Tweet-Bins where the term t_i appear

Tf-idf, The tf-idf weight of a term represents the multiplication of its tf value and its idf value [10].

Also number of Tweets that contains particular term in Tweet-Bin can be encompassed as one of the weight assigning methodologies. Here in algorithm, combination of variations of above approaches have been utilized to prepare a set of keywords (Trendy terms) of Tweet-Bin and to deriver Trendiness distance of Terms.

3.2. Trendiness Distance

Major algorithms / implementations consider terms having high frequency of occurrence in content, but other terms are also contributing to generate Trend. This contribution should not be avoided. One should have some measure that how far other terms from highest occurrence term (Trendy Terms) in content which can be defined as *"Trendiness Distance"*. Trendiness Distance of terms is the distance from Trendy Term in same bin. Trendy term can be considered as Terms with high frequency in all bins, average frequency of all bins, TF-IDF weight of term, count of number of tweets contains term, or combination of all above mention methods. So, in each bin for all terms one can consider highest weight as "super weight" of that bin and that term can be considered as "Super Trendy Term" of that bin. Difference of super weight of bin and individual term weight is the trendiness distance of that term in given bin. Trendiness distance plays significant role in deciding whether to consider that term in set of trendy term in given date-range or time. By analysing results, one can avoid few terms from dataset that do not contribute to generate trend. This will help to reduce data without affecting measurable results.

For any term, one can easily plot *trendiness distance* in each time-interval that gives depth view of behaviour of the term in different time-intervals. Terms with less distance from Super Trendy term can be treated as more likely to be trendy; those terms do have more confidence weight of trendiness than other terms. Variation of trendiness distance of a given term in different timespans gives idea about the behaviour of that term in time intervals, it gives idea about whether term is gaining trendiness (term is getting importance in discussion) or losing trendiness (term is losing importance in discussion) or remains constant (term gets same Constant importance during discussion) with respect to trendiness distance. This behaviour gives idea of ascent or decline of trend of the given term in time span.

Trendiness distance also gives idea about how many terms actually contribute to generate trend and which are those terms which one can omit without affecting actual results. So in post cycles, one can inform to algorithm what not to consider in calculation to reduce the dataset size. As in real life, dataset size always in tons, it would be much practical, labour-reducing and time-saving if one can curtail down dataset size by eliminating few entities without affecting output and hence ultimately needs less computation and storage power.

As a tool, custom coding (C#, SQL Server) has been used to prepare and work with dataset to get intermediate and final results.

4. EXPERIMENTS AND RESULTS

4.1. Flow Diagram

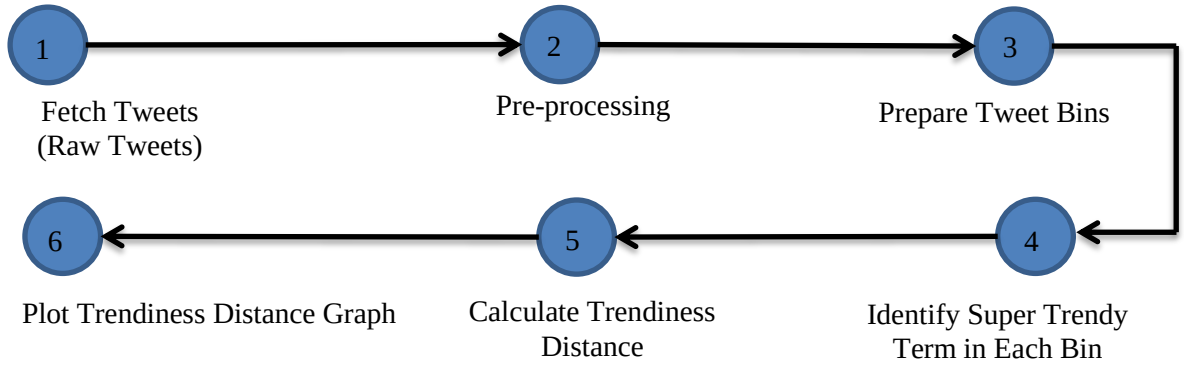

Figure 1 Calculate & Plot Trendiness distance (Flow Diagram)

4.1.1. Fetch Tweets

This step is preparation of Data Set, Collection of row tweets using Twitter API (Indian region Tweets of Specific date range)

4.1.2. Pre-Processing

- Cleaning data: Cleaning is the process to remove unnecessary stuffs clubbed with dataset hence to reduce computation power
- Tokenize tweets

4.1.3. Prepare Tweet Bins

Logically divide processed data into equal sized bins (one can divide data based on equal time interval or based on an equal number of tweets in each bin order by actual tweet time).

4.1.4. Identify Super Trendy Term in Each Tweet-Bin

Term with the highest weight in Bin is considered as Super Trendy Term of given Bin.

4.1.5. Calculate Trendiness Distance

Difference of super weight (Weight of Super Trendy Term of that bin) of bin and individual term weight is the trendiness distance of that term in given bin.

4.1.6. Plot Trendiness Distance Graph

For the given term, plot Trendiness distance in each bin.

4.2. Architecture

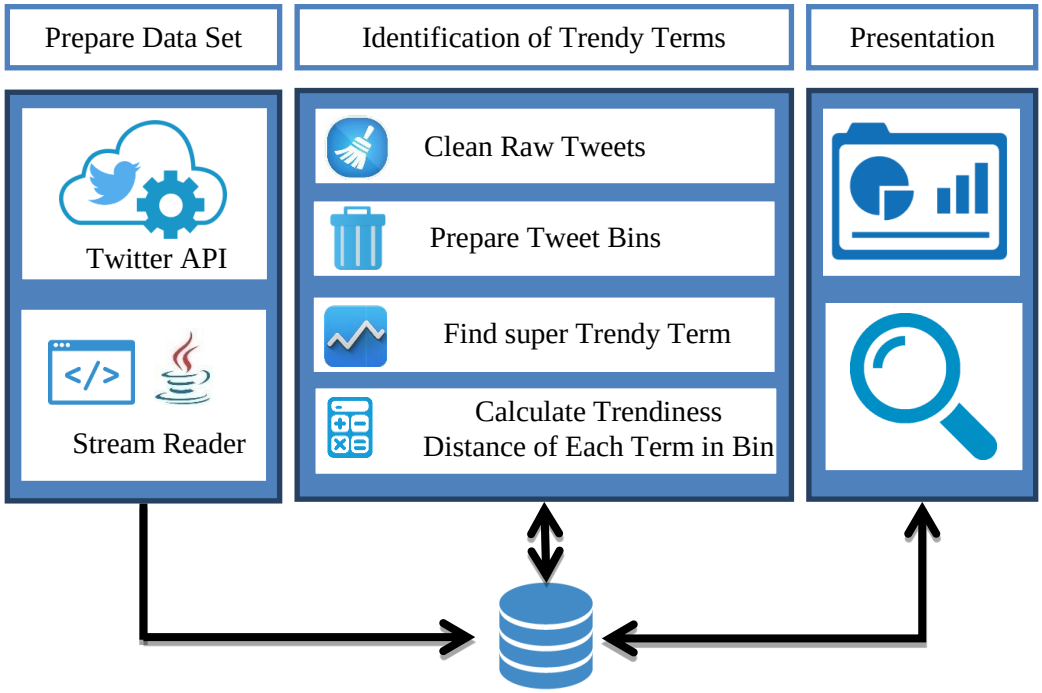

Figure 2 Identify Trendy Terms using Term Weight Trendiness Distance (Architecture)

4.3. Algorithm

```
PREPARE_TWEET_BINS(SIZE_OF_BIN)
1   for each Raw-Tweet in DataSet
2       RawTweet ← REMOVE_STOP_WORDS (RawTweet)
3       RawTweet ← REMOVE_SHORT_WORDS (RawTweet)
4       RawTweet ← REMOVE_USER_DETAIL (RawTweet)
5       RawTweet ← REMOVE_NONTEXT_CONTENT (RawTweet)
6       RANK_TWEET (RawTweet)
7       do if No of Tweets in Tweet-Bin >= SIZE_OF_BIN
8               then    SAVE_BIN_TO_DB(Tweet-Bin)
9                       Tweet-Bin ← NULL
10      end
11      Tweet-Bin ← Tweet-Bin + RawTweet
12  end
```

Figure 3 Algorithm to Prepare Tweet-Bins

Fig.3 shows algorithm which cleans the raw tweets and prepares logically equal Sized Tweet-Bins. Variable SIZE_OF_BIN gives information about number of Tweets in Tweet-Bin or time duration.

PROCESS_BIN (*Tweet-Bin*)

```
1  SEGMENTATION_OF_TWEETS (Tweet-Bin)
2  SUPER_WEIGHT ← FIND_SUPER_TRENDY_TERM (Tweet-Bin)
3  for each TERM in Tweet-Bin
4       UPDATE_TRENDINESS_DISTANCE (TERM, TERM.WEIGHT, SUPER_WEIGHT)
5  end
```

Fig.4 Shows Algorithm to Process the Tweet-Bin, which contains collection of Raw Tweets. It finds Super Trendy Term of given Tweet-Bin, saves Super Weight & Calculates and updates Trendiness distance of each Term in Tweet-Bin.

In this approach, customized logic with few approved methods has been utilized to find out Trendy Terms. Different functions and their algorithmic representation are given below. Different functions of algorithms (Fig. 3 & Fig. 4) are described below. Function REMOVE_STOP_WORDS takes *Tweet* as an argument and removes English, Hindi & Gujarati Stop words from the Tweet and returns modified tweet. Function REMOVE_SHORT_WORDS takes *modified Tweet* as an argument and it removes all words having length less than or equal to predefined minimum length (in current context, 3 has been used as a predefined minimum length) and returns following modified Tweet. Function REMOVE_USER_DETAIL takes *latest modified Tweet* as an argument and it removes all user related information from Tweet as study's primary focus is NOT in 'who is writing Tweet, for whom it is written and who is referred to in that tweet' which finally returns modified Tweet. Function REMOVE_NONTEXT_CONTENT when invoked, it removes all non-text content covering images & hyperlinks from Tweet & returns modified Tweet. Function RANK_TWEET takes *Tweet* as an argument and it emphasizes the particular Tweet (Assigns proper weightage to particular Tweet) based on Re-Tweet (Retweet can be defined in general term as "when you pass another user's Tweet on your personal profile to disseminate the message among your followers". Retweet also can be considered as a method of supporting the idea of original message so that the followers would see the original thought. E.g. one wishes to retweet investment ideas of a professional banker that one believes in, a movie review which one finds worthy or a government scheme that one feels beneficial for one's followers.) and based on Like-Count (It is a true parameter highlighting much important derivations like *Bookmarking Tweets, Expressing Appreciation, Showing Acknowledgement, Automatic Support, Non-Verbal Communication, Personal Involvement, Friendly Support*). This function updates importance of given Tweet in database. Function SAVE_BIN_TO_DB takes bunch of Tweets as an argument and saves it as a separate logical Tweet-Bin in database. Function SEGMENTATION_OF_CONTENT takes a particular Tweet-Bin as argument and performs word-segmentation of given Tweet-Bin. Function FIND_SUPER_TRENDY_TERM accepts a Tweet-Bin as an argument and returns Super Distance which is the weight of Top Trendy term of that Tweet-Bin. This Super Distance is further going to be used to find out Trendiness Distance of other terms in same Tweet-Bin. Function UPDATE_TRENDINESS_DISTANCE accepts three arguments: Term, TermWeight in given Tweet-Bin and Super Distance, this function calculates Trendiness Distance of given term and updates the database.

4.4. Data Set

While working with microblogs, it is always challenging to prepare / identify dataset on which research experiments need to be carried out. Looking at the feasible aspect, It is more practical if one gets real-life Tweets as dataset. Twitter provides API to fetch random sample of Tweets. Twitter provides facility to read base on geographical region, we just need to pass Longitude, latitude & radius to get selected regional tweets. In this experiments, tweets have been accumulated of different Indian cities of specific date range. Two different datasets have been prepared of different cities with different date range.

Given data set is prepared using steam API provided by Twitter. All Tweets are real life tweets from different regions & duration. One can say dataset is an actual society generated random sample. Detail of dataset is given below

- No of Tweets: 4, 44,456
- Geographical regions (India): Delhi, Gujarat, Rajasthan, Maharashtra
- Date range: 01-11-2016 to 05-11-2016

4.5. Experimental Setup

Now the first challenge is how to represent row tweets in such a way that one can utilize information available with these tweets in different mathematical & logical calculations & algorithms. Referring the term "TREND" - it says something "more in discussion". So, genuinely one would be interested to find out what is maximum in discussion? The first step is to pre-process all tweets before one applies algorithm / calculation to intermediate data. Output ought to be expected is "Trendy Terms / Hot Terms / Bursty Words" in discussion.

Following are pre-processing steps carried out on row tweet**s**

- Cleaning data: Cleaning is the process to remove extraneous information available with dataset hence to save computation power
    - Remove stop words (English / Hindi / Gujarati): List of stop words has been collected from different websites and created custom stop word collection.
    - Remove very short words: Another cleaning concept is to remove words with length less than or equal to two (One is expected to assume words with length greater than two do have more probability for contributing to generate trend, also this step is used to reduce dataset size).
    - Remove User Detail –as current study does on focus on who is writing, one just needs to grab what is "Trendy" irrespective of who is writing.
    - Remove hyperlinks & images: Staying to study's primary objective, one is expected to abstract trendy terms from text by avoiding hyperlinks and other multimedia content from tweets.
- Segmentation of tweets

Now analysis is ready with *processed tweets* on which one can apply different methodology to get *Trendy Terms***.** As tweets are written by users at different time intervals, whichever dataset generated is kind of time series data. One needs to logically divide processed data into equal sized bins (One can divide data based on equal time interval or based on equal number of tweets in each bin order by actual tweet time). In carried out experiments, bins of equal no. of tweets as input have been prepared to main algorithm. One needs to measure counts of mentions, hashtags, special symbols or any other quantity that can be considered as timely discussions of tweets which contribute in trend. One should give

more importance or rank to those quantities which are re-twitted / supported by more people i.e. quantities which have been again supported by more people have *considerable significance* towards trend. No importance has been given to any user as study focus mainly on *trend from discussion* irrespective from who is into discussion. But one can give user an importance by considering friends / followers / retweet count of original user tweets.

4.6. Results

The result portion includes thorough findings generated from one's experiments and analysis based upon the methodology one carried to amass communicable knowledge. Things that are or can be known about a given topic or systematic imparting of knowledge, education and training. This portion ought to describe the findings of the carried out thesis organized in a logical sequence without presumption, bias or interpretation. A 'Result-corner' is universally indispensable if paper incorporates data generated from own research. Results are the proof of experiments.

Here as a dataset *Real sample tweets* have been taken into consideration provided by Tweeter API from the date range 12-11-2016 to 22-11-2016. One can note that on 8 November 2016, the Government of India announced the DEMONITISATION, also called NOTEBANDHI in vernacular, of all ₹500 and ₹1,000 currency notes. (The term DEMONITISATION refers to official ban on partial / all existing denomination considering financial reform in long run. In the current context, it is banning of ₹500 and ₹1,000 currency notes.) . This was a *whistle-blower event* in *the then* current time particularly in November and even in subsequent 3-4 months. Considering the direct analogy of the above phenomenon, study logically forced us to grab the above mentioned paradigm shift of Indian economy under study-analysis. During thorough analysis, top trendy terms were extracted which is effective proof of real incident 'DEMONITISATION'. During this cognitive process, some of the more trendy terms were identified: *'people', 'new', 'que', 'india', 'modi', 'money', 'black', 'demonetisation', '2016', 'bank', 'notes', 'cash', '**नोट**', '**देश**', 'govt'* and many more. This is the enlightening testimony of study's main objective and shows why this study is much significant. Different geographical clutters of India like Delhi, Maharashtra, Gujarat, Rajasthan, Panjab were discussing about "Demonetisation 2016 in India" throughout this particular time span. Terms *'que', 'people', 'bank', 'cash'* strongly imply that *people are standing in a queue at Bank to submit cash / notes and talking about this matter on Tweeter.*

Table 1 List of Top Trendy Terms

Sr.	Term	Weight	Sr.	Term	Weight	Sr.	Term	Weight
1	people	9982	26	life	4000	51	says	2690
2	new	9770	27	#demonetisation	3640	52	going	2681
3	just	8989	28	great	3583	53	नोट	2680
4	like	7714	29	watch	3538	54	way	2647
5	que	7403	30	need	3511	55	govt	2511
6	dont	7022	31	2016	3311	56	years	2491
7	love	6809	32	man	3288	57	free	2465
8	good	6485	33	year	3253	58	help	2447
9	india	6403	34	देश	3248	59	work	2418
10	time	5849	35	bank	3246	60	news	2410
11	modi	5729	36	sir	3244	61	#tvpersonality2016	2401
12	day	5621	37	look	3085	62	demonetisation	2396
13	money	5561	38	thank	3028	63	win	2387
14	best	5055	39	notes	3010	64	indian	2387

15	know	5019	40	come	2826	65	media	2374
16	make	4661	41	think	2810	66	क्या	2371
17	trump	4612	42	got	2790	67	say	2347
18	hai	4553	43	thanks	2782	68	follow	2345
19	happy	4532	44	cash	2754	69	delhi	2338
20	video	4395	45	live	2737	70	big	2318
21	today	4317	46	world	2733	71	sex	2311
22	मोदी	4281	47	did	2693	72	things	2303
23	vote	4278	48	really	2692	73	real	2296
24	want	4037	49	rig	2691	74	better	2227
25	black	4020	50	days	2691	75	support	2199

Table 2 Trendiness Distance Values for term ‘que’ which represents ‘Queue’

Bin No	**Trendiness Distance in %**	**Bin No**	**Trendiness Distance in %**	**Bin No**	**Trendiness Distance in %**	**Bin No**	**Trendiness Distance in %**
1	72.50	25	49.59	49	80.00	73	17.39
2	67.50	26	32.56	50	69.63	74	0.00
3	61.01	27	0.00	51	59.46	75	0.00
4	52.90	28	0.00	52	44.17	76	34.75
5	56.04	29	46.48	53	52.10	77	66.67
6	66.67	30	78.23	54	17.58	78	63.33
7	0.00	31	70.15	55	0.00	79	73.18
8	0.00	32	73.28	56	0.00	80	70.47
9	74.25	33	38.41	57	27.97	81	33.87
10	79.53	34	47.01	58	72.73	82	44.94
11	74.75	35	26.17	59	76.11	83	0.00
12	76.67	36	84.62	60	61.48	84	0.00
13	66.04	37	0.00	61	55.14	85	45.10
14	46.62	38	0.00	62	36.00	86	73.47
15	58.17	39	17.36	63	32.26	87	76.36
16	18.75	40	61.02	64	38.41	88	59.18
17	0.00	41	75.23	65	0.00	89	69.63
18	0.00	42	69.40	66	0.00	90	19.81
19	13.74	43	39.82	67	46.88	91	30.53
20	83.15	44	33.33	68	75.47	92	0.00
21	83.77	45	43.31	69	50.91	93	0.00
22	70.07	46	0.00	70	52.38	94	27.73
23	62.76	47	0.00	71	78.73		
24	36.22	48	48.21	72	54.84		

4.7. Results

Here in following charts, two parameters have been incorporated: *1) Significance of the term 2) Trendiness-Distance of the term* in different Tweet-Bins.

Kindly note that in each chart, *Solid shaded region* represents the *Significance of the term* while the other region indicates *Trendiness-Distance of the term.*

Term: demonetisation

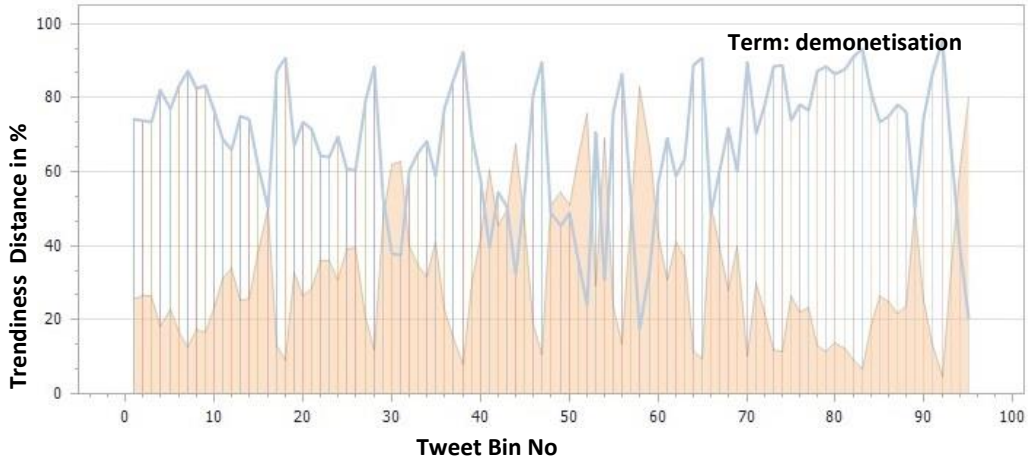

Figure 5(a). Trendiness Distance & Significance of Term "demonetisation" in Different Tweet-Bin

Term: black

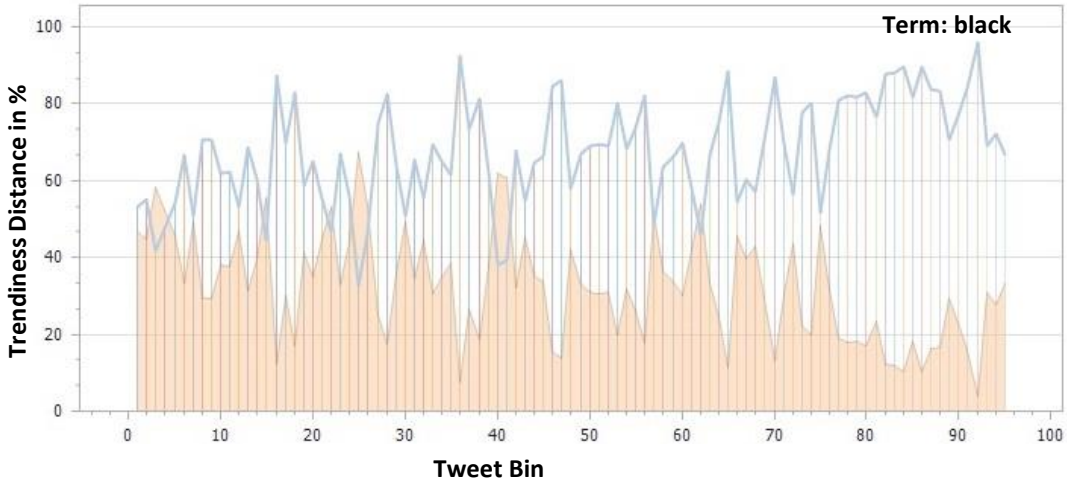

Figure 5(b). Trendiness Distance & Significance of Term "black" in Different Tweet-Bin

Term: cash

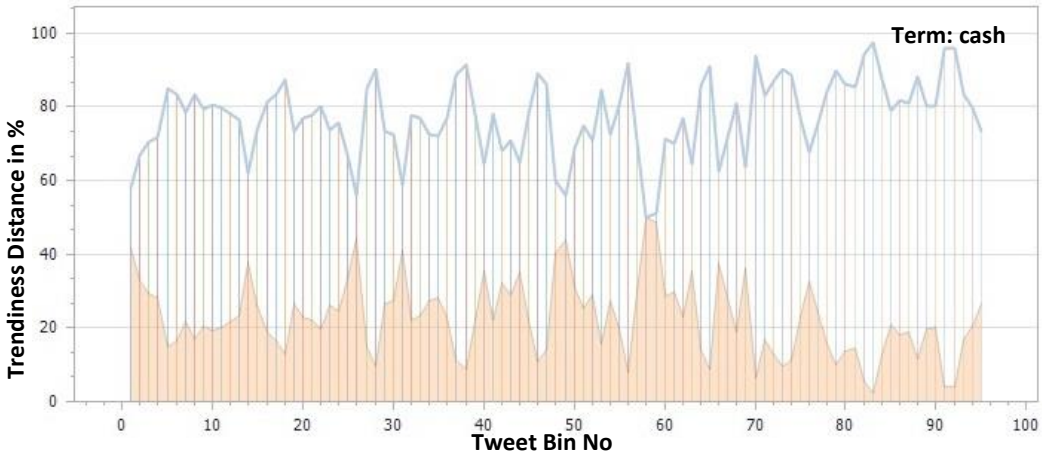

Figure 5(c). Trendiness Distance & Significance of Term "cash" in Different Tweet-Bin

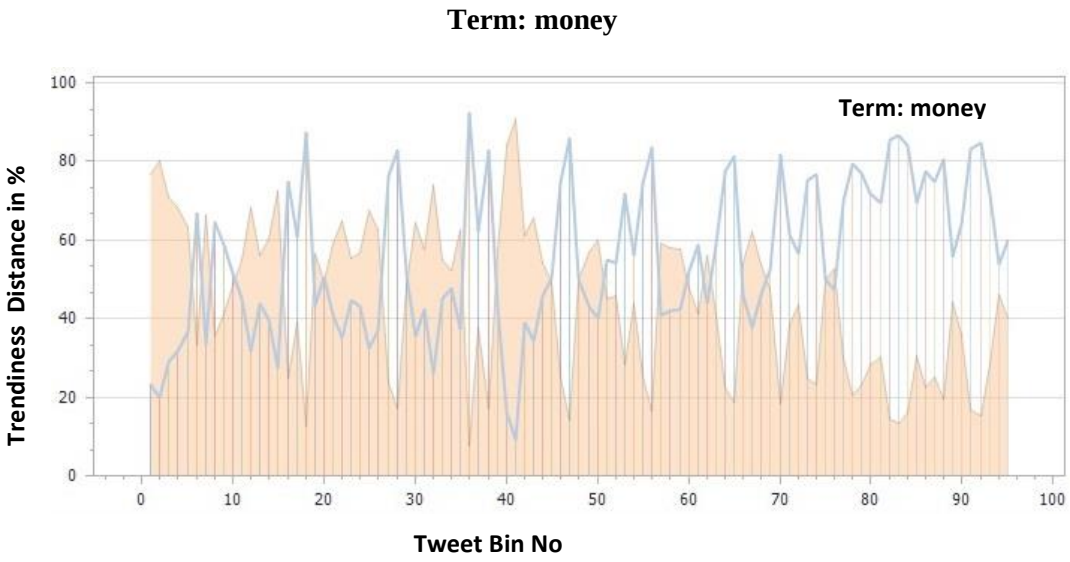

Figure 5(d). Trendiness Distance & Significance of Term "money" in Different Tweet-Bin

5. CONCLUSIONS

As era changes, Microblog has secured a level of proven acceptance as *an established platform* regularly used by multitude to share *emotions, feedbacks, discussions, praises, expectations, sadness & demands*. People are free to write *anything, anytime, anywhere on any diversified topic without any fear***.** Innumerable posts are produced by lots of people daily and thus it becomes quite hard to derive idea about what people are discussing in posts. Considering the usefulness, lot many uses can be drafted here. Naming just few for instance study of various posts from microblogs (1) helps a firm to view feedback of its product / service, (2) helps government to get independent and transparent feedback from denizens, (3) helps political parties to analyse their image in society. But one needs to dig deep in tons of posts. A novel concept *"Term weight – Trendiness distance"* has been proposed that makes identification of trend from millions of post from microblogs speedily as well as easily. This concept further gives confidence of a given term in different time intervals to be trendy, this gives idea about how far given term is from super trendy term. Also the plot of "Trendiness distance" of given term helps to understand about increased / decreased / steady importance of the term in public discussion. In addition to this, one can use this concept to reduce the size of dataset by eliminating terms which do not effect trend. In carried out experiments, not only hashtags but also major entities have been taken into consideration from posts to get real trend of society.

ACKNOWLEDGMENT

This paper would not have seen the light of the day without concrete support of esteemed institute Nirma University. We are sincerely thankful to the Nirma University for providing resources and other facilities to carry out this research work.

REFERENCES

[1] Alexandra Moraru, Janez Brank, Marko Grobelnik, "Trend and anomaly detection in non-structured data", Seventh Framework Program, PlanetData, Network of Excellence, FP7 – 257641 (2012).

[2] Ceren Budak, Divyakant Agrawal, Amr El Abbadi, "Structural Trend Analysis for Online Social Networks", Proceedings of the VLDB Endowment, Vol. 4, No. 10, 2011.

[3] David Alfred Ostrowski, “Identification of Trends in Consumer Behavior through Social Media”, Proceedings on the International Conference on Artificial Intelligence (ICAI) (2012).

[4] Hila Becker, Mor Naaman, Luis Gravano, Beyond Trending Topics: Real-World Event Identification on Twitter, Association for the Advancement of Artificial Intelligence (2011)

[5] Luis González, Iván Pino, “Early identification of social media opinion trends is crucial in crisis management”, d+i is the LLORENTE & CUENCA Centre for Ideas, Analysis, and Trends (2014).

[6] Su Gon Cho and Seoung Bum Kim, Identification of Research Patterns and Trends through Text Mining, International Journal of Information and Education Technology, Vol. 2, No. 3, June 2012.

[7] Gediminas Adomavicius and Jesse Bockstedt, “C-TREND: A New Technique for Identifying Trends in Transactional Data”, Winter Conference on Business Intelligence, 2007.

[8] Clyde Holsapple, Shih-Hui Hsiao, Ram Pakath, “Business Social Media Analytics: Definition, Benefits, and Challenges”, Twentieth Americas Conference on Information Systems, Savannah, 2014

[9] Karen Sparck Jones, "Statistics and retrieval: past and future", International Conference in Computing: Theory and Applications (Platinum Jubilee Conference of the Indian Statistical Institute), Kolkata, IEEE, 2007.

[10] Manning, C.D.; Raghavan, P.; Schutze, H. "Scoring, term weighting, and the vector space model". Introduction to Information Retrieval - 2008.