Mining Twitter Feeds for Top Stories

CS 750 Data Mining Term Paper

Kshitiz Bhattarai
George Mason University

***Abstract*—Twitter, an online social media service, has become increasingly popular in the last few years. Twitter allows user to post short messages of maximum 140 characters also known as "tweets", which have been interestingly used in the past to communicate breaking news, eyewitness accounts and public opinion. One interesting problem in tweet analysis is automatic detection of topics being discussed in tweets. Prior work has explored the use of clustering and topic modeling to cluster tweets by topic, but clustering of Tweets is very difficult due to lack of enough features, use of poor grammar, noisy data, etc. Hence, applying traditional natural language processing algorithm in such data is challenging. Latent Dirichlet Allocation is an unsupervised machine learning technique which identifies latent topic information in documents. Our goal is to use LDA to cluster tweets into fixed categories and as well as summarize each cluster by picking top representative tweet. In this paper, we detail our experiments of finding hidden topic using LDA and measure its effectiveness against other clustering based algorithm based on hashtags labeling and conceptual labeling.**

Keywords-component; tweets, document clustering, lda

I. INTRODUCTION

Twitter is a new form of communication; it allows users to post messages, also known as "tweets" of about maximum 140 characters that have been interestingly used in the past to communicate breaking news, eyewitness accounts and public opinion. One interesting problem in tweet analysis is automatic detection of topics being discussed in tweets. While Twitter allows user to label their tweets sing hashtags so other can follow these interests. These hashtags, which are created by user, may or may not represent the actual topic for that tweet or in some cases represent very narrow topics. For example "Did you watch Heats burn bulls last night #heatalltheway @heat_fan222 http://bit.ly/lbasti6". This tweet belongs to meta-topic of "Sports" or "Personal" and actual topic of "Heat vs Bulls" or "NBA Game" while the hashtag had some classification information but it is very narrow. A second example where hashtags barely help on classification, "How many lucky guys are drinking their green beer today out of this?? #smgirlfriends cc: @mamabritt @dabneyporte", while #smgirlfreinds does not have any classificatory power but "green beer" actually implies it was about celebrating "St. Patrick's Day" which might be considered as Personal category of tweet.

There are roughly around 340 million tweets per day, most of tweets are personal and may not hold significance for business or even users who may just want to browse interesting topics like news, sports, entertainment, technology, finance etc. There have been many works in clustering and classifying tweets but searching for interesting topics within all these tweets is a novelty. Unfortunately, existing algorithm were created for long, structured, grammatically correct text, while tweets are highly non-standard, short, and grammatically incorrect. Hence, applying traditional natural language processing algorithms in such data is challenging. We purpose that after preprocessing tweets we will be able to use existing NLP techniques on this data. Furthermore, we purpose that Latent Dirichlet Allocation can be used to learn hidden topics from tweets and cluster them into fixed categories.

In this paper, we first discuss prior work done on tweet preprocessing and clustering, topic detection, etc. Next we formulate our approach to cluster tweets for fixed topics, dataset, experiments and comparison against other document clustering algorithms. Since our main goal is to summarize tweets, we also discuss how LDA can be used to garner top stories. Finally we offer a discussion of our results.

II. BACKGROUND

Topic modeling is gaining huge attention in text mining communities. LDA [1] is becoming a standard tool for topic modeling. While LDA developed by Blei, Ng and Jordan was specifically for large documents, but in recent years LDA have been extended to social network topic modeling as well. For example, Rosen-Zvi et al. [2] introduced an author-topic model, which can flexibly model authors and their corresponding topic distributions. Connor et. la [3] uses an unsupervised approach to modeling topics. Although they do not provide their test results but their application "TweetMotif" allows user to browse tweets by Themes. Rosa et. al. [4] were able to successfully implement LDA to model topics into six predefined topics. They used hash-tags in the tweets as the gold standard for performance metrics. For summarizing the tweets cluster they proposed an algorithm based on a document novelty selection technique. Beverugen et. al [5] implemented bisecting k-means to produce summaries of Tweets and they found using bigrams as feature selection provides the best cluster validity.

While, we were mostly influenced by Rosa et. al. work on using LDA to model topics but for producing summaries or in our case picking top representative tweets, they used an algorithm based on document novelty selection technique and for they used "Relevance Judgement" for measuring performance of summary produced, which required a lot of money and manpower. So we are purposing that instead of using supervised technique like Rosa et. al., we could use unsupervised summarizing technique (LDA) to pick the cluster representatives.

III. PROBLEM DEFINITION

The most interesting problem in tweet analysis is automatic detection of topic. We purpose Latent Dirichlet Allocation can be used to analyze tweet for learning hidden topic that can be matched into fixed predefined categories. In this project, we plan to use LDA for topic detection, and based on detected topic we plan to summarize or pick top trending stories in a given topic. Since picking top representative tweet of a topic requires the top tweet to be as representative as possible as well as diverse as possible from each other, we plan to use LDA, since it can model latent topics which are representative and diverse from each other.

The major motivation for this idea was from News Websites, like Yahoo and Google news; they have a section called "Top Stories" where they feature top stories from other news sites as well. Similarly, we wanted to mine Twitter feeds for first six predefined topics, namely News, Finance, Sports, Technology, Personal and Entertainment. Based on these topics pick top 3 representative tweets from each topic. So in this project, our main research questions that we plan to address are as follows:

Preprocessing techniques - Since the tweets do not follow old document model used in NLP algorithms. How can we preprocess the raw tweets that could be used for algorithm likes LDA and k-means?

Clustering - Can we cluster tweets into six predefined topics?

Top Stories - Can we summarize tweets of each cluster to come with top tweets like Google News gathers top news?

IV. DATASET

We have collected tweets based on keyword search and user profile from http://twdocs.com/. Our dataset was collected from fixed users for each category as shown in Table I and for the personal we employed keyword search. We collected 25,000 tweets in a span of 5 weeks, where 5000 tweets were collected each week, but used only 5000 for the training and testing.

TABLE I. CATEGORY AND SOURCE

Category	***Source***
News	cnnbrk,nytimes,reuters,usatoday,bbcworld
Finance	bloomberg,cnbc,wallstreetjournal,forbes
Entertainment	hollywood reporter,imdb,tmz,eonline
Technology	AppleNews,microsoft,windowsphone,android
Sports	lasportnews,skysportsnews,sportingnews,espn
Personal	#ilikepeoplewho,#liesihavetoldtomyparents

Twdocs allows exporting twitter feeds from either keyword search or user profile into xml format. Since these tweets contains many twitter specific symbols (like @, #, RT, TT), punctuations etc, we will be heavily preprocessing them.

V. PREPROCESSING TWEET

In order to convert raw tweets to be used by algorithm we performed some normalization steps to reduce the feature space. We experimented with multiple variations of our vocabulary, as shown below.

1. *Method 1:* White Space Tokenization, lowercase conversion, Stopwords removal, un-escaping Html format (conversion from &amp to "&", etc.), short length tweets (length ≤ 3) removal
2. *Method 2:* Previous plus Stemming words and rare terms (whose frequency is less than 3) removed.
3. *Method 3:* Previous plus shortened URL was expanded and their domain was gathered and kept as extra feature.

For preprocessing we employed JFlex [6], which is a fast and easy tool for scanning texts in documents. We ignored most of the punctuation except for "#" and URL related punctuation (like ". : / = ?"), which was kept as some of these hashtags might have classificatory information within. For Stop words list we used [7], which provided a list of mostly used stop words list, but added some extra words which was mostly used in Twitter like "RT", "TT", "ll", "now",etc. Basic steps of preprocessing are shown in Table II.

TABLE II. PREPROCESSING PHASES

Phase	***Tweet***
Raw Tweet	RT @Wiz: #iLikePeopleWho remain true to who they are, & http://bit.ly/doY03A
HTML Unescape	RT @Wiz: #iLikePeopleWho remain true to who they are, & http://bit.ly/doY03A
Lex Tokenization	rt ilikepeople who remain true to who they are http://bit.ly/doY03A
StopWrods Removal	ilikepeople remain true http://bit.ly/doY03A
Stemming + rare terms removal	ilikepeople remain true http://bit.ly/doY03A
URL Expansion	ilikepeople remain true wizkhalifa.com

VI. EXPERIMENTS

For our experiments, we were interested in evaluating our topical clustering techniques, we labelled the data obtained from each source as News, Finance, Entertainment, Technology, Sports, and Personal, this process is hereby referred to as conceptual labeling. We are assuming that everything tweeted by "cnnbrk" will be about News and similarly to each source. Hence, we will be labelling tweets based upon its source and use it as a gold standard for measuring clustering quality. Even though this seems like a strong assumption to make, but we randomly picked 50 tweets from each source and found that class labelling based on user source is accurate.

After clustering the training set, we assign each cluster a class based upon majority rule. To measure we will compute the F-score based on label. To calculate F-score, we first define True-Positive (TP) as correctly clustered tweets. False-Positive (FP), True Negative (TN), and False Negatives (FN) are similarly defined. With these definitions, Precision (P), Recall (R), and F-score can be calculated as :

$$P = \frac{TP}{TP + FP} \quad R = \frac{TP}{TP + FN} \quad F - score = \frac{2PR}{P + R}$$

For LDA algorithm, we decided to modify Gregor Heinrich implementation of LDA with Gibbs sampling [8]. Before we could head into final LDA clustering we needed to see which preprocessing techniques provided the best result. Table III shows the observed F-score for different version of our preprocessing techniques.

TABLE III. PREPROCESSING TECHNIQUES

Classes	*Method 1*	*Method 2*	*Method 3*
News	0.63	0.65	0.65
Finance	0.59	0.84	0.68
Entertainment	0.86	0.49	0.4
Technology	0.77	0.81	0.86
Sports	0.61	0.67	0.45
Personal	0.45	0.72	0.62

Based on the F-measure result from the sample training set on these different approaches we decide to use number two as our preprocessing choice.

Since after preprocessing we saw some of the tweet's feature space or number of words in tweets drop drastically, we did a test on the effect of feature space vs LDA's performance. We only tested it with the minimum requirement for tweet's length to be at max five. Since at 6, we were removing more than 50% of tweets. Based on the result as shown in Table IV, we decided to use minimum number of words per tweets as 3, and removed any tweet that was less than 3.

TABLE IV. FEATURE SPACE VS LDA PERFORMANCE

Minimum Number of Words per Tweet	*Average F-score*
1	0.64
2	0.57
3	0.65
4	0.60
5	0.51

Since our training corpus was huge we decided to see the impact of training set size on LDA's performance. Figure I shows the effect of training size on LDA's measure.

Based on the results we decide to use a subset of training set for rest of the project, 500 tweets from each category. Since on average we received about 0.65 as our average F-measure with different training set size, training set did not have a huge impact on LDA. Since the data set was read in incrementally, the drop you see here is actually due to addition of separate source's tweets which will bring in new keywords. For example first source in our news category is BBC, and second one is CNN. The tweets from BBC were mostly about international news, while CNNBRK was mostly about US news, hence, new never before seen keywords were added upon reading second source. Once LDA received enough amounts of tweets it was able to classify them correctly.

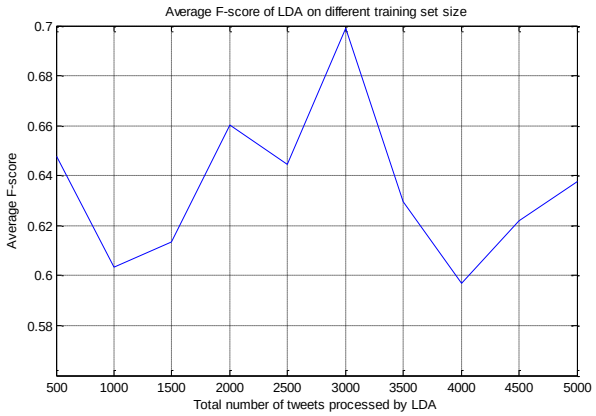

FIGURE 1 LDA PERFORMANCE ON TRAINING SET

For this subset we performed LDA and Table V shows the top words for each topic produced by LDA. Since, we forced LDA to learns only six topics, some of the words returned by LDA on topic does not match. For example for Topic finance, we saw *romnei* and *cnnelect*, but it should actually belong to Topic News.

TABLE V. TOP WORDS FOR EACH TOPIC PRODUCED BY LDA

#	*Top Words*	*Labeled as*
1	Year,watch,top,todai,market,down,take,on,end,win,out, dai,rate,week,stock,share,romnei,cnnelect	Finance
2	Over,report,against,kill,uk,more,soon,presid,state,afgha nistan,tell,obama,syria,detail,peopl	News
3	Appl,new,ipad,android,app,iphon,updat,googl,post,mar ch,blog,announc,releas,launch,tablet	Technology
4	Ssn,live,come,more,leagu,ranger,latest,join,first,goal,te am,sport,action,look,tonight	Sports
5	Video,game,season,exclus,film,sxsw,final,two,show,st ar,tv,chang,review,talk,plaeyfest,photo,music	Entertainment
6	Ilikepeoplewho,liesivetoldmypar,make,go,don,know,ba ck,test,fail,home,keep,realli,good,hous,watch	Personal

Hence, we looked at the confusion matrix generated by LDA (also shown in Table VI), and we can observe that some of the class like news were classified into two different categories, namely politics was learnt as a separate topic than news and was classified as Finance. This meant we ran into problem of non-globular clusters. To solve this we decided to use more number of topics than just six then combine than later on.

TABLE VI. CONFUSION MATRIX FOR SIX TOPICS

	Classified					
True	*Topic 0*	*Topic 1*	*Topic 2*	*Topic 3*	*Topic 4*	*Topic5*
Personal	342	0	4	0	1	1
Entertainment	17	338	21	21	41	61
Sports	59	35	308	5	13	80
Technology	15	13	10	434	3	25
News	19	35	56	13	325	52
Finance	40	22	53	66	88	231

Hence we tested the impact of number of topics on F-score by running LDA from number of topics 6 to 18.

Table VII shows the result obtained from running LDA on different number of topics and then combining the topic based on majority rule.

TABLE VII. F-SCORE FOR DIFFERENT INITIAL NUBMER OF TOPICS

Number of Topics	*Average F-score*
6	0.64
7	0.65
8	0.70
9	0.69
10	0.66
11	0.68
12	0.64
13	0.65
14	0.68
15	0.66
16	0.64
17	0.65
18	0.63

As you can observe, on average F-score did not change a lot. It increased slightly till 8 then went back down but on average it lingered around 0.65. Hence, it implied increasing a number of topic, LDA did not just learnt a new topic, but also misclassified a few more tweets since LDA assigns each word to every topic with some probability, thereby, on average keeping the F-score constant. Since we made a fixed assumption at the beginning that tweets can be labeled with respect to its originating source maybe this was a too rigid of an assumption. Hence, we decided to use reverse hashtags clustering. In which we first go through all the tweets and use the only tweets that have hashtags with them and ignore the rest. Secondly, go through this list and only use top 10 frequent hashtags as our labeling and perform LDA on these tweet dataset with number of topics equal to 10 and each topic had minimum number of tweets of 35 and maximum 50. Below is the result of F-score of each labeled tweets.

TABLE VIII. F-SCORE FOR DIFFERENT HASHTAGS AS TOPICS

Hashtag	*F-score*
#sxsw	0.93
#liesivetoldmyparents	0.99
#bracketinsanity	0.98
#ilikepeoplewho	0.98
#cnnelections	0.9
#ssn	0.96
#imsickof	0.91
#oscars	0.94
#whydoyou	0.98
#apple	0.92

To our surprise LDA performed really well on this dataset. We thought this might have been due to small dataset. Hence we employed twdocs again to gather a big dataset, each hashtag with 300 tweets and used LDA again. Table IX shows the result and we observed good performance again. Hence, using hashtag as label seems to do good for LDA, but some tweets which were labeled with multiple hashtags do suffer. For example, *"#nba #miamiheat #knicks Miami leads 2-0"*. In our experimentation, we did not allow duplicate tweets; we labeled this tweet to the first hashtag. Secondly, hash-tagged labeling seems to do a better job because of the presence of hashtag in each of the tweet vs. the conceptual labeling that google, micrsoft and apple all are Technology.

TABLE IX. F-SCORE FOR DIFFERENT HASHTAGS AS TOPICS

Hashtag	*F-score*
#apple	0.93
#whydoyou	0.81
#google	0.8
#katyperry	0.97
#ssn	0.77
#windows8	0.87
#imsickof	0.85
#miamiheat	0.77
#corporategreed	0.84

VII. COMPARISION AGAINST OTHER ALGORITHM

After finding the best preprocessing technique, best number of topics for LDA, and sample training size. We wanted to compare the performance of LDA against other document clustering methods. To compare against other methods we decided to use K-means with variety on distance measures, Hierarchical Clustering on different linkage. For K-means and Hierarchical clustering, we decide to use Weka.

We exported the training data set as a text file from JAVA and imported it in Weka along with each tweet label. Using Weka's StringToWordVector (SWV) and StringToNGram (SNG) preprocessing method to convert the tweet corpus to bag of words, we performed the same experiment on Weka. Table VII shows the result obtained from Weka's classes to cluster evaluation.

TABLE X. PERFORMANCE OF CLUSTERING ALGORITHMS

	F-score					
Algorithm	*Entmt*	*Finance*	*News*	*Personal*	*Sports*	*Tech*
LDA	0.65	0.84	0.49	0.81	0.67	0.72
Kmeansdifferent distance option						
Manhattan	0.33	0	0.33	0.53	0.02	0.67
Euclidean	0.36	0	0.32	0.53	0.02	0.88
Hierarchical clustering different linkage option						
Single	0.29	0	0	0	0	0
Average	0.29	0	0	0	0	0
Complete	0.29	0.02	0.04	0	0	0
Mean	0.01	0.07	0.1	0.37	0.35	0.57

Secondly, we also tested these algorithms in reverse hash tagged dataset as used in Table IX. We observed a great change in performance in both algorithms. Since, each method used bag of word model, k-means and hierarchical performed better due to the presence of hashtag in each of the tweet. Since these strictly follow old document natural language processing model, having a word in common allows them to cluster it vs. the conceptual labeling.

TABLE XI. F-SCORE FOR DIFFERENT ALGORTIHM

Hashtag	*LDA*	*K-means (Euclidean distance)*	*Hierarchical (Mean Linkage)*
#apple	0.93	0.98	0.63
#whydoyou	0.81	0.46	0.71
#google	0.8	0	0.8
#katyperry	0.97	0.01	0.84
#ssn	0.77	0.87	0.76
#windows8	0.87	0	0.65
#imsickof	0.85	1	0.02
#miamiheat	0.77	0.98	0.8
#corporategreed	0.84	1	0.84

Based on the result we can conclude that LDA outperforms Kmeans and Hierarchical on both dataset.

VIII. CLUSTER SUMMARIZATION

After receiving a cluster of tweets, one natural extension would be to automatically summarize the topics and stories being discussed within the cluster. We purpose that finding these top representative tweets can be obtained from using LDA. Since picking top representative tweet of a topic requires the top tweet to be as representative as possible, as well as diverse as possible from each other, we plan to use LDA, since it can model latent topics which are representative and diverse from each other. The major drawback of using LDA to pick top stories is that we do not have an efficient method to test our methodology.

Hence, we used LDA to mine for 3 topics or 3 top tweets on each category. Table XII shows the top 3 tweets in each category.

TABLE XII. CLUSTER SUMMARIZATION

Topic	*Top Tweets*
Entertainment	PaleyFest: 'Two and a Half Men' (@TwoHalfMen_CBS) Producers Address Season 10, Ashton Kutcher's Return http://t.co/0iwJywar #PaleyFest
	Box Office Report: 'John Carter' Earns Weak $9.8 Mil on Friday, 'Lorax' Will Win Weekend http://t.co/s20sTSGN
	Davy Jones, best known as the lead singer of The Monkees, died this morning after suffering a heart attack. http://t.co/NbVZuENw
Finance	BREAKING: Crude oil jumps over $110 in electronic trading. Crude oil surges on report of pipeline explosion in Saudi Arabia.
	Tomorrow @ 10pET on @CNBC. Watch #60Minutes for an inside look at Carl Icahn. Video: http://t.co/EZ5oC4Y5
	Earnings Alert: Urban Outfitters (URBN) Q4 EPS $0.27 vs. $0.29 Est.; Urban Outfitters (URBN) Q4 Revs. $731m vs. $740m Est
News	Mitt Romney will win the Arizona Republican primary, CNN projects based on exit polls. #CNNElections http://t.co/z0d4H8zR
	#Syria Red Crescent team which entered Baba Amr district #Homs accompanied by UN humanitarian chief Valerie Amos http://t.co/zYLXIhED
	Liquidation "inevitable" for Glasgow's Rangers football club - director Dave King says in statement. Details follow http://t.co/ZvsAfWiI
Tech	In Austin for #SXSW? Visit the @Microsoft lounge & unlock an exclusive Foursquare badge. Follow at http://t.co/QWOBr7p1 first!
	iPad dispute signals new era in trademark troubles - San Francisco Chronicle http://t.co/jRleODI8 #apple
	We're doing a live Twitter chat tonight with @MomItForward's Girls' Night Out about sharing moments. Join in with #gno #TryWindowsPhone
Sports	Good morning! We've got the latest on Chelsea's managerial situation on #SSN this morning & more reaction to Andre Villas-Boas' sacking
	Chris Paul suffers nasal fracture, will wear a protective mask http://t.co/vaQExuvf
	If you can only watch game around lunch, which are you picking: UConn/Syracuse (noon ET) or Baylor/Kansas State (12:30 ET)? #BracketInsanity
Personal	"@ThomasNutbrown: #ImSickOf Justin Bieber and One Direction trending every single day" YESSSS!
	#WhyDoYou make me think there's something between us ? When there's probbly nothing..
	RT @TheLazyRules: #LiesIveToldMyParents Mom: I'm on my way home, did you clean that house like I told you? Kid: Yes! *Starts cleaning*

From Table XI, we can see that each top Tweet in a category are as representative as possible and as well as diverse as possible from each other.

Secondly, we also decided to run a second experiment on 10 stories happening currently. The story headline was obtained from cnn.com and Twdocs was used again to collect 50 tweets from each stories using keyword search for example for Topic 1, we looked for "Judge Trayvon disqualify" and so on. We looked at the following stories.

1. Judge in Trayvon Martin shooting case disqualifies herself, citing conflict of interest.
2. Axl Rose apologies to Cleveland for not attending the Hall of Fame event.
3. Twitter trying to end tech patent wars by announcing its new IPA.
4. Breivik, who killed 77 people last summer faces trial today
5. Chelsea beat FC Barcalona 1-0 fan reactions.
6. Fans reaction to Helm, drummer/singer, in his last days of battle w/cancer
7. People reaction to Kuwait Tire Fire youtube video
8. Microsoft master plan to beat Apple and Google with Nokia Lumia 900
9. Nicollette Sheridan is granted a retrial in her case for unfair dismissal from the show Desperate Housewives
10. Fans reaction to Tupac's Hologram performance

Based on these stories, we ran our LDA again to search for top 10 tweets; Table XII shows the one top tweet in each topic LDA has learnt.

We can observe the similar result with this dataset as well, each representative tweet correctly summarize the topic of interest.

TABLE XIII. CLUSTER SUMMARIZATION ON 10 STORIES

Stories	*Representative Tweet returned by LDA*
1	Judge Quits Travon Martin Case, Cites Conflict: Judge quits Trayvon Martin shooting case, citing conflict of int... http://t.co/xgoEtO0v
2	RT @ARTISTdirect: Axl Rose (@AxlRose) Pens Another Letter to Rock and Roll Hall of Fame + Fans http://t.co/ndtQ1Kxe - he apologizes to Cleveland...
3	How expanding Twitter's pledge could end the patent wars - With just two small changes, Twitter's Innovator's Patent... http://t.co/9aGFdpxi
4	RT @CNNMex: Anders Behring #Breivik, autor confeso de los ataques de #Noruega, dice que entrenó con videojuegos http://t.co/vp4d6X9U
5	Lool "@eljefe39: Ure just talkin"@Real_TMp: Lool! Issorai, Chelsea beat barca"@eljefe39: Leave am like that"@Real_TMp: CHELSEA ======>
6	Fans, musicians pay tribute to Helm - Times Herald-Record: The GuardianFans, musicians pay tribute to HelmTimes ... http://t.co/ocZN9Y7C
7	Kuwait Catastrophe 5 Million Tires: Kuwait Catastrophe 5 Million Tires, A massive fire broke out at a Kuwaiti... http://t.co/tdBwHYzR
8	Microsoft's master plan to beat Apple and Google | http://t.co/2XLab9dk | #microsoft #apple #google #windows #windows8
9	Nicollette Sheridan Gets Retrial Date For Wrongful Termination Lawsuit: Desperate Housewives star Nicollette She... http://t.co/On6nIUke
10	(http://t.co/TePKb6Vv) Dr. Dre -- I'd LOVE to See a Jimi Hendrix Hologram: Dr. Dre says he... http://t.co/EEYviKKU http://t.co/9qx66TU0

IX. CONCLUSION

In this paper, we looked at how to preprocess tweets so that they could be used by Latent Dirichlet Allocation algorithm. We used conceptual labeling to label the tweet based on its source as well as hashtag labeling where the first hashtag will be the label for the tweet. We experimented with many preprocessing techniques and found that rare terms removal helps LDA to learn better while URL expansion to domain name deteriorates performance, which can be explained by the frequent use of common domain names like google.com, youtube.com in tweets of different classes.

We observed that forcing LDA to learn just fixed number of topics may not yield the best result as we saw in multiple runs that tweets about elections was being clustered with finance. Furthermore, we thought that this was due to presence of non-globular clusters and hence, we tried different number of topics but LDA's performance did not change much, which can be explained by two factors. First, LDA cannot handle class imbalance problem, hence, by increasing number of topics if each topic does not have sufficient number of tweets it will fail to learn it as a separate topic. Secondly, since LDA assign a word to each topic with some probability hence, LDA did not just learn one new topic but misclassified some as well.

On the other hand, we saw a huge increase in performance on hash tagged labeled tweet dataset vs conceptually labeled, which is due to the fact that LDA is a bag of word model and since, in hash tagged dataset each class has at least one common hashtag.

In our performance comparison tests we tested LDA against Simple kMeans and Hierarchical Clustering using Weka toolkit and found that LDA outperforms both of them.

We then experimented with summarizing and picking the top tweet in each cluster using LDA and we observed that most of the results obtained are as representative and diverse as possible from each other. One of the interesting observations of the picked representative tweet is that they are usually the longest one, as each word on a tweet contributes some value to a Topic. Hence, the longest tweet will usually be closer to its Topic than the shorter one.

X. FURTHER WORK

During preprocessing, we observed that removing some stop words and rare terms, the tweet feature space drops drastically. So for future we would like to use WordNet Vocabulary to add synonym to increase the feature space of tweets. Secondly, Twitter suffers from topic drift phenomena, which implies the topic that are trending today might not be important tomorrow and many even be non-existent a few weeks from now. Hence, trying to create an online LDA algorithm that could learn continuously and adapt to trending tweets would be a great research topic.

Due to time limitation we were not able to compare other clustering algorithm namely Bisecting k-means and DBSCAN, we would like to compare them in future. Also we were unable be come up with an algorithm that could prove the picked top representative tweet using our technique is the best one, hence in future we would like to come up with an efficient algorithm.

REFERENCES

[1] D.M. Blei, A.Y.Ng and M.I. Jordan, Latent Dirichlet Allocation. The journal of Machine Learning Research, 3:993-1022, 2003

[2] M. Rosen – Zvi, T. Griffiths, P. Smyth, and M. Setyvers. The author-topic model for authors and documents. In UAI'04: Proceedings of the 20th Conference on Uncertainty in Ai, pages 487-494, 2004.

[3] Brendan O'Connor, Michel Krieger, and David Ahn. TweetMotif: Exploratory Search and Topic Summarization for Twitter. In Proceedings of the International AAAI Conference on Weblogs and Social Media, Washington, DC, May 2010.

[4] Kevin Dela Rosa, Rushin Shah, Bo Lin, Anatole Gershman, Robert Frederking, Topical Clustering of Tweets

[5] Gary Beverungen, Jugal Kalita, Evaluating Methods for Summarizing Twitter Posts, WSDM'11, February 9-12,2011, Hong Kong, China

[6] http://jflex.de, Accessed February 15, 2012

[7] http://www.lextek.com/manuals/onix/stopwords1.html, Accessed February 15, 2012

[8] http://www.arbylon.net/projects/LdaGibbsSampler.java