Events Insights Extraction from Twitter Using LDA and Day-Hashtag Pooling

Muhammad Haseeb UR Rehman Khan
s1826098@s.tsukuba.ac.jp
Tsukuba University
Ibaraki, Tsukuba, Japan

Kei Wakabayashi
kwakaba@slis.tsukuba.ac.jp
Tsukuba University
Ibaraki, Tsukuba, Japan

Satoshi Fukuyama
s1721691@s.tsukuba.ac.jp
Tsukuba University
Ibaraki, Tsukuba, Japan

ABSTRACT

News extraction from Twitter data is a hot topic. But can we extract much more than just news? The purpose of this research is to find, either news is the only information which can be extracted from Twitter data or it contains much more insights about real life events. So, we introduce a technique for analysis of Twitter's raw content. After pre-processing of tweets data, we apply hashtag pooling and extract topics using available topic modeling algorithm Latent Dirichlet Allocation (LDA) without modifying its core machinery. In the second part, estimated number of tweets per day and correlated top hashtags for each topic are calculated using day-hashtag pooling. Finally, the continues time series graph is constructed for topic analysis. Our findings show interesting results of bursty news detection, topic popularity, people's way to perceiving an event, real-life event's transition over time and before & after affects of a specific event.

CCS CONCEPTS

• **Information systems → Data analytics**.

KEYWORDS

LDA, Hashtag Pooling, Topic Modeling, Time Series Analysis

ACM Reference Format:
Muhammad Haseeb UR Rehman Khan, Kei Wakabayashi, and Satoshi Fukuyama. 2019. Events Insights Extraction from Twitter Using LDA and Day-Hashtag Pooling. In *The 21st International Conference on Information Integration and Web-based Applications & Services (iiWAS2019), December 2–4, 2019, Munich, Germany.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3366030.3366090

iiWAS2019, December 2–4, 2019, Munich, Germany

ACM ISBN 978-1-4503-7179-7/19/12…$15.00
https://doi.org/10.1145/3366030.3366090

1 INTRODUCTION

Twitter is a very unique source of information where millions of users try to sum up an event, trend or their emotions into 140 characters. Diverse users of twitter freely express their thoughts which leads to many topics. Extracting trends from tweet's data could be very handy to know and understand better about real-life events because of huge dataset available and people's interest in it. The application area of twitter is vast including many useful domains such as real-time events detection [11], sentiment predication analysis [12], understanding public health opinions [6], time series topic popularity variation [3] and it's comparison with traditional media [13]. Over 85% of topics are headline news or persistent news in nature when tweets data is classified for trending topics [8]. These topics aren't just only news but also contain reasons and effects of specific events. Also, people's interest is directly proportional to intensity of a specific event and its effects on people's life. As we know millions of tweets are tweeted everyday so it is impossible to extract topics manually. Twitter has hashtag information to follow the trending topics and frequency of tweets per hashtag can give us some information about popularity of a hashtag. But, hashtag is a user generated string and can lead to many topics or sometimes irrelevant information related to one specific topic. So, we develop a technique to find most of the useful information from twitter's raw data using already existing research, topic modeling and hashtag pooling. A new, under-examined but useful type of hashtag pooling known as day-hashtag pooling also greatly effects the most essential part of this research "graph analysis".

LDA is widely used for text classification of documents to topics but, it is not very efficient for short text documents, so hashtag pooling is used for making relatively big documents. Another problem is, LDA doesn't model variation of topics so we need a way to see the topics variation in time series manner to follow trends. Therefore, we introduce a way to transform LDA generated topics of twitter's data to time series trend analysis along with finding the correlations of these topics with hashtags.

2 RELATED WORK

Koike [7] proposed a method that draws a time-series graph to find the bursty topic detection from twitters data individually as well as with correlated news by using Dynamic topic model [1]. Koike applied the DTM to extract 50 topics from a subset of news articles and twitter about *The London Olympic game*. Even though DTM allows the distribution of topics and words to be changed over time, DTM has a drawback in the computational cost, which particularly prevents us to increase the number of topics K to hundreds or thousands. In Twitter, we can believe that there are many diverse topics including almost anything the people talk in their life. Consequently, the limitation of the DTM is a critical issue for the extraction of topics from Twitter.

Latent dirichlet allocation (LDA) [2] is also commonly used for topic modeling problems in natural language processing. It is efficient and results are promising for long documents with regular vocabulary and grammatical structure, such as news articles and scientific research papers. Tweets, on the other hand, are short, contain URLS, usernames, hashtags, and emojis. Spelling mistakes and non-standard abbreviations are also commonly seen in tweets. All these things make it difficult to apply LDA on tweets data without any external help. Applying linguistic preprocessing may somewhat help [4]. But still the main problem is the length of documents which is just 140 characters or less in our case. An intuitive solution to this problem proposed by [5] is tweet pooling which later was proved experimentally [9]. Results shows that hashtag pooling is the best one among others.

Hashtag pooling is making documents based on hashtags and all the tweets with one hashtag form a single document. And any tweet having more than one hashtag is added to the tweet pool of each of those hashtags.

3 PROPOSED METHOD

The goal of this research is to developer a technique in such a way that topic trends in tweets data can be extracted efficiently and analyzed visually. The first step towards our goal is to clean the tweets as much as possible in pre-processing, then we use hashtag pooling to make our documents relatively bigger in size as compared to single tweets. Next step is to apply LDA on hashtag pooled tweets data and extract the topics distribution and top words for each topic which convey the meaning of that specific topic. Before the inference, we apply day-hashtag pooling so that we could be able to track the topic trend on time series graph. Day-hashtag pooling to some extent is a combination of hashtag pooling and temporal pooling proposed by Mehrotra et al. [9] and even though author showed the possibility of hashtag-time pooling scheme but it is almost ignored in past. All the tweets with one hashtag on a specific date are grouped together to make one document and this pooling scheme plays a very important role in this research. Inference in our case, is estimating the total number of tweets belong to each topic in each document. As day-hashtag pooling is applied so estimated number of tweets can be calculated with this formula.

$$N_{dk} = \theta_{dk} \times T_d \tag{1}$$

Where N_{dk} is estimated number of tweets of topic k from document d. θ_{dk} is probability of topic k occurring in document d. And T_d is total number of tweets in document d. θ_{dk} is calculated using Dirichlet distribution by applying LDA on input data.

The final step is to make time series graphs of estimated tweets. Top words and hashtags are also extracted.

4 EXPERIMENT

To achieve our goal of developing a method for trends analysis from twitter data. We used the Tweets2011[1] dataset of more than 3 million English tweets sampled between January 23rd to February 8th, 2011. As the original dataset consists of all the publicly available tweets in that period of time which means tweets are in many languages, we used python library *langdetect*[2] to extract English tweets only. Usually tweets data is very messy so some preprocessing was desirable as a first step for cleaning of this data. So, we removed stop words (the, a, an, in and more) using *nltk.corpus*[3] python package, special characters ($, @, %, & and more), URLs, and words having only two characters because mostly two characters words do not have concrete meaning(up, vs, ha, RT and more). Next step applied on this dataset was hashtag pooling and after applying it we got 275,836 hashtag pooled documents. Then LDA implementing the stochastic variational Bayesian method of [10] in Java with 1000 number of topics, 1000 docs per batch also known as mini-batch size and 1000 number of iteration was trained on hashtag pooled documents. The experiment environment was Ubuntu 16.04 for the OS, 2 Intel 8on E5-2630 (2.40 GHz) 8 cores for the CPU, and Python and Java for the implementation. The training of LDA took one hour 17 minutes and 54 seconds and for the inference part the time for calculating θ_{dk} of documents for topics was just 16 minutes and 20 seconds.

5 RESULTS

Three parts need to be explained in this section. As we have merged and implemented different techniques in this research so it is a good idea to discuss the results step-vise and ultimately combine all the results.

[1] https://trec.nist.gov/data/tweets/
[2] https://pypi.org/project/langdetect/
[3] http://www.nltk.org/api/nltk.corpus.html

The parts are divided into three subsections and discussed in details bellow. First we'll be discussing the topics generated by training of LDA topic model. Secondly, our unique concept of top hashtags for correlated topics is explained. In the last part of results section, Time-series graphs along with top hashtags will be analyzed.

Topics

As already mentioned in experiment section, 1000 topics as an input is used in this research and top 10 words of most of the topics are self explanatory and by just looking at those words we can come-up with topics titles.

	Topic	Top words
1	Music	song, listening, club, track, right, radio, home, hot, high, ill
2	Education	learning, education, past, language, driven, brush, lessons, intelligent, digg, arts
3	USA	american, gov, spread, brotherhood, reform, barackobama, decades, democracy, 500, damon
4	Climate	moon, fine, weather, baro, rising, speed, officialkimora, waning, sunrises, mostly
5	Gadgets	gps, laptop, battery, watch, charger, nike, wifi, tablet, color, high
6	Football	deal, suarez, club, carroll, kenny, player, transfer, players, request, luis
7	Justin Bieber	newmusiclive, justin, bieber, made, tuesday, say, pattie, beiber, till, belieber
8	Photography	photos, m4w, gallery, photographer, camera, w4m, stunning, fleur, photographic, kitty
9	Gaming	xbox, trailer, ops, 360, famous, beta, protests, capcom, unlocked, brief
10	Violence	kills, dead, weekly, iron, headshots, transforming, architects, slayer, tix, attack

Table 1: Top words of topics

In Table 1, very few of the total topics and top words of these topics are shown and we can interpret these words into a title. For example in Table 1 topic 1, from words like song, listening, track, and radio, it's obvious that this topic is related to music. Table 1 topic 2, words like learning, education, language, lessons, and intelligent refers to education in general and american, gov, barackobama, democracy is somehow related to USA in topic 3. So we can claim that we can easily come up with topics titles from the top words of LDA generated topics which can be proved from other examples too.

Top Hashtags

	Topic	Top hashtags
1	Music	nowplaying-2/3, nowplaying-2/8, np-1/27, nowplaying-1/27, nowplaying-2/2
2	Education	8days-2/3, bring5friends-1/23, happybirthdayharry-2/1, twitition-1/27, welovestyles-1/23
3	USA	egypt's-2/5, egypt-1/28, scariestwordsever-2/5, jan25-1/29, sfo-2/6
4	Climate	news-1/32, nowwatching-2/4, aquarius-2/2, zodiacfacts-2/1, unknownwhat-1/26
5	Gadgets	twalue-1/25, lfc-1/30, americanidol-1/27, teamfollowback-2/8, worstpickuplines-2/2
6	Football	gwo-1/29, mbteamcl-2/2, global-2/7, nufc-1/31, aquarius-1/26
7	Justin Bieber	nmlbelieber-1/28, hosting-2/4, hosting-1/29, muchmusic-1/31, hosting-2/5
8	Photography	news-1/23, thegame-2/2, np-2/2, fail-2/7, neversaynever3d-1/24
9	Gaming	twibbon-1/24, twibbon-2/7, magistream-2/5, blackandyellow-2/7, thegame-2/2
10	Violence	jan25-2/2, egypt-2/3, jan25-2/1, jan25-2/4, jan25-2/3

Table 2: Top hashtags of topics

LDA generated topics seems self-explanatory but is this information enough to claim that twitter users truly talked about these topics? We can extract some supporting evidence to proof that people were really interested in these topics and actually tweeted about these topics. So in the inference part we create a new dataset from our original twitter dataset by using day-hashtag pooling technique. In this dataset, all the hashtags having more than 10 tweets are included to make a relatively large and efficient dataset for topics correlation with hashtags. All the tweets with one hashtag of a single day are merged into one document and in total there are 9686 documents. LDA model which was trained on original dataset is applied to this dataset for hashtag relation with topics.

We calculated the θ_{dk} of each topic for every document. Then using equation 1, estimated number of tweets for each topic are calculated. Once estimated number of tweets are calculated we can easily conclude which document in other words hashtag is more relevant to which topic. In Table 2, some topics have strong correlation with their corresponding top hashtags, which states that people were interested and actually talked about these topics e.g Music and Violence(in egypt). But we can also see random hashtags as top hashtags for some of the topics which means people maybe used the words related to these topics in general but didn't explicitly had interest in these topics.

Time Series Graph

Previous two subsections shows the importance of LDA generated topics and correlation of these topics with hashtags. But one piece of information is still missing which is how high the number of estimated tweets is. And the estimated number of tweets on a time-series graph to understand it easily. There were 425906 tweets in total of 9686 documents of inference dataset. As mentioned above, 1000 topics for LDA were used and total of 17 days(Jan 23 - Feb 8) were considered in inference part. So around 25 tweets per topic each day is the equally average value. A very basic generalization is used here just to see the popularity of a topic which is, if the number of estimated tweets is higher than the average value then it implies people were more interested in this topic.

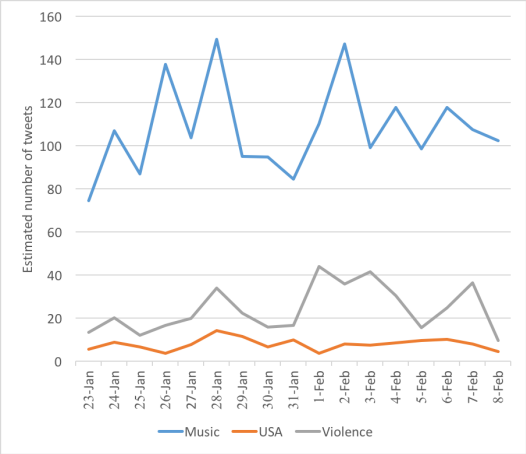

Figure 1: Estimated number of tweets for topic 1, 3, 10

From Table 1 and 2: topic 1, 3 and 10 have strong correlation of top words and top hashtags. But calculating estimated number of tweet per day and visualizing it shows the actual popularity of the topics. In Figure 1, Music which is topic 1 in Table 1 & 2 was way much popular topic as compared to others whereas topic USA which is topic 3 in Table 1 & 2 is bellow average which means people were not much interested in USA's statements about egypt on jan25.

The most interesting part of this research is **Time Series Topic Analysis**.

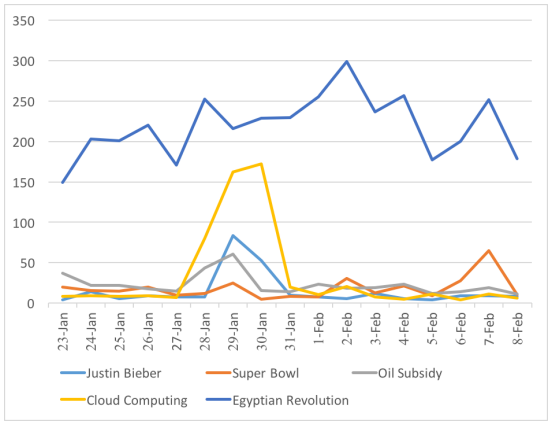

Figure 2: Topics transition over time

From Figure 2, which has date at x-axis and estimated number of tweets on y-axis, we can find the transition of actual events over time and people's interest in these events. we can also find the bursty topics and how high the burst value is by this graph analysis.

For example, in Figure 2, topic "cloud computing" which is topic number 4 in Table 3 has a very high burst on Jan 29-30 and relevant tweet confirmed the reason of this burst which was a real life event. We can observe the similar bursty trends with topic justin bieber, super bowl and oil subsidy which are topic 1, 2 and 3 respectively in Table 3.

An interesting thing observed in this research is, we can discover the actual reasoning and some insights of an event which may not be found by news articles or other source of information in real life e.g. social media played a very important role in the egyptian revolution 2011 which can be seen in Figure 2 as "Egyptian Revolution" and topic 5 in Table 3.

6 CONCLUSION

In this research, we develop a technique to extract topics trends transitions in graphical representation from twitters data without modifying the original machinery of LDA along with the help of hashtag pooling. Calculating estimated number of tweets for each topic tells us the actual popularity of the topics. This paper also shows that with this technique we can detect not only bursty topics but also the level and interval of burstiness. Top hashtags correlation with topics reflects the focus of topics. Analyzing top words of topics, top correlated hashtags and estimated number of tweets all

	Top Words	Top hashtags	Sample Relevant Tweet	Topic Statement
1	wacky, guitar, soca, ensemble, orgy	nmlbelieber-1/29, nmlbelieber-1/30, muchmusic-1/29, nml-1/29, nml-1/30	Justin Bieber on NewMusicLive this Tuesday has made me a NMLBELIEBER	Justin Bieber performed on New Music Live
2	ugh, alcoholic, failure, cans, woke	superbowl-2/6, superbowl-2/7, steelers-2/7, steelers-2/6, sb45-2/7	There are many Super Bowl Parties this Sunday around the Plymouth area at the bars and restaurants.	Super bowl parties
3	oil, obama, cnn, funding, response	cars-2/3, cars, 2/7, us-2/1, whatif-1/27, us-1/29	NYTimes: Obama's Bid to End Oil Subsidies Revives Debate	Oil price subsidy statement by obama
4	computing, could-computingexpo, billion, infrastructure, inc	cloud-1/28, news-2/3, cloud-1/29, cloud-2/2, services-2/2	Data Center Links: PEER 1, Telx, IBM, Unisys: IBM Launches $42 Million Cloud Computing Cente...	Cloud Computing service launch by IBM
5	news, video, facebook, live, blog	egypt-1/31, jan25-2/2, jan25-2/1, egypt-1/29, egypt-2/1	"Internet is a gift from God for all of ""Egyptians"". They shut it down and We were just ""Gyptians"	Social Media played an important role in egyptian revolution

Table 3: Topic statement from top words, top hashtags and relevant tweet of topics

together, we can even find the reasons and after effects of an event or at least what and how people's reaction was about a specific event happened in real-life.

We did counter-check our results with the original tweets and information available on other platforms e.g. news articles, blog posts etc.

7 ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number 19K20333 and 16H02904.

REFERENCES

[1] David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*. ACM, 113–120.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

[3] Satoshi Fukuyama and Kei Wakabayashi. 2018. Extracting time series variation of topic popularity in microblogs. In *Proceedings of the 20th International Conference on Information Integration and Web-based Applications & Services*. ACM, 365–369.

[4] Bo Han, Paul Cook, and Timothy Baldwin. 2012. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*. Association for Computational Linguistics, 421–432.

[5] Liangjie Hong and Brian D Davison. 2010. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*. acm, 80–88.

[6] Amir Karami, Alicia A Dahl, Gabrielle Turner-McGrievy, Hadi Kharrazi, and George Shaw Jr. 2018. Characterizing diabetes, diet, exercise, and obesity comments on Twitter. *International Journal of Information Management* 38, 1 (2018), 1–6.

[7] Daichi Koike, Yusuke Takahashi, Takehito Utsuro, Masaharu Yoshioka, and Noriko Kando. 2013. Time series topic modeling and bursty topic detection of correlated news and twitter. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. 917–921.

[8] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web*. AcM, 591–600.

[9] Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. 2013. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 889–892.

[10] David Mimno, Matt Hoffman, and David Blei. 2012. Sparse stochastic inference for latent Dirichlet allocation. *arXiv preprint arXiv:1206.6425* (2012).

[11] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*. ACM, 851–860.

[12] Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Fourth international AAAI conference on weblogs and social media*.

[13] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing twitter and traditional media using topic models. In *European conference on information retrieval*. Springer, 338–349.