Published in final edited form as:
Proceedings (IEEE Int Conf Bioinformatics Biomed). 2019 November ; 2019: 1544–1547. doi:10.1109/bibm47256.2019.8983167.

Clustering and topic modeling over tweets: A comparison over a health dataset

Juan Antonio Lossio-Ventura[α], **Juandiego Morzan**[β], **Hugo Alatrista-Salas**[β], **Tina Hernandez-Boussard**[α], **Jiang Bian**[θ]

[α]Department of Medicine, Biomedical Informatics, Stanford University, USA

[β]School of Engineering, Universidad del Pacífico, Lima, Peru

[θ]Health Outcomes & Biomedical Informatics, University of Florida, USA

Abstract

Twitter became the most popular form of social interactions in the healthcare domain. Thus, various teams have evaluated Twitter as an additional source where patients share information about their healthcare with the potential goal to improve their outcomes. Several existing topic modeling and document clustering applications have been adapted to assess tweets showing that the performances of the applications are negatively affected due to the nature and characteristics of tweets. Moreover, Twitter health research has become difficult to measure because of the absence of comparisons between the existing applications. In this paper, we perform an evaluation based on internal indexes of different topic modeling and document clustering applications over two Twitter health-related datasets. Our results show that Online Twitter LDA and Gibbs LDA get a better performance for extracting topics and grouping tweets. We want to provide health practitioners this comparison to select the most suitable application for their tasks.

Index Terms—

topic modeling; clustering; internal cluster indexes; natural language processing; Twitter

I. Introduction

There are several online social networking platforms and services that allow parties to connect to share information, such as Facebook, Twitter, LinkedIn, YouTube, Instagram, among others. Twitter enables users to publish and read short messages, named “tweets” composed of 140 characters (now, 280 characters). Twitter users often share their opinions, feelings, thoughts, and personal activities. With over 500 million tweets posted each day, Twitter has become a very powerful data source to get real-world insights. In the health domain, Twitter has increasingly been adopted by users to share information and interact with other users with similar symptoms, disorders; attracting the attention of clinical researchers with the potential goal to improve patients’ outcomes [1]–[4]. Moreover, several studies have been demonstrated the use of Twitter as low-cost source for public health

jlossio@stanford.edu .

surveillance [5], such as for influenza vaccination [6], mental health [7], public mood [8], suicide [9], gender discrimination [10], etc.

These research works have focused on the design of natural language processing (NLP) methods to digest and analyze large amounts of text. Topic modeling and clustering are techniques among the proposed NLP methods, used to infer patients' interests, track new health-related stories, and identify emerging health topics. Clustering methods aim at grouping documents into clusters [11], [12]. They have different applications in information retrieval such as event detection, text summarization [13], [14]. Generally, the methods are based on representing text as a bag-of-words, and grouping texts on the basis of their lexical similarity. Topic modeling methods seek to extract topics from a set of text documents based on statistical techniques. Each topic is defined as a distribution over a set of words. Topic modeling and clustering have similar characteristics: both are based on unsupervised learning, they need a number of topics/clusters to be specified beforehand, and do not require labels. Also, a major problem in topic modeling and clustering methods is to determine the number of topics/clusters. Although many algorithms have been suggested to tackle the problem of determining the number of clusters, there does not appear to be a single method proven to be the most reliable, possibly due to the high complexity in real-world datasets. Thus, task-specific method for determining the number of clusters is always preferred, e.g., biomedical literature [15]. There are two kinds of cluster evaluation metrics which are called external and internal validation indexes. External indexes measure the quality based on already annotated datasets. Internal indexes evaluate the result on information intrinsic to the data alone. The latter is useful when there is no annotated dataset available. However, despite the abundance of NLP techniques available in the literature, there are several challenges when it comes to the analysis of tweets due to its noisy nature and inconsistent user reliability. This prevents the tweets from being employed to their full potential. Moreover, Twitter health research has become difficult to measure because of the absence of comparisons between the existing applications.

To the best of our knowledge, various studies have been devoted to content analysis of health-related tweets, however, none has carried out a deep content comparison of topic modeling and clustering methods over health datasets. In this paper, we want to address the problem of how effectively several standard topic modeling and clustering methods perform on health-related tweets. We test and compare several state-of-the-art applications on an unbalanced dataset composed of two subsets: Human Papillomavirus (94.6%) and Lynch Syndrome (5.4%). Our experiments are validated based on internal evaluation indexes due to the lack of available annotated datasets.

II. Methods

A. Tweets collection

Our health dataset is composed of two subsets: the human papillomavirus (HPV) and the lynch syndrome tweets. The extraction strategy considered keywords and hashtags containing common generic HPV and lynch syndrome names and slang terms. Table I shows a description of our collection. Our tweets collection are composed of 140 characters.

We applied several rules to preprocess the tweets collection: 1) text was changed to lowercase; 2) suppression of repeated tweets; 3) suppression of stop-words; and 4) omission of links from the tweets.

B. Applications

1. *Topic modeling*: we set up six well-known available methods used for short texts: (i) Latent Semantic Indexing (LSI)[1] [16], (ii) Latent Dirichlet Allocation (LDA)[2] [17], (iii) LDA with Gibbs Sampling (GibbsLDA)[3] [18], (iv) Online LDA[4] [19], (v) Biterm (BTM)[5] [20], and (vi) Online Twitter LDA[6] [21].
2. *Clustering*: we used *k*-means as algorithm on two different dataset representations: (i) TFIDF representation [22] and (ii) Doc2Vec[7] [23].

C. Analysis of applications

1. *Configuration*: for topic modeling, "*k*" (i.e., number of topics) will range from 2 to 50. In our work, topic modeling results are used to classify tweets to a particular topic. Each tweet is represented by a feature vector, where each component of the vector is the probability of the tweet to belong to a given topic. For instance, k=2 means the size of the feature vector is 2; for k=50 is 50. We then use an *argmax* function to determine the most prominent topic of each tweet. The clustering algorithm, *K-means*, uses two document representations: TFIDF and Doc2Vec. Both set the number of features (bag-of-words) equal to 100 for comparison purposes, with a "*k*" (i.e., number of clusters) also ranging from 2 to 50. Note that "*k*" is indistinctively used as number of clusters and topics.
2. *Evaluation*: we evaluated all topic modeling and clustering algorithms using 100, 500, and 1,000 iterations. The initial number of iterations is recommended in [24] and is a default value in the applications. To evaluate the performance of the topic modeling and clustering methods, we have employed two internal validity indexes: Calinski-Harabasz index (CH) [25] and Silhouette Coefficient (SC) [26]. Calinski and Harabasz index has demonstrated in several works to be an effective measure for determining the most appropriate number of clusters [27]. On the other hand, Silhouette Coefficient is one of the most well-known measures and one of the fewest measures independent from the number of clusters. In the next paragraphs we explain the principles of the internal indexes.

Calinski-Harabasz index: Calinski-Harabasz index: also known as the Variance Ratio Criterion, it can be used to evaluate the clustering model, where a higher CH value relates to a model with better defined clusters. The CH_k value is given by the ratio between average

1 https://radimrehurek.com/gensim/models/lsimodel.html
2 https://radimrehurek.com/gensim/models/ldamodel.html
3 https://nlp.stanford.edu/software/tmt/tmt-0.4/
4 https://radimrehurek.com/gensim/models/ldamulticore.html
5 https://github.com/xiaohuiyan/BTM
6 https://github.com/jhlau/online_twitter_lda
7 https://radimrehurek.com/gensim/models/doc2vec.html

inter-cluster dispersion matrix (B_k) and intra-cluster dispersion matrix (W_k) as defined in Formula 1.

$$\mathbf{CH_k} = \frac{\mathrm{B_k}}{\mathrm{W_k}} \times \frac{n-k}{k-1} \tag{1}$$

where n is the total number of points and k the number of clusters. The B_k value is based on the distance between clusters and is defined as:

$$\mathbf{B_k} = \sum_{i}^{k} n_i \cdot dist^2(c_i - c)$$

where n_i is the number of elements of cluster C_i, c_i is the center of C_i, and c is the center of the complete dataset. W_k is based on the distance within clusters and is defined as:

$$\mathbf{W_k} = \sum_{i=1}^{k} \sum_{x \in C_i} dist^2(c_i, x)$$

where x is a point of cluster C_i. Note that to obtain well separated and compact clusters, B_k is maximized and W_k minimized. Therefore, the maximum value of CH indicates a suitable partition for the dataset.

Silhouette Coefficient: Silhouette Coefficient: studies the separation distance between the resulting clusters. SC computes for each point a width depending on its membership in any cluster. This silhouette width is then an average over all observations. SC value has a range of [−1, 1], where −1 represents poor clustering quality or poorly defined clusters and 1 high clustering quality or well-defined clusters. The SC_k value for a single sample is defined in Formula 2.

$$\mathbf{SC_k} = \frac{1}{n} \times \sum_{i}^{n} \frac{b_i - a_i}{max(a_i, b_i)} \tag{2}$$

where n represents the total number of elements in a cluster, a_i is the average distance between an element i of the cluster and all other elements within the same cluster, b_i represents the average distance between the element i of the cluster and all other elements in the nearest cluster.

In summary, higher clustering quality of a particular algorithm tends to yield higher predictive performance on information retrieval tasks. For this reason, we seek to identify the algorithms that maximize the overall clustering quality (i.e., internal indexes).

III. Experiments and results

As we have stated, the focus of this study is to compare the performance of applications using the internal indexes CH and SC, over the content of Twitter. In this section we show the results obtained for k={2,5,10,50}. CH and SC quantify the performance of a clustering algorithm based on two aspects: the similarity of tweets within the same cluster (cohesion), and the difference between the tweets of different clusters. Tables II, III, IV and V show the CH and SC results of each method for 2, 5, 10, and 50 number of clusters/topics (*"k"*) respectively. In all cases, the best values are obtained by Online Twitter LDA followed by GibbsLDA.

IV. Discussion

Although the dataset is an unbalanced corpus, the results suggest that Online Twitter LDA followed by GibbsLDA characterize well the tweets in topics. Therefore, the clusters formed are more compact in terms of CH and SC, since they have a higher density (within the cluster) and a greater degree of separation. Also, variations of LDA perform better since they are improvements based on LDA. Note that when the number of topics increases, the metrics obtained tend to decrease as also shown in Fig 1. The reason is that our tweet collection is composed of two subsets: lynch syndrome and HPV tweets. Thus, two topics are quite marked and differentiated. Therefore, it is reasonable that clusters will be denser and more defined for a smaller k (which adheres to the nature of the dataset).

On the other hand, clustering results are lower than topic modeling. Hence, topic modeling algorithms might be providing more interesting and sophisticated insights than the single vectorial representation of tweets. Also, clustering with Doc2Vec representation has better results than TFIDF for smaller k. Nevertheless, as the number of k increases, this behavior is reversed, and TFIDF shows better metrics.

Finally, we can see the greater number of iterations in the experiments the better results obtained of the internal indexes. Therefore, the results obtained with the experiments of 1,000 iterations are usually better, and especially with topic modeling that are trained and perform better as the number of iterations increases.

V. Conclusions

In this paper, we conducted a deep comparison of different topic modeling and document clustering applications on a Twitter health-related dataset composed of two subsets: HPV and lynch syndrome tweets. We set up LSI, LDA, LDA with Gibbs Sampling, Online LDA, Biterm, Online Twitter LDA, and K-means based on TFIDF and Doc2Vec document vectorizations. They were evaluated considering two internal indexes: Calinski-Harabasz index and Silhouette Coefficient. The best results were obtained by Online Twitter LDA, which was able to group better the tweets in the extracted topics. Overall, this comparison provides encouraging results towards the application of topic modeling over health-related tweets.

As future work, we plan to complete our evaluation with external indexes. Currently, we are computing the Adjusted Rand index, Normalized Mutual Information, Homogeneity index, Completeness, and V-measure over the same dataset. Our ultimate goal is to do a complete evaluation of the available applications and identify whether they are able to answer healthcare questions such as: most discussed health topic in a tweet collection, the interaction of healthcare professionals with patients, most discussed topic of patients, among others.

Acknowledgment

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Number R01CA183962. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1]. Sinnenberg L, Buttenheim AM, Padrez K, Mancheno C, Ungar L, and Merchant RM, “Twitter as a tool for health research: a systematic review,” American J. of public health, vol. 107, no. 1, pp. e1–e8, 2017.

[2]. De Martino I, DApolito R, McLawhorn AS, Fehring KA, Sculco PK, and Gasparini G, “Social media for patients: benefits and drawbacks,” Current reviews in musculoskeletal medicine, vol. 10, no. 1, pp. 141–145, 2017. [PubMed: 28110391]

[3]. Zhang L, Hall M, and Bastola D, “Utilizing twitter data for analysis of chemotherapy,” International journal of medical informatics, vol. 120, pp. 92–100, 2018. [PubMed: 30409350]

[4]. Vraga EK, Stefanidis A, Lamprianidis G, Croitoru A, Crooks AT, Delamater PL, Pfoser D, Radzikowski JR, and Jacobsen KH, “Cancer and social media: A comparison of traffic about breast cancer, prostate cancer, and other reproductive cancers on twitter and instagram,” Journal of health communication, vol. 23, no. 2, pp. 181–189, 2018. [PubMed: 29313761]

[5]. Paul MJ and Dredze M, “Social monitoring for public health,” Synthesis Lectures on Information Concepts, Retrieval, and Services, vol. 9, no. 5, pp. 1–183, 2017.

[6]. Huang X, Smith MC, Jamison AM, Broniatowski DA, Dredze M, Quinn SC, Cai J, and Paul MJ, “Can online self-reports assist in real-time identification of influenza vaccination uptake? a cross-sectional study of influenza vaccine-related tweets in the usa, 2013–2017,” BMJ open, vol. 9, no. 1, p. e024018, 2019.

[7]. Coppersmith G, Harman C, and Dredze M, “Measuring post traumatic stress disorder in twitter,” in Proceedings of the AAAI Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1–4, 2014., 2014.

[8]. Ofoghi B, Mann M, and Verspoor K, “Towards early discovery of salient health threats: A social media emotion classification technique,” in Biocomputing 2016: Proceedings of the Pacific Symposium. World Scientific, 2016, pp. 504–515.

[9]. Braithwaite SR, Giraud-Carrier C, West J, Barnes MD, and Hanson CL, “Validating machine learning algorithms for twitter data against established measures of suicidality,” JMIR mental health, vol. 3, no. 2, p. e21, 2016. [PubMed: 27185366]

[10]. Alatrista-Salas H, Hidalgo-Leon P, and Nunez-del Prado M, “Documents retrieval for qualitative research: Gender discrimination analysis,” in 2018 IEEE Latin American Conference on Computational Intelligence (LA-CCI). IEEE, 2018, pp. 1–6.

[11]. Lossio Ventura JA, Hacid H, Ansiaux A, and Maag ML, “Conversations reconstruction in the social web,” in Proceedings of the 21st International Conference on World Wide Web, ser. WWW ‘12 Companion. New York, NY, USA: ACM, 2012, pp. 573–574.

[12]. Lu Y, Zhang P, Liu J, Li J, and Deng S, “Health-related hot topic detection in online communities using text clustering,” Plos one, vol. 8, no. 2, p. e56221, 2013. [PubMed: 23457530]

[13]. Lossio-Ventura JA, Hacid H, Roche M, and Poncelet P, “Communication overload management through social interactions clustering,” in Proceedings of the 31st Annual ACM Symposium on Applied Computing, ser. SAC ‘16. New York, NY, USA: ACM, 2016, pp. 1166–1169.

[14]. Yin J, Chao D, Liu Z, Zhang W, Yu X, and Wang J, “Model-based clustering of short text streams,” in Proc of the 24th ACM SIGKDD Int Conference on Knowledge Discovery & Data Mining, ser. KDD ‘18. New York, NY, USA: ACM, 2018, pp. 2634–2642.

[15]. Lossio-Ventura JA, Bian J, Jonquet C, Roche M, and Teisseire M, “A novel framework for biomedical entity sense induction,” Journal of biomedical informatics, vol. 84, pp. 31–41, 2018. [PubMed: 29935347]

[16]. Deerwester S, Dumais ST, Furnas GW, Landauer TK, and Harshman R, “Indexing by latent semantic analysis,” J. of the American society for information science, vol. 41, no. 6, pp. 391–407, 1990.

[17]. Blei DM, Ng AY, and Jordan MI, “Latent dirichlet allocation,” J. of machine Learning research, vol. 3, no. Jan, pp. 993–1022, 2003.

[18]. Wei X and Croft WB, “Lda-based document models for ad-hoc retrieval,” in Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006, pp. 178–185.

[19]. Hoffman M, Bach FR, and Blei DM, “Online learning for latent dirichlet allocation,” in advances in neural information processing systems, 2010, pp. 856–864.

[20]. Yan X, Guo J, Lan Y, and Cheng X, “A biterm topic model for short texts,” in Proc of the 22nd Int Conference on World Wide Web, ser. WWW ‘13. New York, NY, USA: ACM, 2013, pp. 1445–1456.

[21]. Lau JH, Collier N, and Baldwin T, “On-line trend analysis with topic models:# twitter trends detection topic model online,” Proceedings of COLING 2012, pp. 1519–1534, 2012.

[22]. Salton G and Buckley C, “Term-weighting approaches in automatic text retrieval,” Information processing & management, vol. 24, no. 5, pp. 513–523, 1988.

[23]. Le Q and Mikolov T, “Distributed representations of sentences and documents,” in International conference on machine learning, 2014, pp. 1188–1196.

[24]. Halko N, Martinsson P-G, and Tropp JA, “Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions,” SIAM review, vol. 53, no. 2, pp. 217–288, 2011.

[25]. Caliński T and Harabasz J, “A dendrite method for cluster analysis,” Communications in Statistics-theory and Methods, vol. 3, no. 1, pp. 1–27, 1974.

[26]. Rousseeuw PJ, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” Journal of computational and applied mathematics, vol. 20, pp. 53–65, 1987.

[27]. Anderson MJ, “A new method for non-parametric multivariate analysis of variance,” Austral ecology, vol. 26, no. 1, pp. 32–46, 2001.

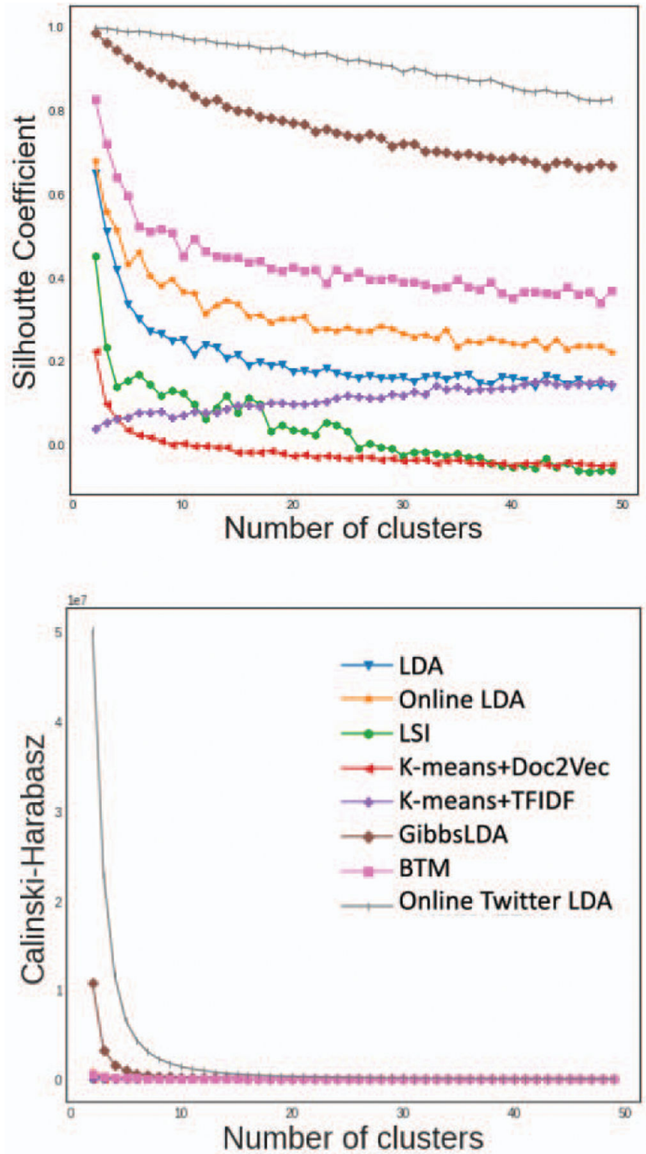

Fig. 1:
Silhouette Coefficient and Calinski-Harabasz metrics with 100 iterations, for "k" ranging from 2 to 50.

Author Manuscript

Author Manuscript

Author Manuscript

TABLE I:

Details of our health-related tweets.

Subset	HPV	Lynch syndrome
No. of tweets	271,533	15,438
No. of users	99,227	4,492
Collection period	Jan 2014 – Mar 2016	Oct 2016 – Nov 2017
No. of unique hashtags	14,875	1,649
No. of tweets with hashtag	115,859	10,224
No. of tokens before preprocessing	1,767,920	147,144
No. of tokens after preprocessing	1,042,063	96,437

TABLE II:

Internal index results for *k*=2.

	100 iterations		***500 iterations***		***1,000 iterations***	
	CH	***SC***	***CH***	***SC***	***CH***	***SC***
LSI	86,260	0.41	86,260	0.40	86,260	0.40
BTM	634,412	0.74	661,242	0.75	604,629	0.72
LDA	515,737	0.68	486,522	0.68	521,198	0.69
GibbsLDA	**10,767,060**	**0.97**	**10,514,730**	**0.98**	**9,932,722**	**0.97**
Online LDA	849,068	0.77	834,351	0.76	938,428	0.78
Online Twitter LDA	**50,110,500**	**0.99**	**53,291,260**	**0.99**	**50,730,260**	**0.99**
K-means+Doc2Vec	31,196	0.20	31,196	0.20	31,196	0.20
K-means+TFIDF	5,764	0.04	5,764	0.04	5,764	0.04

Author Manuscript

Author Manuscript

Author Manuscript

TABLE III:

Internal index results for k=5.

	100 iterations		500 iterations		1,000 iterations	
	CH	**SC**	**CH**	**SC**	**CH**	**SC**
LSI	50,641	0.40	51,961	0.40	51,961	0.40
BTM	165,515	0.60	171,041	0.60	175,937	0.61
LDA	69,526	0.37	71,240	0.38	71,741	0.38
GibbsLDA	**967,683**	**0.91**	**1,016,640**	**0.93**	**1,010,773**	**0.92**
Online LDA	173,255	0.61	170,659	0.60	185,005	0.62
Online Twitter LDA	**6,554,339**	**0.98**	**10,117,270**	**0.98**	**10,989,530**	**0.99**
K-means+Doc2Vec	13,998	0.04	13,998	0.04	13,998	0.04
K-means+TFIDF	4,722	0.07	4,722	0.07	4,722	0.07

TABLE IV:

Internal index results for k=10.

	100 iterations		500 iterations		1,000 iterations	
	CH	**SC**	**CH**	**SC**	**CH**	**SC**
LSI	25,346	0.34	25,539	0.34	25,532	0.35
BTM	55,110	0.41	61,790	0.43	67,856	0.53
LDA	24,498	0.27	24,102	0.27	24,860	0.27
GibbsLDA	**239,457**	**0.83**	**266,065**	**0.85**	**272,035**	**0.86**
Online LDA	70,824	0.54	69,418	0.55	69,892	0.55
Online Twitter LDA	**1,400,045**	**0.96**	**1,925,903**	**0.97**	**2,035,547**	**0.97**
K-means+Doc2Vec	7,617	0.02	7,617	0.02	7,617	0.02
K-means+TFIDF	3,758	0.07	3,758	0.07	3,758	0.07

TABLE V:

Internal index results for *k*=50.

	100 iterations		500 iterations		1,000 iterations	
	CH	**SC**	**CH**	**SC**	**CH**	**SC**
LSI	3,894	0.21	3,925	0.22	3,907	0.19
BTM	9,501	0.35	10,089	0.37	10,507	0.37
LDA	3,006	0.14	2,801	0.12	2,960	0.14
GibbsLDA	**18,188**	**0.65**	**21,256**	**0.68**	**22,014**	**0.69**
Online LDA	9,835	0.44	10,322	0.45	10,299	0.46
Online Twitter LDA	**39,014**	**0.79**	**62,749**	**0.85**	**66,051**	**0.87**
K-means+Doc2Vec	2,028	−0.02	2,028	−0.02	2,028	−0.02
K-means+TFIDF	2,200	0.17	2,200	0.17	2,200	0.17