How people talk about health? Detecting Health Topics from Twitter Streams

Carmela Comito
National Research Council of Italy
Institute for High Performance Computing and Networking
{carmela.comito}@icar.cnr.it

Clara Pizzuti
National Research Council of Italy
Institute for High Performance Computing and Networking
{clara.pizzuti}@icar.cnr.it

Nicola Procopio
National Research Council of Italy
Institute for High Performance Computing and Networking
{nicola.procopio}@icar.cnr.it

ABSTRACT

The paper proposes an online clustering algorithm for detecting health-related topics. The method extracts from the tweets relevant terms and incrementally groups them by taking into account both term occurrences and tweet age. A detailed experimentation on the tweets posted by users in US shows that the method is capable to group tweets addressing common health issues into the pertinent topic, outperforming traditional topic model approaches, like Doc-p and LDA.

CCS Concepts

•Information systems → Clustering; Data stream mining;

Keywords

Twitter, Topic Detection, e-Health

1. INTRODUCTION

Social media, in the last years, constitute an important data source for analyzing and extracting in real-time information related to events occurring in the world. In the public health area, especially, physician and patient communities could take a great advantage since the available huge data can be gathered faster and at a lower cost, compared to the traditional sources, mainly surveys.

Recently, social network data has been explored to monitor and analyze health issues with applications in disease surveillance and epidemiological studies. By far the first and most common healthcare application in social media is influenza. Seminal works like [1, 2, 3, 4, 5, 6] have shown that the tweets can be used to track and predict influenza and detect depression [7, 8]. To this purpose, a variety of techniques have been proposed: starting from capturing the overall trend of a particular disease outbreak by monitoring social media [1, 2], many other approaches appeared such as the ones based on linear regression [3], supervised machine learning [5, 6] and social network analysis [4].

In this context, an important problem is how to effectively detect health-related topics within Twitter's data. Public health-related topics are difficult to identify in large conversational datasets like Twitter. Traditional topic modeling methods like *LDA* [9] and *Doc-p* [10] have implemented topic modeling to Twitter data. LDA [9] is a topic model that relates words and documents through latent topics. It associates with each document a probability distribution over topics, which are distributions over words. The topic and word distributions have to be estimated by using Bayesian Inference. *Doc-p* [10] is an on-line clustering method that, for each document, considers the list of words, and computes the cosine similarity of the *tf-idf* representation of a new text with respect to all those already examined. If the best similarity is above a threshold θ, then the story is assigned to the cluster with the best match, otherwise a new cluster is generated.

The main limitation of such approaches is that they heavily depend on the frequency distribution of words to generate topics, failing to detect low frequency topics that instead are common in the faced scenario. In fact, public health-related topics and discussions use less frequent words, and are therefore more difficult to identify using traditional topic modeling. Dedicated health topic modeling approaches, such as the Ailment Topic Aspect Model (ATAM), have thus been proposed to discover ailments from tweets [6]. (ATAM) is an LDA-like technique able to detect the symptoms and possible treatments for ailments that people discuss on Twitter. It implements a machine learning based classification algorithm used for identifying health-related tweets by exploiting n-gram word features to train a linear kernel SVM classifier. However, ATAM still suffers of the same LDA limitations, failing to detect topics with low frequency.

In this work, we propose an incremental clustering approach for detecting ailments and health topics from tweets that, differently from traditional topic modeling approaches, is able to detect low-frequency topics of discussion since the grouping of tweets in topics is not based on word frequencies, but on the similarity of the words and hashtags used in the tweets. We do this by designing an online clustering algorithm which automatically infers interesting patterns and regularities in the tweets text by leveraging on the frequency of lexically similar terms and temporal closeness of the tweets. The information extracted from similar tweets, that is those dealing with a common health subject,

BDIOT' 17, December 20-22, 2017, London, UK

ISBN 978-1-4503-5430-1...$15.00

DOI: http://dx.doi.org/10.1145/12345.67890

is summed up into the centroid of the cluster they have been assigned to. We refer to clusters as topics and to centroids as topic summary. A topic summary consists of the terms contained in the already analyzed tweets and assigned to a particular cluster, along with the times of cluster generation and last updating. In order to deal with the high streaming rate, we use a representation of tweets and cluster centroids that dynamically builds the term vocabulary.

A thorough experimentation on the tweets posted by users in US during the period September 2015 - April 2016 shows that the method is capable to group tweets addressing common health issues into the pertinent topic, outperforming traditional topic model approaches like Doc-p and LDA.

The paper is organized as follows. Section 2 introduces the concepts used to model the clustering problem of tweet streams. Section 3 reports the experimental results. Finally Section 4 concludes the paper.

2. METHODS

In this paper, we address the problem of how to detect health-related topics within social media, especially Twitter. We aim to identify which are the health topics discussed on Twitter and how people talk about them.

We formulate the problem as an incremental clustering algorithm that sequentially analyzes the tweet streams and clusters together similar tweets. Each cluster is considered as a topic of discussion.

The proposed algorithm, referred as HealthS-Tweet (Health Surveillance through Twitter), is a specialized version for the healthcare domain of a previous work we presented in [11]. In the following we describe the data model and the rationale of the algorithm HealthS-Tweet.

We represent a tweet tw posted by a user u at time t from a location l, as a tuple, $tw = (id, u, l, t, fv)$ where id is the tweet identifier and $fv = (w_u, w_b, h_u, h_b)$ is a feature vector that summarizes its content with the list of words (w_u) and hashtags (h_u), along with the corresponding bigrams (w_b and h_b), i.e. the adjacent item pairs appearing in the tweet text.

The objective of the proposed approach is to group the tweets about the same topic into a cluster, referred as a *health topic*, HT. The main element of a health topic HT is the topic summary S (it is actually the cluster centroid). The summary S of a health topic HT has a representation similar to that of a tweet, extended with term occurrences to take into account the tweets examined so far and assigned to HT. Specifically, the summary is a tuple $S = (ht, t_0, t_c, fv_T)$, where ht is the topic label, t_0 is the creation time of the topic, t_c is the time stamp of the last time a tweet was added to ht, and $fv_T = (fv, ff)$. fv is the feature vector analogous to the tweet feature vector, while $ff = (f_{w_u}, f_{w_b}, f_{h_u}, f_{h_b})$ is the list of frequencies corresponding to the features.

The rationale of the HealthS-Tweet algorithm is as follows. At the beginning, the tweet tw_1 is considered the first topic (cluster) HT_1, and the centroid coincides with the feature vector of the tweet, where the frequencies are set to 1 and the time stamp is the same of tw_1. As a new tweet tw_i appears at time t_i, the algorithm selects the features from tw_i and computes the similarity between the tweet and the existing topics at the time stamp t_i. If HT_c is the cluster whose centroid has maximum similarity with tw_i, and this similarity value is higher than the fixed threshold ϵ, the content of the tweet is added to HT_c by updating the summary. Otherwise a new health topic is generated from tw_i. These steps are repeated until the stream of tweets continues.

An important step of the algorithm is the similarity function used to assign a tweet to a topic. The function we adopt has been proposed in [11] and it is based on the frequency of lexically similar terms and the temporal closeness of the tweets. For each word and hashtag, both unigram and bigram of a tweet t_w, the intersection with the terms in the summary feature vector and their union are computed by summing up the frequencies of the terms appearing in the intersection and union, respectively. The ratio between the intersection and the union is then multiplied by a temporal factor to bias the similarity function towards health topics temporally closer to the tweet t_w.

3. EXPERIMENTAL STUDY

In this section, we show the effectiveness of HealthS-Tweet by running the method on a set of health tweets. Specifically, we first evaluate the ability of the algorithm in detecting health-related topics identifying the best parameter setting. After that, we perform a temporal analysis to detect the prevalence of ailments over time. We also analyze trends of individual terms, by counting the number of tweets containing a term, normalized by the total number of tweets in the dataset from that time period.

Finally, we compare HealthS-Tweet with three state-of-the-art methods, *LDA* [9], *Doc-p* [10], and *SFPM* [12], on a dataset of tweets described in the next section.

3.1 Twitter dataset

The dataset used for the experimental evaluation comes from the data collected in the context of a study about influenza surveillance in US [13]. The data consists of about 10,000 tweets posted in the period September 2015 - April 2016 in US. This dataset of tweets has been obtained by first filtering tweets having health-related keywords (including flu-related words) using the Twitter streaming API, and then using Amazon Mechanical Turk, a crowdsourcing service, [14] to distinguish relevant health tweets from spurious ones.

Figure 1 shows the total number of tweets per topic for the top 10 topics, expressed as percentage on the overall number of tweets. As can be noted from the graph, there is a huge prevalence of flu and influenza terms. In fact, in the dataset there is only a small percentage of other ailments not reported in the graph. We highlight that we distinguish flu and influenza, even if the two terms are used interchangeably for the same disease.

3.2 Evaluation of HealthS-Tweet

In a first set of experiments we aimed to identify the best setting of parameters for the HealthS-Tweet algorithm. Specifically, an important algorithm parameter is the similarity threshold, ϵ. To this purpose we performed a set of experiments by using different similarity threshold values. Table 1 summarizes the differences in terms of number of topics detected, number of singletons and average size of the clusters for different ϵ values. The results reveal that the number of obtained clusters increases with the ϵ values since the higher the similarity threshold ϵ, the lower the probability that the lexicon of the tweets meets the similarity threshold.

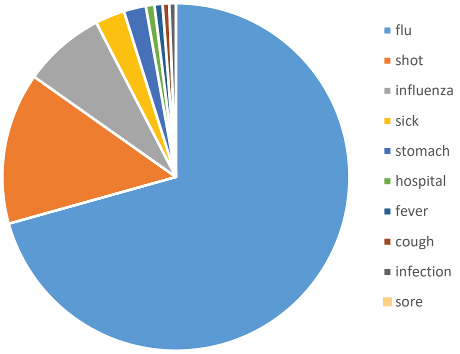


Figure 1: Percentage of tweets per topics.

Accordingly, the number of singleton increases while the average size of clusters decreases. The best value of similarity function resulted to be 0.3 because of the lowest number of produced singletons and the significant average size of the clusters. With high similarity thresholds the algorithm produces too many small clusters, mostly for the more popular topics, like influenza and flu, as it is confirmed by the average small size of the produced clusters, around 3 for both 0.5 and 0.8 similarity thresholds.

To have a more detailed view of the topics identified with different threshold similarity values, Figure 2 shows the number of topics detected for disease by using different similarity threshold values. The detected topics range from more seasonal diseases like flu, influenza, allergy and consequent remedies, like vaccine, to common diseases, like cancer, to rare ones linked to exceptional events, like bird flu and dengue fever.

ϵ	Number of clusters	Average size	Number of singletons
0.3	1181	12.88	686
0.5	5576	3.86	5057
0.8	6137	3.49	5764

Table 1: Evaluation of HealthS-Tweet for different ϵ.

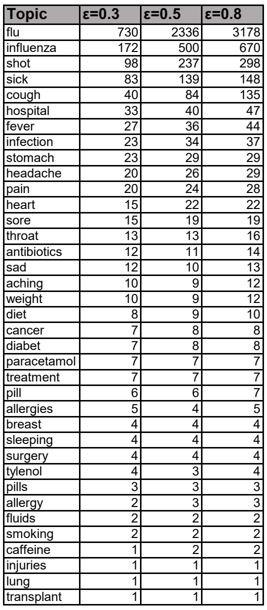
3.2.1 *Quality performance evaluation*

We used two performance indices to evaluate the quality of the clustering obtained by HealthS-Tweet with different similarity values, the entropy and the silhouette metrics, as detailed in the following.

The entropy of a cluster allows to understand cluster homogeneity by means of the term distribution across the tweets. The entropy value decreases when tweets share a similar vocabulary, whereas it increases when the vocabulary varies among tweets grouped in the cluster. Thus it gives information on how people discuss about a topic, by checking whether people talk about a topic by using similar terms. We computed the entropy for each feature f in the feature vector fv_T of the summary S of each health topic HT. Specifically, we computed the variation of the feature throughout the tweets in the cluster. The higher the entropy, the more different the feature from tweet to tweet within the cluster. Entropy is computed as:

$$H(f)_{HT} = -\sum_{i=1}^{|f|} p_i log\ p_i, \quad p_i = \frac{f_{f_i}}{N} \tag{1}$$

where $H(f)_{HT}$ is the Shannon's entropy of feature f for the topic HT, f_{f_i} is the size of the value i of feature f (in other words its frequency), $|f|$ is the number of distinct values, N is the total size of feature f (the sum of the frequencies of the $|f|$ different values of the feature), and p_i is the observed probability of the value i.



Topic	ε=0.3	ε=0.5	ε=0.8
flu	730	2336	3178
influenza	172	500	670
shot	98	237	298
sick	83	139	148
cough	40	84	135
hospital	33	40	47
fever	27	36	44
infection	23	34	37
stomach	23	29	29
headache	20	26	29
pain	20	24	28
heart	15	22	22
sore	15	19	19
throat	13	13	16
antibiotics	12	11	14
sad	12	10	13
aching	10	9	12
weight	10	9	12
diet	8	9	10
cancer	7	8	8
diabet	7	8	8
paracetamol	7	7	7
treatment	7	7	7
pill	6	6	7
allergies	5	4	5
breast	4	4	4
sleeping	4	4	4
surgery	4	4	4
tylenol	4	3	4
pills	3	3	3
allergy	2	3	3
fluids	2	2	2
smoking	2	2	2
caffeine	1	2	2
injuries	1	1	1
lung	1	1	1
transplant	1	1	1

Figure 2: Number of produced topics for disease with different ϵ values.

Figure 3 shows that the entropy of the clusters obtained with HealthS-Tweet algorithm is rather low and smoothly decreases with the similarity threshold, proving the good homogeneity of the clusters produced by the proposed algorithm. This is clearly more evident for higher similarity values.

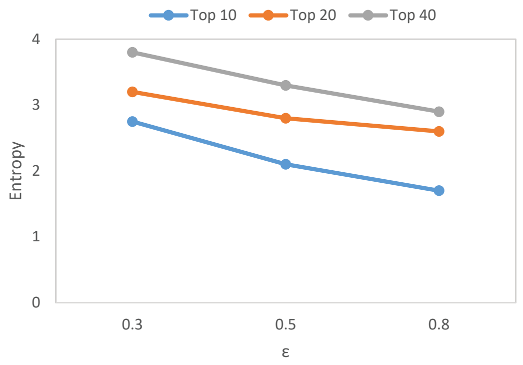


Figure 3: Entropy w.r.t ϵ for different top K topics.

Another important characteristic of a good clustering algorithm is the separability. To this aim we used the silhouette index, which is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette value ranges from -1 to 1. A high value indicates that the object is very similar to the other objects of its own cluster and it is poorly alike to objects of neighboring clusters. A good clustering is one where most objects have a high silhouette value.

The results achieved by the algorithm when using a similarity threshold of 0.3 are shown in Figure 4. We notice that the separability of the clusters is really good: the silhouette value is quite high and remains rather stable with the top k detected topics, whereas it is very low for $\epsilon = 0.5$ and $\epsilon = 0.8$.

We can conclude that, even if the threshold $\epsilon = 0.3$ achieved an entropy that is slightly higher than the other values, it represents the optimal value for the used dataset since with this threshold the algorithm achieved the highest separability and it is able to identify significant topics.

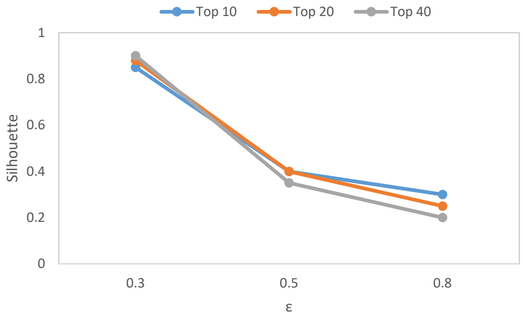


Figure 4: Silhouette w.r.t ϵ for different top K topics.

3.2.2 *Temporal Analysis*

In this section we aim to identify patterns and regularities for the discovered topics, e.g., analyze the evolution of disease over time or the prevalence of a disease in a specific time period. For the analysis we chose topics that present seasonal regularities like flu, influenza and flu shot, and topics that do not a-priori exhibit any regularities (e.g., cancer). For each week, we computed the number of tweets containing the words flu and influenza, and the number of detected topics containing the same words. We refer the former as weekly tweet rate and the latter as weekly topic rate.

Clearly, we expect that the trend of the weekly flu/influenza topic rate follows the weekly flu/influenza tweet rate. To this purpose we use as baseline models the data on flu and influenza illness like provided by the American Centers for Disease Control and Prevention (CDC) [15], and the number of tweets with the words influenza and flu, respectively (see Figure 5).

As reported on the CDC website, the timing of flu can vary in different parts of the country and from season to season. Most of the time flu activity peaks between December and February. The peak week of flu activity in terms of influenza-like illness (ILI) for the 2015-2016 season was the week ending March 12, 2016 (referred to as week 10). Figure 5(a) reports the CDC flu data exhibiting a peak at week 10. Accordingly, Figures 5(b) and 5(c) show a peak in the number of tweets about flu and influenza within the range of week 10, precisely week 9 for flu and week 11 for influenza.

Figures 6(a) and 6(b) show the weekly topic rate of flu and influenza that, in accordance to the CDC report and the tweets trend of Figure 5, present a peak at week 9 and 11, respectively. Both trends then have a deep at week 16 and another peak at week 21. While the first peak is in correspondence of the peak of the seasonal flu, the one at week 21 is related to the *bird flu* outbreak. Figure 7(a) shows that also the flu shot topic rate exhibits the same peaks at week 9 and 21 even if it has other peaks in between the two.

We also analysed the temporal pattern of a non seasonal ailment like the cancer. However, as shown in Figure 7(b), it exhibits a trend that follows that of flu shot since many topics started on this subject because people were concerned about the fact that vaccine could cause cancer.

3.3 Comparison with related approaches

In this section we compare the proposed approach with the three methods *LDA* [9], *Doc-p* [10], and *SFPM* [12].

SFPM, *Soft Frequent Pattern Mining*, [12] is a soft version of the well known frequent pattern mining approach that finds frequent patterns in association rules by taking into account the simultaneous co-occurrences between any number of terms.

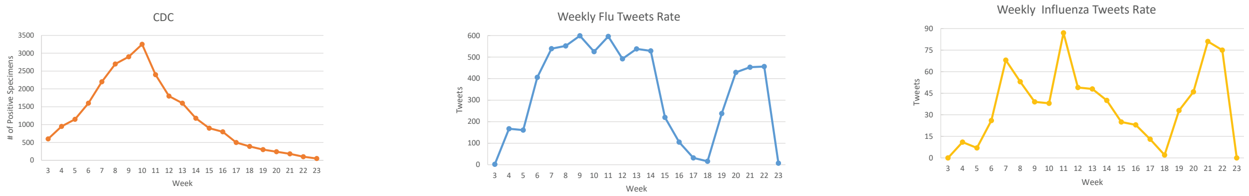
The software implementing these methods has been provided by Aiello et al. [12]. We executed them by using the following parameter setting as it yields the best result. *LDA* needs the expected number of clusters and keywords, which have been fixed to 200 and 15, respectively. The input for *Doc-p* is the similarity threshold, set to 0.8. As regards *SFPM* we maintained the parameters set by the authors.

To confirm the good performance of HealthS-Tweet, we compared it against the contestant algorithms also in terms of quality performance metrics. Specifically, the clustering produced by the different approaches are compared by using the quality indices introduced in the previous section. To explore the variation of the performance when more topics are produced, we studied the performance metrics as the number K of top results considered varies.

Figure 8(a) shows that the average cluster homogeneity, measured in terms of entropy, decreases with the number of top produced topics for all the approaches. However, HealthS-Tweet exhibits, together with LDA, the best results since the decrease is rather narrowed; in fact, as can be seen on the graph, the entropy increases shortly when K augments. We remind that a small entropy corresponds to a high cluster coherence, thus an increase of entropy results in coherence decrease.

Figure 8(b) shows that cluster separability remains pretty constant with the number of top detected topics, for all the algorithms. Doc-p, LDA and HealthS-Tweet achieved very high silhouette values and perform very similarly. However, HealthS-Tweet achieved the best separability when the top detected topics increased.

The table in Figure 9 shows the number of topics detected by each method for different ailments. In this case only the topics with a size greater than 10 are considered. As can be noted, HealthS-Tweet largely outperforms contestant methods in the number of detected topics. This is even more evident for the topics concerning flu and influenza. One can also note that contestant approaches are not able to detect topics regarding less frequent tweeted ailments, like allergy and diabetes.

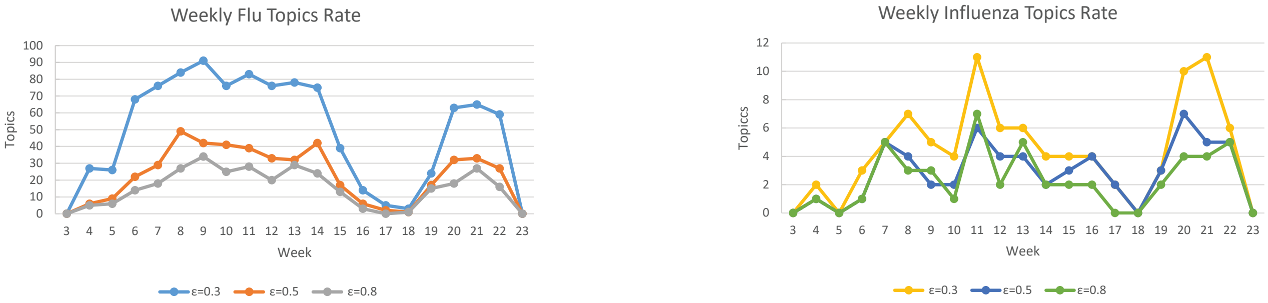


(a) CDC data

(b) Flu Tweets

(c) Influenza Tweets

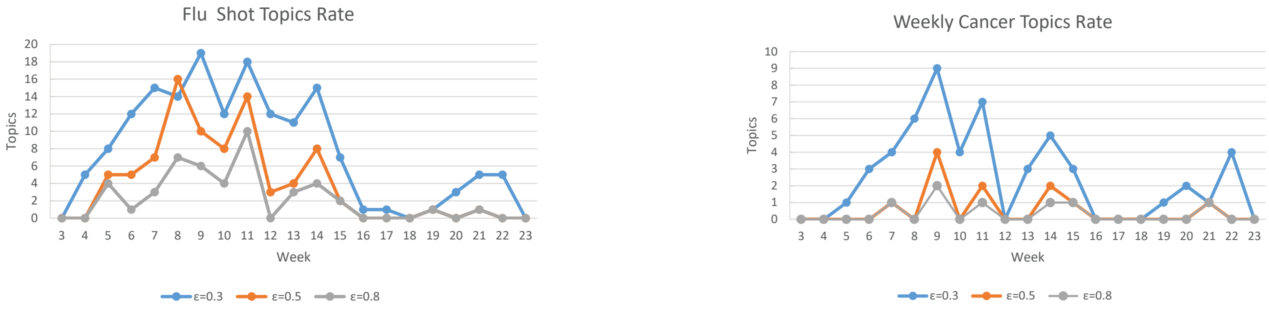
Figure 5: Baseline models



(a) Flu Topics

(b) Influenza Topics

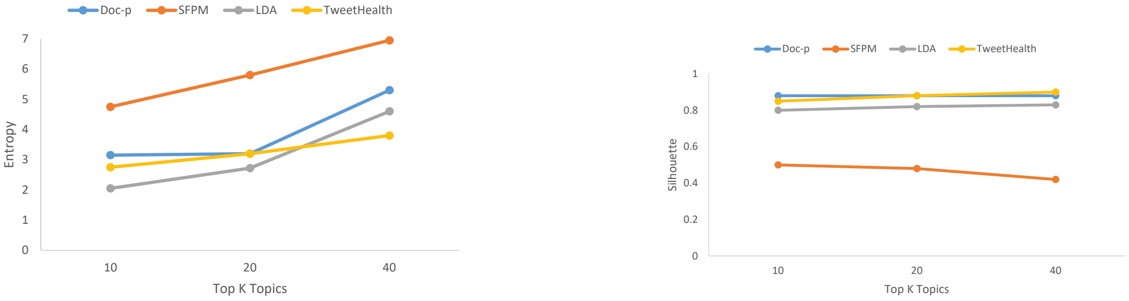
Figure 6: Weekly Topics Rate of Flu and Influenza



(a) Flu Shot Topics

(b) Cancer Topics

Figure 7: Weekly Topics Rate of Flu Shot and Cancer

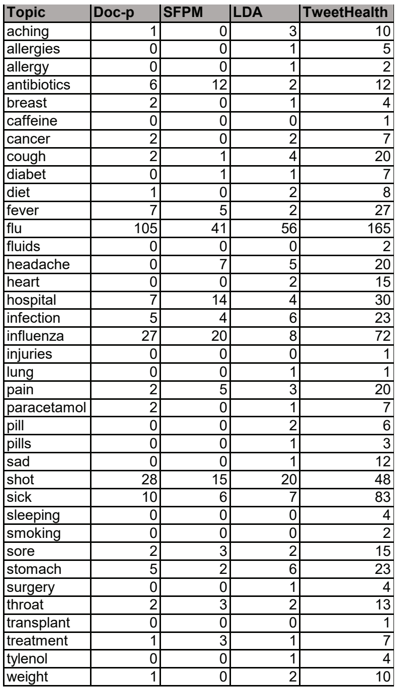


(a) Entropy

(b) Silhouette

Figure 8: Entropy and Silhouette w.r.t. the top K topics for the different algorithms

The overall experimental evaluation proved the effectiveness and efficacy of the HealthS-Tweet algorithm, showing its ability to detect a large number of health topics, including the ones with low frequency terms, largely outperforming traditional topic modeling methods. We can conclude by saying that the above results are encouraging. We have shown that user-generated content on Twitter does indeed provide relevant health-related data, whose identification



Topic	Doc-p	SFPM	LDA	TweetHealth
aching	1	0	3	10
allergies	0	0	1	5
allergy	0	0	1	2
antibiotics	6	12	2	12
breast	2	0	1	4
caffeine	0	0	0	1
cancer	2	0	2	7
cough	2	1	4	20
diabet	0	1	1	7
diet	1	0	2	8
fever	7	5	2	27
flu	105	41	56	165
fluids	0	0	0	2
headache	0	7	5	20
heart	0	0	2	15
hospital	7	14	4	30
infection	5	4	6	23
influenza	27	20	8	72
injuries	0	0	0	1
lung	0	0	1	1
pain	2	5	3	20
paracetamol	2	0	1	7
pill	0	0	2	6
pills	0	0	1	3
sad	0	0	1	12
shot	28	15	20	48
sick	10	6	7	83
sleeping	0	0	0	4
smoking	0	0	0	2
sore	2	3	2	15
stomach	5	2	6	23
surgery	0	0	1	4
throat	2	3	2	13
transplant	0	0	0	1
treatment	1	3	1	7
tylenol	0	0	1	4
weight	1	0	2	10

Figure 9: Number of produced topics for the different algorithms.

may allow to tailor health interventions more effectively.

4. CONCLUSIONS

The paper presented an online clustering algorithm to detect health topics from Twitter streams. The algorithm incrementally groups tweets dealing with the same disease in topics, through a similarity function that takes into account the terms occurring in the tweet and their frequencies. The experimental evaluation on the tweets posted by users in US showed that the method is capable to group tweets addressing common health issues into the pertinent topic, outperforming traditional topic model approaches like Doc-p and LDA. The detected topics are very different, spanning from seasonal diseases like flu and allergy, to common diseases, like cancer, or exceptional events like bird flu. The method we proposed can be considered as a valuable contribution to the emerging area of the e-health that exploits social media to improve public health management. As for future work we plan to explore sentiment analysis techniques to catch users feelings and moods (e.g., fear, concern, happiness).

5. REFERENCES

[1] A. Culotta, "Towards detecting influenza epidemics by analyzing twitter messages," in *Proceedings of the First Workshop on Social Media Analytics*, ser. SOMA '10. ACM, 2010, pp. 115–122.

[2] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '10. ACM, 2010, pp. 1029–1038.

[3] V. Lampos, T. De Bie, and N. Cristianini, *Flu Detector - Tracking Epidemics on Twitter.* Springer Berlin Heidelberg, 2010, pp. 599–602.

[4] A. Sadilek, H. Kautz, and V. Silenzio, "Modeling spread of disease from social interactions," in *In Sixth AAAI International Conference on Weblogs and Social Media (ICWSM*, 2012.

[5] T. Nguyen, D. T. Nguyen, M. E. Larsen, B. O'Dea, J. Yearwood, D. Phung, S. Venkatesh, and H. Christensen, "Prediction of population health indices from social media using kernel-based textual and temporal features," in *Proceedings of the 26th International Conference on World Wide Web Companion*, ser. WWW '17 Companion, 2017, pp. 99–107.

[6] M. J. Paul and M. Dredze, "Discovering health topics in social media using topic models," *PLoS ONE*, vol. 9, no. 8, 2014.

[7] Y. Zhang, J. Tang, J. Sun, Y. Chen, and J. Rao, "Moodcast: Emotion prediction via dynamic continuous factor graph model," in *ICDM 2010, The 10th IEEE International Conference on Data Mining*, 2010, pp. 1193–1198.

[8] M. De Choudhury, S. Counts, and E. Horvitz, "Major life changes and behavioral markers in social media: Case of childbirth," in *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, ser. CSCW '13. ACM, 2013, pp. 1431–1442.

[9] Y. W. Teh, D. Newman, and M. Welling, "A collapsed variational bayesian inference algorithm for latent dirichlet allocation," *Adv. Neural Inf. Process. Syst*, 2007.

[10] S. Petrović, M. Osborne, and V. Lavrenko, "Streaming first story detection with application to twitter," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.* Association for Computational Linguistics, 2010, pp. 181–189.

[11] N. P. Carmela Comito, Clara Pizzuti, "Online clustering for topic detection in social data streams," in *28th IEEE International Conference on Tools with Artificial Intelligence, ICTAI*, 2016, pp. 362–369.

[12] L. M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Goker, I. Kompatsiaris, and A. Jaimes, "Sensing trending topics in twitter," *IEEE Transactions on Multimedia*, vol. 15, no. 6, pp. 1268–1282, 2013.

[13] (2017) Using machine learning to analyze twitter for real time influenza surveillance. [Online]. Available: https://medium.com/@justinzcai/

[14] (2017) Amazon mechanical turk. [Online]. Available: https://www.mturk.com

[15] (2017) Centers for disease control and prevention. [Online]. Available: //www.cdc.gov/flu/about/season/flu-season-2015-2016.htm