Dynamic Topic Modeling using Social Network Analytics

Shazia Tabassum[1], João Gama[1], Paulo Azevedo[1], Luis Teixeira[2], Carlos Martins[2], and Andre Martins[2]

[1] INESC TEC, University of Porto, Rua Dr. Roberto Frias, Porto, Portugal
https://www.inesctec.pt/
[2] Skorr, Portugal
https://skorr.social/

Abstract. Topic modeling or inference has been one of the well-known problems in the area of text mining. It deals with the automatic categorisation of words or documents into similarity groups also known as topics. In most of the social media platforms such as Twitter, Instagram, and Facebook, hashtags are used to define the content of posts. Therefore, modelling of hashtags helps in categorising posts as well as analysing user preferences. In this work, we tried to address this problem involving hashtags that stream in real-time. Our approach encompasses graph of hashtags, dynamic sampling and modularity based community detection over the data from a popular social media engagement application. Further, we analysed the topic clusters' structure and quality using empirical experiments. The results unveil latent semantic relations between hashtags and also show frequent hashtags in a cluster. Moreover, in this approach, the words in different languages are treated synonymously. Besides, we also observed top trending topics and correlated clusters.

Keywords: Topic modelling · Social network analysis · Hashtag networks.

1 Introduction

Social media applications such as Twitter, Facebook, Instagram, Google, Linkedin have now become the core aspect of people's lives. Consequently, these are growing into a dominant platform for businesses, politics, education, marketing, news and so forth. The users are interested in which of such topics or products is one of the primary questions of research in this area. Inferring topics from unstructured data has been quite a challenging task.

Typically, the data gathered by the above applications is in the form of posts generated by the users. Posts can be short texts, images, videos, messy data such as concatenated words, URLs, misspelled words, acronyms, slangs and more. Classification of posts into topics is a complex problem. While topic modeling algorithms such as Latent semantic analysis and Latent dirichlet allocation are originally designed to derive topics from large documents such as articles, and

books. They are often less efficient when applied to short text content like posts [1]. Posts on the other hand are associated with rich user-generated hashtags to identify their content, to appear in search results and to enhance connectivity to the same topic. In [18] the authors state that hashtags provide a crowd sourcing way for tagging short texts, which is usually ignored by Bayesian statistics and Machine learning methods. Therefore, in this work, we propose to use these hashtags to derive topics using social network analysis methods, mainly community detection.

Moreover, the data generated from social media is typically massive and high velocity. Therefore, we tried to address the above issues by proposing an approach with the contributions stated below:

1. We propose fast and incremental method using social network analytics.
2. Unlike conventional models we use hashtags to model topics which saves the learning time, preprocessing steps, removal of stop words etc.
3. Our model categorises tags/words based on connectivity and modularity. In this way the tags/words are grouped accurately even though they belong to different languages or new hashtags appear.
4. We employ dynamic sampling mechanisms to decrease space complexity.

Rest of the paper is organised as follows: In section 2 we presented a brief overview of the related works. Section 3 details the data set and some statistics about it. The methodology is described in section 4. The experiments and results are discussed in section 5. Finally, section 6 summarizes conclusions and some potential future works.

2 Related Work

Research works focusing on topic modelling are mostly based on inferring abstract topics from long text documents. Latent dirichlet allocation [5] is one of the most popular techniques used for topic modelling where the topic probabilities provide an explicit representation of a document. However, it assumes fixed number of topics that a document belongs to. Other well known models include Latent semantic analysis [8], Correlated topic models [4], Probabilistic latent semantic indexing [10]. Word2Vec [13] is another popular word representation techniques. This model outputs a vector for each word, so it is necessary to combine those vectors to retrieve only one representation per product title or post, since there is the need to have the entire sentence representation and not only the values of each word. Word2Vec output dimensions can be configurable, and there is no ideal number for it since it can depend on the application and the tasks being performed. Moreover, these types of models are very common and can be expensive to train. However, traditional topic models also known as flat text models are incapable of modeling short texts or posts due to the severe sparseness, noise and unstructured data [11], [18].

Recently, several researchers have focused on specifically hashtags clustering. In [14] the authors clustered hashtags using K-means on map reduce to find the

structure and meaning in Twitter hashtags. Their study was limited to understanding the top few hashtags from three clusters. They found the top hashtags to be understandable as they are popular and while increasing the number of clusters the hashtags are dispersed into more specific topics. In another interesting work, multi-view clustering was used to analyse the temporal trends in Twitter hashtags during the Covid-19 pandemic [7]. The authors found that some topic clusters shift over the course of pandemic while others are persistent. Topic modelling was also applied on Instagram hashtags for annotating images [2]. In [3] the authors clustered twitter hashtags into several groups according to their word distributions. The model was expensive as Jensen-Shannon divergence was calculated between any two hashtags from the data. However, they considered a very small data set and calculated the probabilities for top 20 frequent hashtags while the structure and quality of clusters was not analysed.

While most of the models above were run on small-scale data sets crawled from one of the social media applications, we used a considerably large one which is composed of data from several micro blogging applications and also visualised the quality and structure of our clusters. Moreover, our approach is dynamic considering community detection for clustering tags.

3 Case Study

An anonymized data set is collected from a social media activity management application. The data set ranges from January to May 2020; comprises of 1002440 posts with 124615 hashtags posted by users on different social networking platforms (Twitter, Facebook, Instagram, Google). The content of posts is not available, instead the posts are identified with posts IDs and the users are identified with anonymous user IDs. Figure 2 displays the distribution of hashtags vs posts. A few hashtags are used by large number of posts and many different hashtags are discussed by only some users. This satisfies a power law relation which is usually seen in most of the real world social networks [17]. Each post can include one or more hashtags or none. The number of posts per day is given in Figure 1 which shows the seasonality of data. As one can observe there is decreased activity on weekends (Saturday and Sunday) compared to other days with the peaks on Fridays. The data in the last week of April had not been available which can be seen as an inconsistency in the curve with abnormally low activity close to zero. Figure 3 displays the top ten trending hashtags in the given data set. This type of analysis with the help of topic modelling or trending hashtags can be used to detect events. In the figure, the top two of frequent hashtags are relating to Covid19. What we need from our model is to cluster these hashtags and also the one's that are less frequent (such as covid, covid 19, corona etc.) to be classified as one topic relating to Covid19. Similarly, with the other tags and their related posts. In order to achieve this we followed the methodology briefed below.

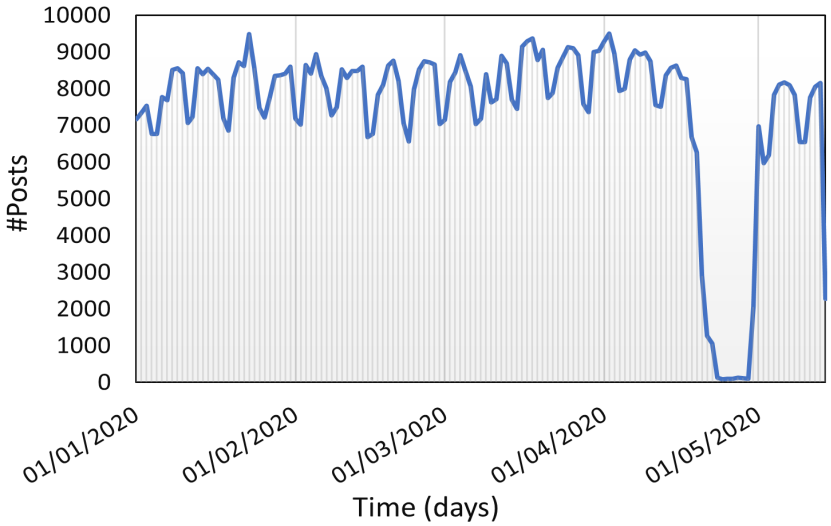

Fig. 1. Temporal distribution of posts per day

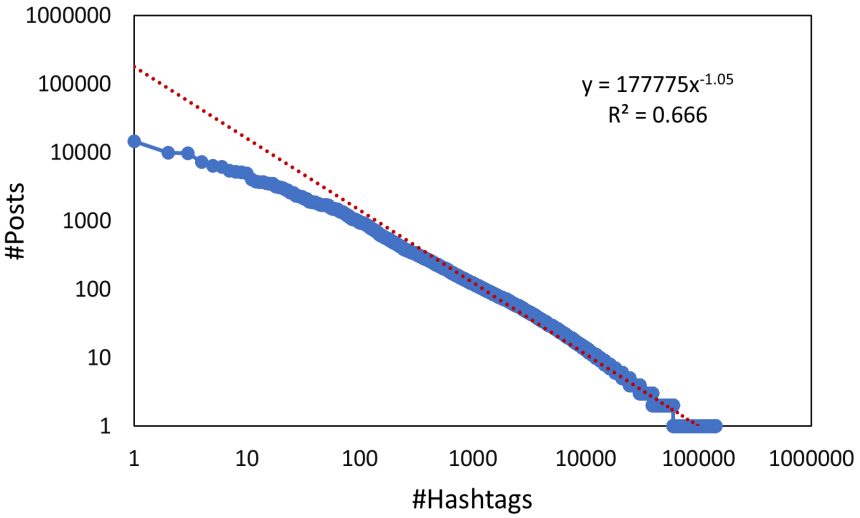

Fig. 2. Posts vs hashtags distribution (blue line). Power curve following given function(red)

4 Methodology

Text documents share common or similar words between them, which is exploited in calculating similarity scores. However, topic modeling in hashtags is unlike

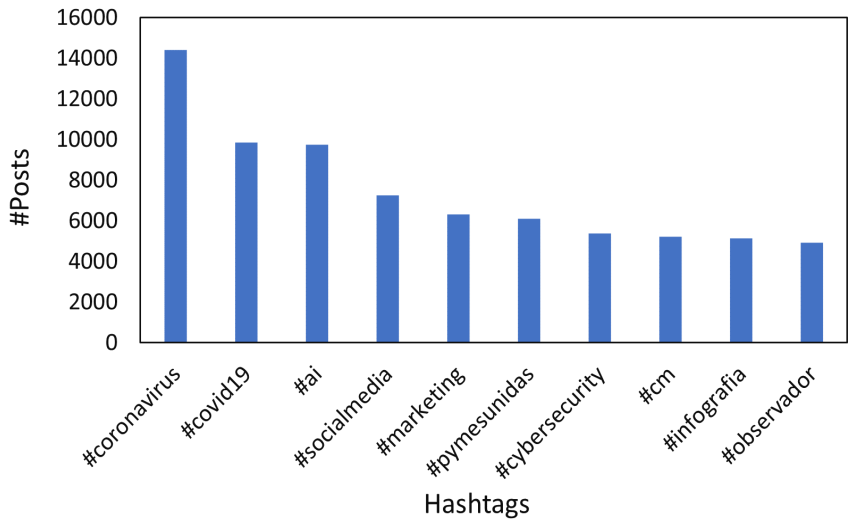

Fig. 3. Top ten trending hashtags distribution

documents. Therefore, here we considered the hashtags to be similar based on their co-occurrence in a post.

The first step in the process is to build a co-occurrence network from the streaming hashtags incrementally. The hashtags that needs to stay in the network are decided based on the choice of the sampling algorithm in Section 4.3. There after the communities are detected in the network as detailed in Section 4.4.

4.1 Problem Description

Given a stream of posts $\{p_1, p_2, p_3...\}$ associated with hashtags $\{h_1, h_2, h_3...\}$ arriving in the order of time, our approach aims to categorize similar posts or hashtags into groups or clusters called topics at any time t. Each post can be associated with one or more hashtags.

4.2 Hashtag Co-occurrence Network

In our graph based approach, we constructed the network of hashtags by creating an edge e between the ones that have been tagged together in a post. Therefore $e = (h_i, h_j, t)$ where $i, j \in \mathbb{N}$ and t is the time stamp when it occurred.

4.3 Stream Sampling

As posts are temporal in nature generating in every time instance, so are the hashtags. Also, there are new hashtags emerging over time. Moreover, the context for grouping hashtags may change over time. For example, hand sanitizers and

face masks were not as closely related as with the onset of covid19. Therefore, we employed the approach of exploiting the relation between hashtags based on the recent events or popular events by using the real-time dynamic sampling techniques below.

Sliding Windows Sometimes applications need recent information and its value diminishes by time. In that case sliding windows continuously maintain a window size of recent information [9]. It is a common approach in data streams where an item at index i enters the window while another item at index $i - w$ exits it. Where w is the window size which can be fixed or adaptive. The window size can be based on number of observations or length of time. In the later case an edge (h_i, h_j, t) enters window while an edge $(h_i, h_j, t - w)$ exits.

Space Saving The Space Saving Algorithm [12] is the most approximate and efficient algorithm for finding the top frequent items from a data stream. The algorithm maintains the partial interest of information as it monitors only a subset of items from the stream. It maintains counters for every item in the sample and increments its count when the item re-occurs in the stream. If a new item is encountered in the stream, it is replaced with an item with the least counter value and its count is incremented.

Biased Random Sampling This algorithm [16] ensures every incoming item m in stream goes into the reservoir with probability 1. Any item n from the reservoir is chosen for replacement at random. Therefore, on every item insertion, the probability of removal for the items in the reservoir is $1/k$, where k is the size of reservoir. Hence, the item insertion is deterministic but deletion is probabilistic. The probability of n staying in the reservoir when m arrives is given by $(1 - 1/k)^{(t_m - t_n)}$. As the time of occurrence or index of m increases, the probability of item n from time t staying in reservoir decreases. Thus the item staying for a long time in the reservoir has an exponentially greater probability of getting out than an item inserted recently. Consequently, the items in the reservoir are super linearly biased to the latest time. This is a notable property of this algorithm as it does not have to store the ordering or indexing information as in sliding windows. It is a simple algorithm with $O(1)$ computational complexity.

4.4 Community Detection

Community detection is very well known problem in social networks. Communities can be defined as groups, modules or clusters of nodes which are densely connected between themselves and sparsely connected to the rest of the network. The connections can be directed, undirected, weighted etc. Communities can be overlapping (where a node belongs to more than one community) or distinct. Community detection is in its essence a clustering problem. Thus, detecting communities reduces to a problem of clustering data points. It has a wide scope of applicability in real-world networks.

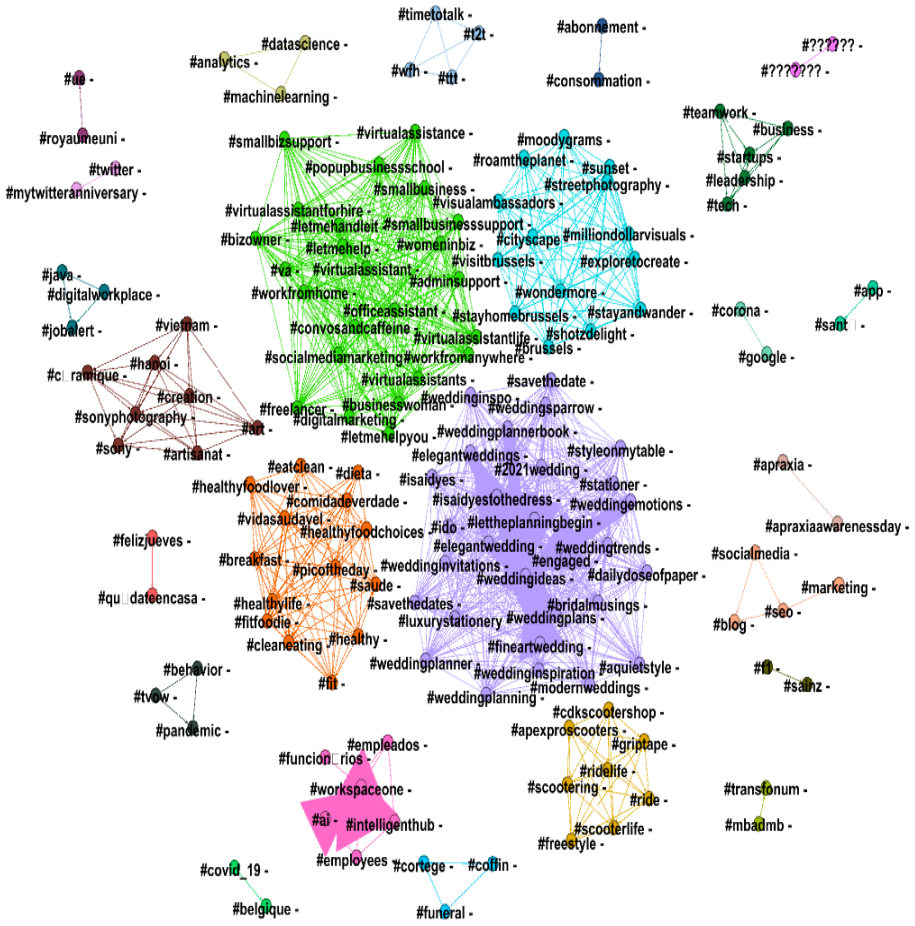

Fig. 4. Sliding Window

In this work, we applied the community detection algorithm proposed by Blondel et al. [6] on every dynamic sample snapshot discretely. However, an incremental community detection algorithm can also be applied on every incoming edge. Nevertheless, the technique mentioned above is a heuristic based on modularity optimization. **Modularity** is a function that can be defined as the number of edges within communities minus the number of expected edges in the same at random [15] as computed below.

$$Q = \frac{1}{2m} \sum [A_{ij} - \frac{k_i k_j}{2m}] \delta(c_i, c_j), \tag{1}$$

where m is the number of edges, k_i and k_j represent, respectively, the degree of nodes i and j, A_{ij} is the entry of the adjacency matrix that gives the number

of edges between nodes i and j, $\frac{k_i k_j}{2m}$ represents the expected number of edges falling between those nodes, c_i and c_j denote the groups to which nodes i and j belong, and $\delta(c_i, c_j)$ represents the Kronecker delta. Maximizing this function leads to communities with highly connected nodes between themselves than to the rest of the network. However in very large networks the connections are very sparse and even a single edge between two clusters is regarded as strong correlation. Therefore, a resolution parameter is used to control high or low number of communities to be detected. Modularity is also used as a quality metric as shown in Table 1.

The above said algorithm has a fastest runtime of $O(n.log_2 n)$, where n is the number of nodes in the network. In our case n is very small compared to the total number of nodes in the network, for instance n is equal to the number of hashtags in a sliding window.

5 Experimental Evaluation

The experiments are conducted to evaluate the above method of detecting topic clusters in the data detailed in Section 3. To facilitate visual evaluation and demonstration, the size of samples is fixed to be 1000 edges. In the case of sliding windows the window size is based on number of observations i.e. 1000 edges. However, a time window such as edges from recent one day/one month can also be considered. The resolution parameter in community detection for all the methods is set to 1.0. The detected clusters are shown in Figures 4, 5, and 6. The figures represent sample snapshots in the end of stream. Each cluster with a different color in the figure represents a topic. Sliding windows and biased sampling considers repetitive edges as the frequency or weight of an edge which is depicted as thick arrows or lines in the figures. The thicker edge represents stronger connection between two hashtags. The hashtags with thicker edges are considered top hashtags in their cluster as they are most frequent.

The choice of sampling algorithm has different trade offs. For finding the most frequent or trending topics from the stream over time, space saving is a relevant choice; however, it is computationally expensive compared to the other two though it is space efficient and the fastest one of its genre. The one with least time complexity among the three is biased sampling but lacks in terms of structure in this case, with a very sparse graph.

5.1 Results Discussion

We see that the clusters in the figures clearly make sense in terms of synonymy and polysemy (for example in Figure 5, synonyms such as covid, covid 19, etc., are grouped in one blue cluster on the right and polysemy words are sharing two clusters green and orange in the center). The clusters formed by sliding windows are more denser than the other two. Quantitative metrics of these graphs are displayed in Table 1. The bias to low degree hashtags has increased the number of components and decreased the density. Nevertheless, a large cluster of the

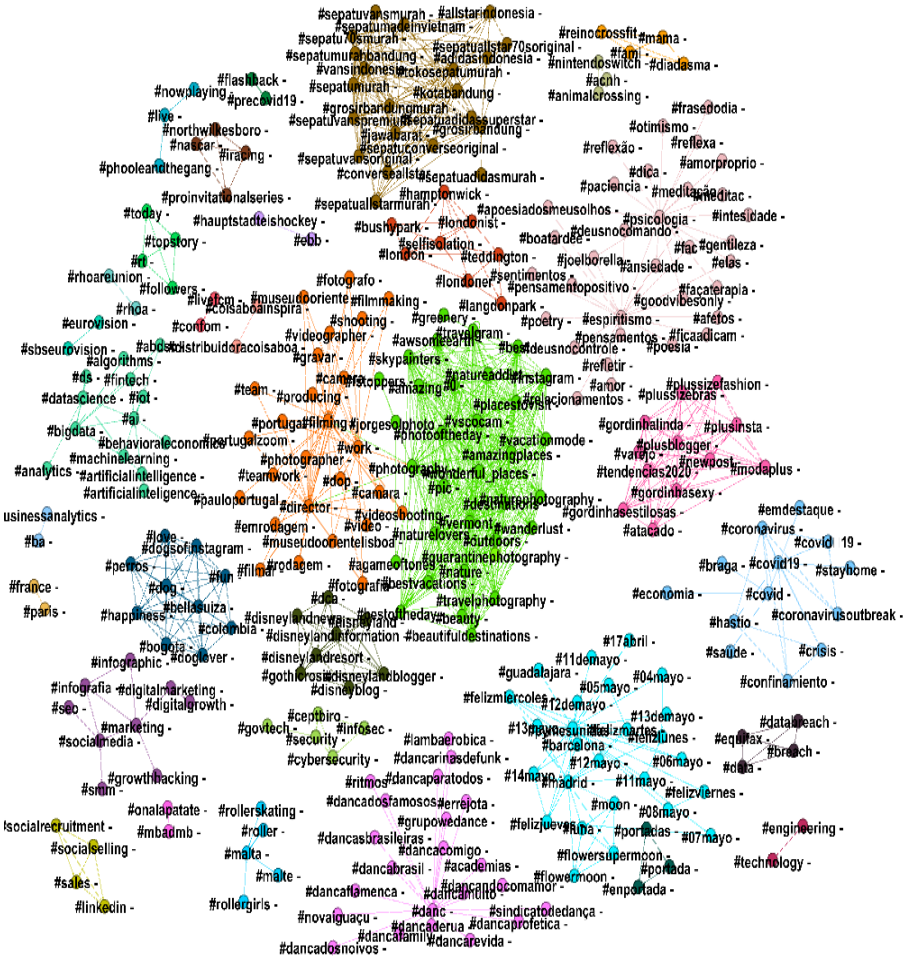

Fig. 5. Space Saving

popular topic "covid19" can only be seen in space saving because sliding window and biased sampling collect data from the end of stream that is from the month of May, where it has low occurrence in our data. Moreover, the top trending hashtags from Figure 3 in analysis are found inside communities of space saving (Figure 5). However, they are also found in sliding window and biased sampling as we increase the sample size.

The results also show correlated clusters that share common hashtags as seen in Figure 5, the green and orange cluster. Another important feature is that tags in different languages (Portugues, Spanish, or more) are still clustered semantically, such as in Figure 5 "confinamiento" belongs to covid cluster. Moreover, acronyms such as "ai" belongs to artificial intelligence cluster.

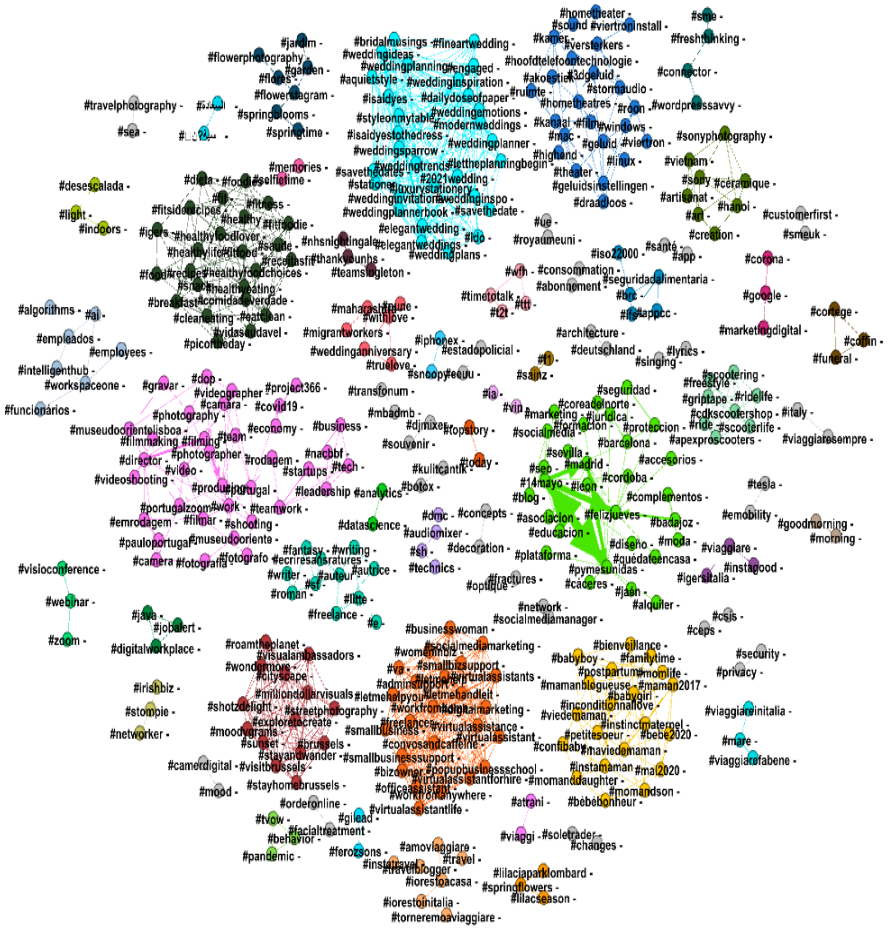

Fig. 6. Biased Random Sampling

The posts and users relating to these hashtags can be further investigated for numerous applications. Each post is associated with multiple hashtags, therefore each post can be assigned to a number of topics.

6 Conclusion and Future Work

In this work, we have presented a fast and memory efficient approach for incrementally categorising posts into topics using hashtags. We proved the efficacy of method over a large data set. We discussed how the different sampling algorithms can effect the outcome. Further, we considered their biases and trade offs. We analysed the seasonality and trending hashtags in the data. We compared

	Average degree	Avg. weighted degree	Density	Modularity	#Clusters
Sliding window	6.6	6.667	0.087	0.723	25
Space saving	3.413	3.413	0.022	0.806	33
Biased sampling	2.687	2.747	0.015	0.872	61

Table 1. Network properties

their outcomes in terms of semantics and structure of clusters. To facilitate comprehensibility we preferred network visualisation layouts over the conventional presentation using tables.

There can be many potential applications as an advancement of this work. The users posting in particular topics can be classified accordingly to analyse their preferences for product marketing and identifying nano influencers to enhance their engagement. On the availability of posts text we can implement other topic models and improve them using our approach. Further, we intend to analyse the trend of topics overtime and the evolution of communities. Additionally, predicting hashtags for the missing ones using our topic model.

Acknowledgements

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project UIDB/50014/2020.

References

1. Alash, H.M., Al-Sultany, G.A.: Improve topic modeling algorithms based on twitter hashtags. In: Journal of Physics: Conference Series. vol. 1660, p. 012100. IOP Publishing (2020)
2. Argyrou, A., Giannoulakis, S., Tsapatsoulis, N.: Topic modelling on instagram hashtags: An alternative way to automatic image annotation? In: 2018 13th international workshop on semantic and social media adaptation and personalization (SMAP). pp. 61–67. IEEE (2018)
3. Bhakdisuparit, N., Fujino, I.: Understanding and clustering hashtags according to their word distributions. In: 2018 5th International Conference on Business and Industrial Research (ICBIR). pp. 204–209. IEEE (2018)
4. Blei, D., Lafferty, J.: Correlated topic models. Advances in neural information processing systems **18**, 147 (2006)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research **3**, 993–1022 (2003)
6. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment **2008**(10), P10008 (2008)
7. Cruickshank, I.J., Carley, K.M.: Characterizing communities of hashtag usage on twitter during the 2020 covid-19 pandemic by multi-view clustering. Applied Network Science **5**(1), 1–40 (2020)

8. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science **41**(6), 391–407 (1990)
9. Gama, J.: Knowledge Discovery from Data Streams. Chapman and Hall / CRC Data Mining and Knowledge Discovery Series, CRC Press (2010)
10. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. pp. 50–57 (1999)
11. Hong, L., Davison, B.D.: Empirical study of topic modeling in twitter. In: Proceedings of the first workshop on social media analytics. pp. 80–88 (2010)
12. Metwally, A., Agrawal, D., El Abbadi, A.: Efficient computation of frequent and top-k elements in data streams. In: International conference on database theory. pp. 398–412. Springer (2005)
13. Mikolov, T., Yih, W.t., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies. pp. 746–751 (2013)
14. Muntean, C.I., Morar, G.A., Moldovan, D.: Exploring the meaning behind twitter hashtags through clustering. In: International Conference on Business Information Systems. pp. 231–242. Springer (2012)
15. Newman, M.E.: Finding community structure in networks using the eigenvectors of matrices. Physical review E **74**(3), 036104 (2006)
16. Tabassum, S., Gama, J.: Sampling massive streaming call graphs. In: ACM Symposium on Advanced Computing. pp. 923–928 (2016)
17. Tabassum, S., Pereira, F.S., Fernandes, S., Gama, J.: Social network analysis: An overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **8**(5), e1256 (2018)
18. Wang, Y., Liu, J., Huang, Y., Feng, X.: Using hashtag graph-based topic model to connect semantically-related words without co-occurrence in microblogs. IEEE Transactions on Knowledge and Data Engineering **28**(7), 1919–1933 (2016)