Filter hashtag context through an original data cleaning method

Didier Henry, Erick Stattner, Martine Collard

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 00 (2016) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

The 9th International Conference on Ambient Systems, Networks and Technologies (ANT 2018)

Filter hashtag context through an original data cleaning method

Didier Henry[a], Erick Stattner[a], Martine Collard[a]

[a]*LAMIA, University of French West Indies, 97110 Pointe A Pitre, Guadeloupe, France*

Abstract

Nowadays, social networks are one of the most used means of communication. For example, the social network Twitter has nearly 100 million active users who post about 500 million messages per day. Sharing information on this platform is unique because messages are limited in characters number. Faced with this limitation, users express themselves briefly and use sometimes a hashtag that summarizes the general idea of the message. Nevertheless, hashtags are noisy data because they do not respect any linguistic rule, may have several meanings, and their use is not under control. In this work, we tackle the problem of hashtag context which may have useful applications in several fields like information recommendation or information classification. In this paper, we propose an original data cleaning method to extract the most relevant neighbor hashtags of a hashtag. We test our method with a dataset containing hashtags related to several topics (such as sport, music, technology, etc.) in order to show the efficacy and the robustness of our approach.

Peer-review under responsibility of the Conference Program Chairs.

Keywords: social media ; hashtag ; context ; data cleaning

1. Introduction

In just a few years, social media has become a huge platform for exchange and sharing that collects writing, opinions, thoughts and events of humanity. However, while these spaces of public exchange are today firmly anchored in our modern societies, we observe that ways of communicating have also evolved considerably through these media.

In particular, the limitation in terms of size imposed by certain platforms and the encouragement of rapid and impulsive sharing often incite users to share short messages, consisting of short words and accompanied by emoticons. Moreover, these messages are often associated with *hashtags*, that is to say one or several words preceded by the symbol '#', for instance: *#xboxone, #iphone, #tesla*.

Hashtags aim to join discussions on emergent topics[1], to identify or track conversations on the same event[2] or even to serve as the symbol of a community[3,4]. Thus, hashtags may be useful to find a related content on social media but do not respect any linguistic rule: *#musicislife, #yummm, #tlot*. In addition, a single hashtag may have several

* Corresponding author. Tel.: +590-059-048-3074 ; fax: +590-059-048-3086.
E-mail address: didier.henry@univ-antilles.fr

Peer-review under responsibility of the Conference Program Chairs.

meanings, for example, *#controller* can refer to a job or to an input device used for playing video games. Moreover, a hashtag present in a message may be inappropriate simply because their creation and sharing are not controlled. Therefore, the understanding of hashtags may be difficult for machines and humans because they are noisy data in terms of the lexicon, syntax and semantic.

In recent years, numerous researchers[5,6,7,8] have expressed interest in classification or category recognition of messages posted on social media. Michelson et al.[9] note that hashtags are ungrammatical and noisy. They have proposed a method linking a tweet to a tree of Wikipedia categories. Next this insight, Genc et al.[10] have shown that a Wikipedia-based technique produces better classification messages accuracy than Latent Semantic Analysis[11] (LSA) and String Edit Distance[12] (SED). In their work, Ferragina et al.[13] have focused on semantic hashtags classification. They have used a Hashtag-Entity Graph, where entities are linked to Wikipedia categories and a support vector machine classifier. Thus, we observe that Wikipedia has been used in several approaches and seems to be useful to both messages and hashtags classification task.

Messages and hashtags classification problem is closely linked to messages and hashtags recommendations. Li et al.[14] have proposed a method to suggest hashtags by using WordNet and a Euclidean similarity distance. In their approach, Godin et al.[15] have used a Latent Dirichlet Allocation (LDA) model to recommend hashtags from the message. Lu et al.[16] have introduced a recommendation tweets system based on user previous tweets. In a recent work, Gong et al.[17] have adopted a convolutional neural networks to perform the hashtag recommendation problem.

Messages classification may have useful applications for studies in the field of information diffusion. For instance, Romero et al.[18] have remarked that information propagation is different according to the topic. Myers et al.[19] have noticed that messages about education, art or work have a shorter reach than others. Several works[20,21] have observed discounts and promotions dissemination in social networks. Other researchers[22,23] have proposed diffusion models taking into consideration the message topic.

In this work, we are interested in the contextualisation of hashtags by other hashtags. We argue that such a context may improve information understanding both machines and humans, and may be useful in information recommendation/classification field. Nevertheless, some hashtags neighbour may be incoherent with the hashtag context. Our aim is to extract the most relevant hashtags related to a hashtag, so indirectly find messages dealing with close topic. To the best of our knowledge, not any work has focused on the hashtag context filtering.

The social network Twitter is a good case study to address this kind of problem. Indeed, Twitter has been a pioneer in the appearance of hashtags, and also contains a wide variety of users: professionals, individuals, politicians, associations, unions, companies, etc. In addition, this platform has nearly 100 million active users and 500 million messages are posted every day.

The rest of the article is organised as follows: Section 2 describes the proposed methodology. Section 3 details the experiments and the obtained results. Section 4 is dedicated to the discussion. Finally, Section 5 concludes and presents our future directions.

2. Methodology

Our aim is to filter the hashtag context which may contain irrelevant hashtags. We propose a data cleaning method based on the generation of three hashtags context: the *time context*, the *artificial context* and the *recent context*. In order to obtain the cleaned context, we choose hashtags present in at least two contexts (see Figure 1). Algorithm 1 presents this approach. Our method turns out to be language-independent as long as there are Wikipedia pages related several topics for the target language. In addition, our method is suitable for parallel computing.

Function `getCleanedContext(`*hashtag*`)`**:**
    $HRecentList \longleftarrow getRecentContext(hashtag, nbTweets, nbRelatedHashtags, dateSince, dateUntil)$;
    $HTimeList \longleftarrow getTimeContext(hashtag, nbTweets, nbRelatedHashtags, dateSince1, dateUntil1, dateSince2, dateUntil2, dateSince3, dateUntil3)$;
    $HArtificialList \longleftarrow getArtificialContext(hashtag, nbTweets1, nbRelatedHashtags, nbTopicsWords, dateSince, dateUntil)$;
    $relatedHashtagsList \longleftarrow keepCommonHashtags(HRecentList, HTimeList, HArtificialList)$;
    return *relatedHashtagsList*;

Algorithm 1: Algorithm for cleaning the hashtag context

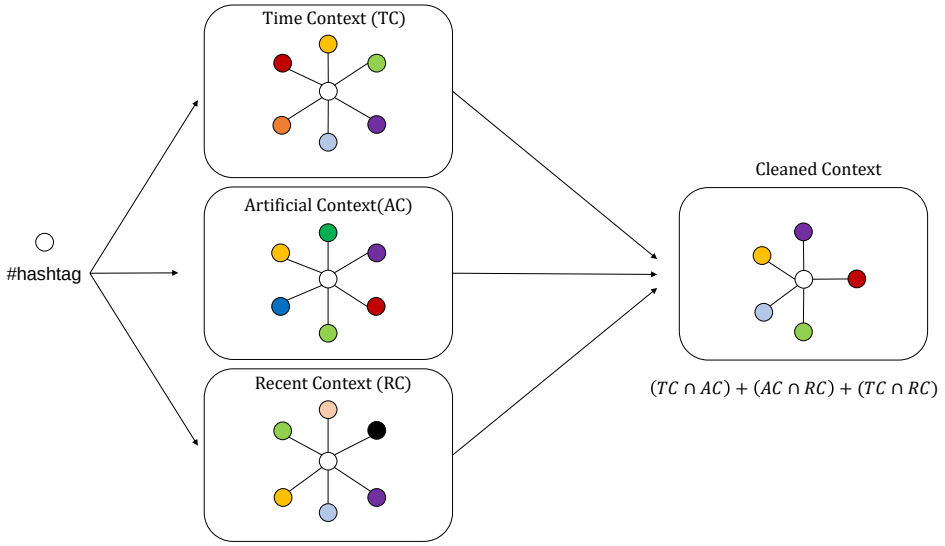


Fig. 1. Methodology of hashtag context filtering.

Recent context. As its name suggests, the recent context is generated by keeping the top x of the most present hashtags from N recent tweets. This context is the baseline context used in our experiments. In algorithm 2 we give a description of this concept.

Function `getRecentContext`(*hashtag, nbTweets, nbRelatedHashtags, dateSince, dateUntil*)**:**
  HRecentList ⟵ *getRelatedHashtags*(*hashtag*, *nbTweets*, *nbRelatedHashtags*, *dateSince*, *dateUntil*);
  return *HRecentList*;

Function `getRelatedHashtags`(*hashtag, nbTweets, nbRelatedHashtags, dateSince, dateUntil*)**:**
  tweetsList ⟵ *getTweetsList*(*hashtag*, *nbTweets*, *dateSince*, *dateUntil*);
  hashtagsListGlobal ⟵ *getHashtags*(*tweetsList*);
  countOccurrences(*hashtagsListGlobal*, *hashtagsList*, *hashtagsListCount*);
  mergeSort(*hashtagsList*, *hashtagsListCount*);
  index ⟵ *length*(*hashtagsList*) − 1;
  nbRelatedTags ⟵ 0;
  while *index*≥*0* and *nbRelatedTags*<*nbRelatedHashtags* **do**
    relatedHashtagsList.add(*hashtagsList.get*(*index*));
    nbRelatedTags + +;
    index − −;
  end
  return *relatedHashtagsList*;

Algorithm 2: Recent context algorithm.

Time context. The time context is based on the idea that the meaning of the hashtag changes little over time. Thus, we propose a temporal filter in order to keep hashtags neighbour invariant over time. The temporal filer consists to generate a top x related hashtags from N tweets on three disjointed time intervals. To build the time context composed up to x linked hashtags, we select hashtags present in at least two contexts (see Figure 2). Algorithm 3 introduces the time context.

Function `getTimeContext`(*hashtag, nbTweets, nbRelatedHashtags, dateSince*1*, dateUntil*1*, dateSince*2*, dateUntil*2*,dateSince*3*, dateUntil*3)**:**
  *HTimeList*1 ⟵ *getRelatedHashtags*(*hashtag*, *nbTweets*, *nbRelatedHashtags*, *dateSince*1, *dateUntil*1);
  *HTimeList*2 ⟵ *getRelatedHashtags*(*hashtag*, *nbTweets*, *nbRelatedHashtags*, *dateSince*2, *dateUntil*2);
  *HTimeList*3 ⟵ *getRelatedHashtags*(*hashtag*, *nbTweets*, *nbRelatedHashtags*, *dateSince*3, *dateUntil*3);
  HTimeList ⟵ *keepCommonHashtags*(*HTimeList*1, *HTimeList*2, *HTimeList*3);
  return *HTimeList*;

Algorithm 3: Time context algorithm.

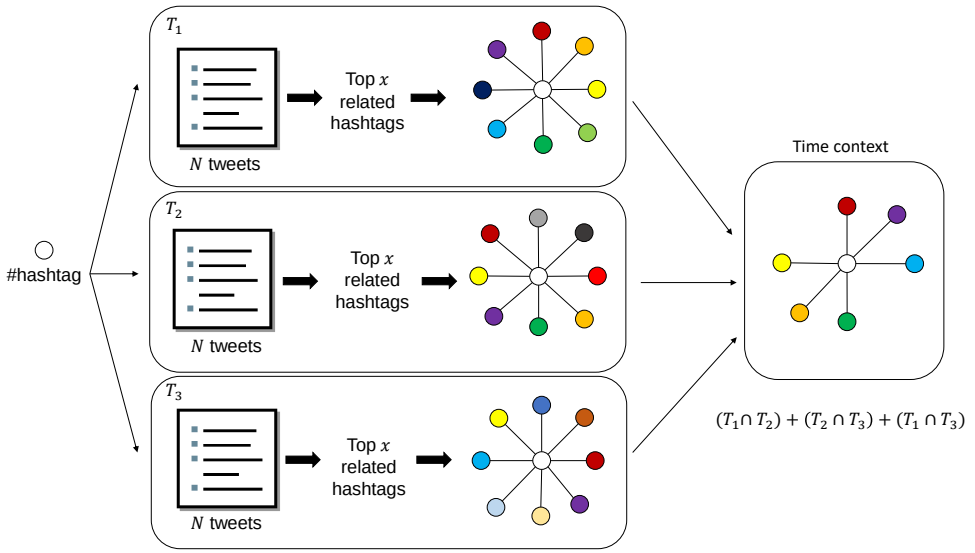


Fig. 2. Methodology of the *time context* generation.

Artificial context. The concept is to create artificially several noisy contexts for a hashtag, then extract the most present hashtags in all contexts. To that end, we need words related several topics. We have used Wikipedia to extract 8 bags of English words (non-alphabetic characters, names, and stop words have been removed) linked to 8 topics: food, games, health, movies, music, politics, sport, and technology. In order to build the artificial context, we select randomly λ words in each bag of words, next, we transform them in hashtags. Then, we create hashtag pairs composed of the hashtag that we want to contextualise and hashtags from the word bags. We select the Top x related hashtags for each hashtag pair from K tweets. Finally, the artificial context is created by selecting up to x hashtags the most present in each noisy context (see Figure 3). Algorithm 4 describes the artificial context.

```
Function getArtificialContext(hashtag, nbTweets1, nbRelatedHashtags, nbTopicsWords, dateSince, dateUntil):
    wordsTopicsList ⟵ getRandomWordsTopicsList(nbTopicsWords);
    for i=0 to wordsTopicsList.size()-1 do
        keywords ⟵ hashtag+" #"+wordsTopicsList.get(i);
        HListTempo ⟵ getRelatedHashtags(keywords, nbTweets1, nbRelatedHashtags, dateSince, dateUntil);
        for j=0 to HListTempo.size()-1 do
            hashtagsListGlobal.add(HListTempo.get(j));
        end
    end
    countOccurrences(hashtagsListGlobal, hashtagsList, hashtagsListCount);
    mergeSort(hashtagsList, hashtagsListCount);
    index ⟵ length(hashtagsList) − 1;
    nbRelatedTags ⟵ 0;
    while index≥0 and nbRelatedTags<nbRelatedHashtags do
        HArtificialList.add(hashtagsList.get(index));
        nbRelatedTags + +;
        index − −;
    end
    return HArtificialList;
```

Algorithm 4: Artificial context algorithm.

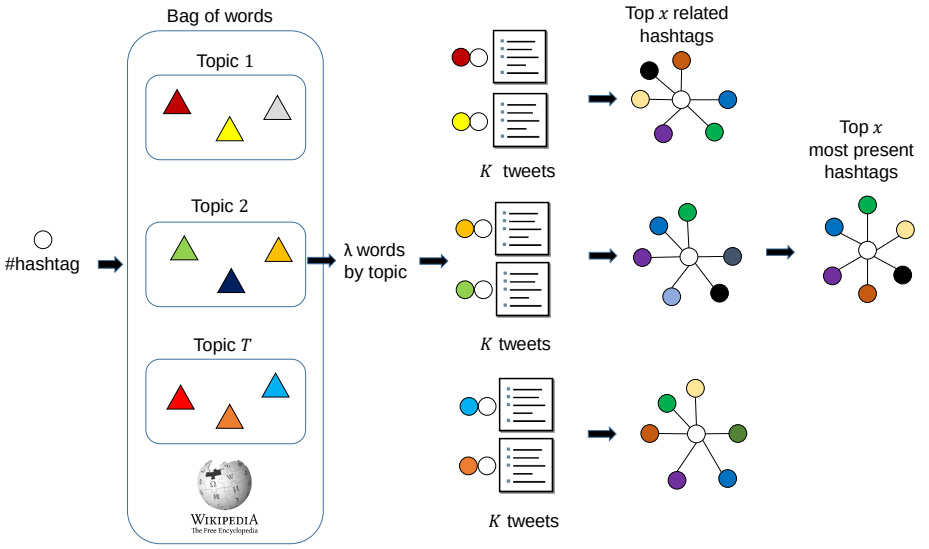


Fig. 3. Methodology of the *artificial context* generation.

3. Experiments and Results

In our experiments, we use *GetOldTweets*[1] library which provides access (without limits) to older tweets in function of keywords by choosing a start date and an end date. We apply our method on a subset of dataset[2] provide by Ferragina et al.[13] containing hashtags related to 8 topics: food, games, health, movies, music, politics, sport and technology. In order to set parameters N and x mentioned above, we have generated baseline contexts of 50 hashtags by varying the number N of tweets used. A human expert has judged the relevance of each hashtag presents in top 20 related hashtags (see Figure 4). We observe that the top 11 related hashtags seem more relevant when 100 tweets are used. To compare the baseline context with the cleaned context, we choose to set N to 100, x to 10, λ and K are also set at 10 in order to get up to 100 tweets by topic. T_1, T_2 and T_3 are set to 2015, 2016 and 2017 respectively.

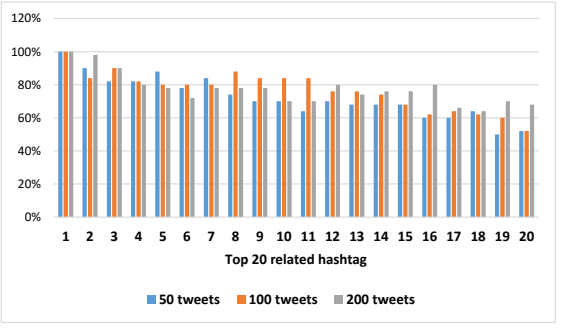


Fig. 4. Average relevance of hashtags in function of rank and number of tweets used to generate the baseline context.

We have selected randomly 65 hashtags for each topic, 520 hashtags in all, to compare the baseline context and the cleaned context. Two human judges (judge 1 is a Twitter user, while judge 2 is not.) have evaluated the two contexts for each hashtag and have noted if one of the contexts is better or if they are equivalent (i.e. equal) both in terms of coherence and readability. For instance, regarding the hashtag *#maternalhealth* (see Table 1), the cleaned

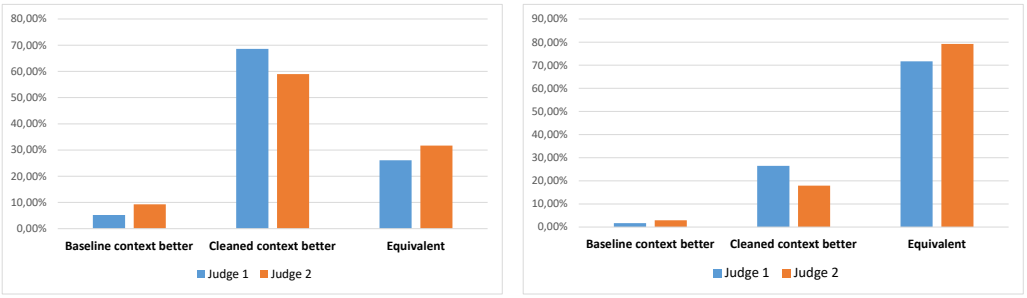
[1] https://github.com/Jefferson-Henrique/GetOldTweets-java
[2] http://acube.di.unipi.it/hashtag-datasets/

context is more relevant and more understandable than the baseline context. Concerning the hashtag *#ps4*, contexts are equivalent in terms of coherence but the cleaned context is more explicit.

Hashtags	Baseline context	Cleaned context
#maternalhealth	#MaternalHealth #SincerelyVee #SWOP2017pic #Malawi #equality #TeamVee #globalhealth #GBV #LamazeAdvocacy17 #Nkhatabay	#MaternalHealth #globalhealth #women #pregnancy
#ps4	#PS4 #BO3 #XB1 #xbox #gaming #twitch #PSN #Destiny2 #XboxOne #PS4sharepic	#PS4 #xbox #XboxOne #gaming #twitch #Gaming

Table 1. Examples of baseline context and cleaned context.

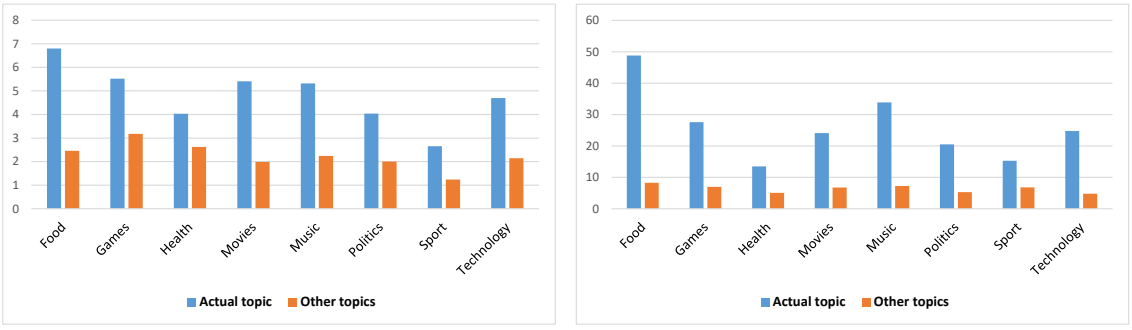
According to judges evaluation on the 520 hashtags, our data cleaning method improve the coherence of hashtag context in 64% of cases on average (see Figure 5). In addition, the cleaned context is more understandable than the baseline context in 22% of cases on average. These results show that our filter is effective on several topics.



(a) (b)

Fig. 5. Comparaison of baseline context and cleaned context on coherence (a) and readability (b).

Moreover, we argue that our artificial context may be helpful for information classification task. Indeed, we have noted that the number of tweets found according to the subject during the generation of the artificial context is larger when the hashtag pair is relevant. In other words, there are more tweets found when both hashtags belong to the same topic. In order to illustrate this fact , we have used 50 hashtags randomly selected (for each topic) and we have set K to 10 and λ to 10 in a first experiment (see Figure 6 (a)), then, we have set K to 100 and λ to 5 (see Figure 6 (b)). We remark the same trend on both experiences.



(a) (b)

Fig. 6. Average number of tweets found when hashtags in a hashtag pair belong to the same topic or do not belong to the same topic.

As shown in Figure 6 (b), 49 tweets are found on average when a hashtag related to food is associated with another hashtag (generate from the word bag of food) related to food, while 8 tweets are found on average when a hashtag related to food is associated with a hashtag related a different topic.

To observe the complexity of our data cleaning method, we perform several runtimes by using 100 hashtags randomly selected beforehand. We have used a laptop with an Intel processor i7-7500U (2 cores, 4 threads) and 8gb of ram running Ubuntu 16.04. We remark that the runtime increases linearly with the number of tweets used (see Figure 7) .

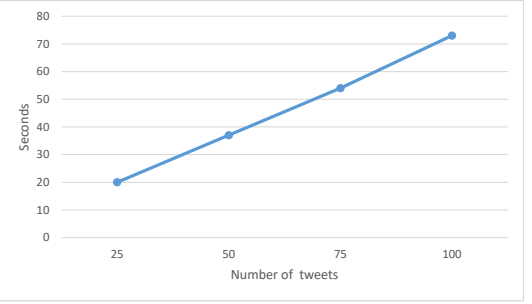


Fig. 7. Execution time in function of the number of tweets used.

4. Discussion

In this paper we have addressed the problem of hashtag context, by proposing a new and original data cleaning method that aims to extract relevant neighbours hashtags of a hashtag. As shown in the paper, the relevance here means consistency and readability.

This work as part of the more general task of cleaning and preprocessing data. Indeed, the evolution of the means of communication, and especially messages exchanged on online platforms such as Facebook or Twitter, allows today the diffusion short messages, with many abbreviations, punctuated by a lot of emoticons and dotted with hashtags and suitcase words. In this context, many data mining tasks fail when they are applied on this kind of message. Thus this cleaning work is fundamental to give meaning to the words and expressions collected and to attempt to optimise downstream data mining tasks.

Although this work only focuses on a cleaning method, the gain in meaning that provides approach could be used to address various text analysis tasks such as text classification, sentiment and opinion analysis, text clustering. Moreover, we have shown that the runtime increases linearly with the number of tweets, which could allow to address large datasets.

5. Conclusion and Perspectives

In this paper, we introduced a data cleaning methodology (independent of the language and suitable for parallel computing) to filter hashtag context. We tested our method on 520 hashtags related to 8 topics and we observed an improvement of relevance of the hashtag context. Indeed, on the one hand, we remarked that about 64% of cleaned context are more consistent than baseline context and, on the other hand, about 22% of cleaned context are more readable than baseline context. In addition, we showed that our method has a linear complexity, so may be used on large datasets. Therefore, our filter could be helpful to improve information classification or information recommendations in Twitter.

As perspectives, we plan to complete the experiments on a larger number of tweets. The goal is to consolidate results and check for the trends when the dataset is larger. In this context, we also want to study the scalability, and especially the behaviour of the method when faced with very large datasets. In the longer term, we plan to demonstrate also the interest of the approach in the improvement of the results obtained with traditional text mining tasks such as text classification, sentiment analysis, opinion mining or text clustering.

References

1. Huang, J., Thornton, K.M., Efthimiadis, E.N.. Conversational tagging in twitter. In: Proceedings of the 21st ACM conference on Hypertext and hypermedia. ACM; 2010, p. 173–178.
2. Davidov, D., Tsur, O., Rappoport, A.. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In: Proceedings of the fourteenth conference on computational natural language learning. Association for Computational Linguistics; 2010, p. 107–116.
3. Laniado, D., Mika, P.. Making sense of twitter. The Semantic Web–ISWC 2010 2010;:470–485.
4. Starbird, K., Palen, L.. Voluntweeters: Self-organizing by digital volunteers in times of crisis. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM; 2011, p. 1071–1080.
5. Garcia Esparza, S., O'Mahony, M.P., Smyth, B.. Towards tagging and categorization for micro-blogs. In: Paper presented at the 21st National Conference on Artificial Intelligence and Cognitive Science (AICS 2010), Galway, Ireland, 30 August-1 September, 2010. 2010,.
6. Kinsella, S., Wang, M., Breslin, J.G., Hayes, C.. Improving categorisation in social media using hyperlinks to structured data sources. In: Extended Semantic Web Conference. Springer; 2011, p. 390–404.
7. Posch, L., Wagner, C., Singer, P., Strohmaier, M.. Meaning as collective use: predicting semantic hashtag categories on twitter. In: Proceedings of the 22nd International Conference on World Wide Web. ACM; 2013, p. 621–628.
8. Yang, S.H., Kolcz, A., Schlaikjer, A., Gupta, P.. Large-scale high-precision topic modeling on twitter. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2014, p. 1907–1916.
9. Michelson, M., Macskassy, S.A.. Discovering users' topics of interest on twitter: a first look. In: Proceedings of the fourth workshop on Analytics for noisy unstructured text data. ACM; 2010, p. 73–80.
10. Genc, Y., Sakamoto, Y., Nickerson, J.. Discovering context: classifying tweets through a semantic transform based on wikipedia. Foundations of augmented cognition Directing the future of adaptive systems 2011;:484–492.
11. Landauer, T.K., Dumais, S.T.. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review 1997;104(2):211.
12. Ristad, E.S., Yianilos, P.N.. Learning string-edit distance. IEEE Transactions on Pattern Analysis and Machine Intelligence 1998;20(5):522–532.
13. Ferragina, P., Piccinno, F., Santoro, R.. On analyzing hashtags in twitter. In: Ninth International AAAI Conference on Web and Social Media. 2015,.
14. Li, T., Wu, Y., Zhang, Y.. Twitter hash tag prediction algorithm. In: ICOMP11-The 2011 International Conference on Internet Computing. 2011,.
15. Godin, F., Slavkovikj, V., De Neve, W., Schrauwen, B., Van de Walle, R.. Using topic models for twitter hashtag recommendation. In: Proceedings of the 22nd International Conference on World Wide Web. ACM; 2013, p. 593–596.
16. Lu, C., Lam, W., Zhang, Y.. Twitter user modeling and tweets recommendation based on wikipedia concept graph. In: Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence. 2012,.
17. Gong, Y., Zhang, Q.. Hashtag recommendation using attention-based convolutional neural network. In: IJCAI. 2016, p. 2782–2788.
18. Romero, D.M., Meeder, B., Kleinberg, J.M.. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In: WWW. ACM; 2011, p. 695–704.
19. Myers, S.A., Zhu, C., Leskovec, J.. Information diffusion and external influence in networks. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2012, p. 33–41.
20. Son, I., Lee, D., Kim, Y.. Understanding the effect of message content and user identity on information diffusion in online social networks. In: PACIS. 2013, p. 8.
21. Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A.. Twitter power: Tweets as electronic word of mouth. Journal of the American society for information science and technology 2009;60(11):2169–2188.
22. Guille, A., Hacid, H.. A predictive model for the temporal dynamics of information diffusion in online social networks. In: Proceedings of the 21st international conference on World Wide Web. ACM; 2012, p. 1145–1152.
23. Zhou, Y., Zhang, B., Sun, X., Zheng, Q., Liu, T.. Analyzing and modeling dynamics of information diffusion in microblogging social network. Journal of Network and Computer Applications 2017;86:92–102.