A topic detection and visualisation system on social media posts

Stelios Andreadis, Ilias Gialampoukidis, Stefanos Vrochidis, and Ioannis Kompatsiaris

Information Technologies Institute, Centre for Research and Technology Hellas
{andreasisst,heliasgj,stefanos,ikom}iti.gr,
WWW home page: http://mklab.iti.gr/

Abstract. Large amounts of social media posts are produced on a daily basis and monitoring all of them is a challenging task. In this direction we demonstrate a topic detection and visualisation tool in Twitter data, which filters Twitter posts by topic or keyword, in two different languages; German and Turkish. The system is based on state-of-the-art news clustering methods and the tool has been created to handle streams of recent news information in a fast and user-friendly way. The user interface and user-system interaction examples are presented in detail.

Keywords: Topic Detection and Visualisation, Twitter posts, keyword-based search, topic-based filtering

1 Introduction

The increasing number of social media users ranges from young people to the elderly people, having the need to identify and detect interesting topics and events. This is a challenging task due to the large amount of posts that are published on a daily basis. Topic detection aims at grouping together social media posts or text documents in general that discuss about the same topic-event. Topics are not considered general categories, such as politics, sports or lifestyle, but particular thematic areas and trending topics which are updated. Topic labels also need regular updates, according to the most recent documents that are collected. Fields of application for topic detection vary from media monitoring and news recommendation to content linking in the security domain.

In the social media domain, several monitoring tools have appeared, offering search by keyword, statistics and management of multiple accounts, such as Hootsuite[1] and TweetReach[2]. In addition, Social Mention[3] displays top keywords, hashtags, and sites and Twazzup[4] shows the top influencers (based on followers), the most retweeted photos, and the top 10 keywords in response to a

[1] https://hootsuite.com/
[2] https://tweetreach.com/
[3] http://socialmention.com/
[4] http://new.twazzup.com/

search by query. Other tools that allow the social media user to handle efficiently large streams of posts include IceRocket, TweetDeck, Twitonomy, Followerwonk and SumAll. Contrary to these approaches, our tool is able to cluster a selection of recent social media posts by topic, as identified from a set of concepts that are assigned to each post. Each detected topic is visualised using a cloud of concepts and named entities, as identified in the collection. Moreover, the tool offers a language option and is able to filter results by a list of suggested keywords. The tool has been developed for the purposes of the H2020-KRISTINA[5] project, building on top of other relevant topic detection services (e.g. FP7-MULTISENSOR[6]), hence the supported languages are German and Turkish.

Topic detection assumes a vector representation of a text document and is usually considered as a clustering problem [1], in absence of training sets. One of the most popular topic modeling methods is Latent Dirichlet Allocation (LDA), which requires as input the number of topics. On the other hand, density based approaches [2–4] do not require a priori knowledge of the number of clusters, but they are less effective than LDA in text clustering. Moreover, LDA has been generalised to nonparametric Bayesian approaches, such as the hierarchical Dirichlet process [5] and DP-means [6]. The estimation of the correct number of topics has been examined in [7], where the DBSCAN-Martingale has shown better performance than other state-of-the-art methods [8]. Other topic detection approaches in the social media domain involve graph-based approaches [9], where a graph clustering algorithm is applied on the graph of text documents and the decision, whether to link two social media posts or not, is based on the output of a classifier, which assigns or not the candidate items in the same cluster. Contrary to this graph-based approach, we cluster social media posts in an unsupervised way, using the DBSCAN-Martingale to estimate the number of clusters and then LDA to assign posts to topics.

Our paper is structured as follows. In Section 2 we describe the overall system architecture and we refer to the methods which are involved in the data analysis. In Section 3 we present the user interface of the topic detection tool, as well as user interaction modes with concrete usage scenarios.

2 Topic Detection System

The overall system architecture of our proposed tool is presented in Figure 1. The front page offers search by keyword as a database query. Our MongoDB database currently contains 48,589 tweets of 8 well-known journalism organizations in Germany and Turkey, crawled from May 2, 2017 up to now by exploiting the Twitter API [7]. Topic detection is a service, developed in Java, that builds on top of open source code[8], which is based on the combination of density-based clustering with LDA, as proposed in [7]. The module estimates the number of

[5] http://kristina-project.eu/en/
[6] https://www.multisensorproject.eu/
[7] https://dev.twitter.com/
[8] https://github.com/MKLab-ITI/topic-detection

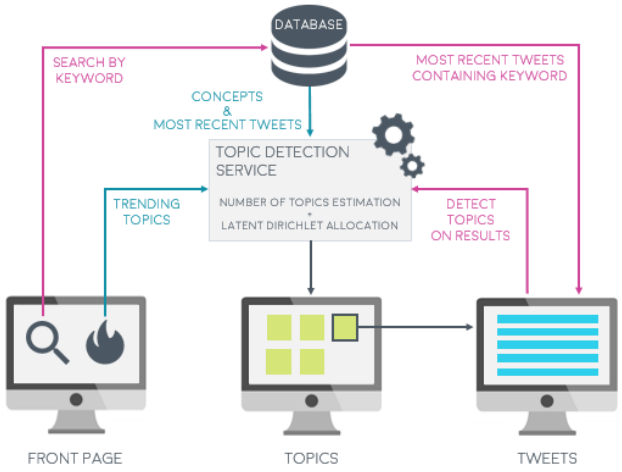

Fig. 1. System architecture

clusters (topics) and the estimation is followed by Latent Dirichlet Allocation [10], so as to assign social media posts to topics. Topic detection takes place either on the most recent (using a timeframe option) posts or on the ones that are provided by a search by keyword. The second approach has also been presented in [11], as query-based topic detection. Contrary to these approaches, the method is more efficient, since it has been recently observed that a faster version of the so called "DBSCAN-Martingale" [7] appears when sampling the density levels of DBSCAN from a skewed distribution rather than the uniform distribution [12].

3 Topic Detection Interface and Interaction Modes

Our Topic Detection Interface[9] has been designed to offer the end user an efficient and intuitive way to search for recent tweets of interest and to automatically identify topics in the resulted content. The application was implemented using common Web technologies, i.e. HTML5, CSS, JavaScript, jQuery, and PHP to communicate with the database, as well as the Kendo UI Core framework[10].

3.1 Topic detection Interface

The front page, as depicted in Figure 2, includes a one-line text input field, common to the user's experience of popular search engines. The user is able to type a set of keywords and then initiate a search for tweets that contain one or

[9] http://mklab-services.iti.gr/KRISTINA_topic_detection/
[10] https://github.com/telerik/kendo-ui-core

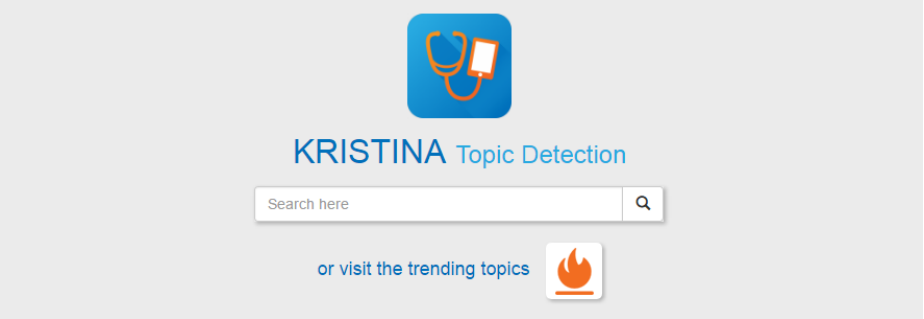

Fig. 2. Front page of the user interface

more of these keywords. Pressing Enter on keyboard or clicking the magnifier icon will navigate to the results page. Alternatively, by clicking the flame button under the search input field, the user can visit the trending topics, which will be described later in this section.

The results page, shown in Figures 3 and 4, consists of two basic components; a toolbar on top of the user interface that can be seen in detail in Figure 3 and the main panel that covers most of the page. Describing the toolbar from left to right, there is a logo that returns to the front page, a text input field to facilitate more searches and the number of results, either tweets or topics. Moreover, the toolbar includes a dropdown selection of keywords that will be explained along with the topic detection and a dropdown module to select language. The default language is German, but the user also has the option to perform search or topic detection in Turkish.

Regarding the main panel, the outcome of the search is shown as a list of white boxes, where every box represents a tweet (Figure 4). Please note here that the user interface connects directly to the database and finds the five hundred most recent posts that include any of the keywords defined by the user. For each tweet, a variety of information is presented. In detail, the main text of the post, together with images or active links if existing, the username of the author, which links to the user account's Twitter page, the date when the tweet was published, and a link to the original post in Twitter (by clicking on the grey Twitter icon). All results are sorted by their creation date, from most recent tweet to oldest.

After search results are displayed, the user is provided with two different options, in the form of two buttons on the right. The first button with the flame icon navigates to the trending topics, while the second button performs topic detection on the resulted tweets. When the latter is clicked, the topic detection service is called, receiving the ids of the tweets as input and returns the topics detected. Then, tweets are removed and the main panel is filled with topics in a grid-like view (Figure 3). Each topic is represented as a square, white box with its most frequent concepts or named entities, as extracted in the context of the H2020-KRISTINA project, in a colorful word cloud, where the more recurring terms appear in larger font. These terms are also added to the aforementioned

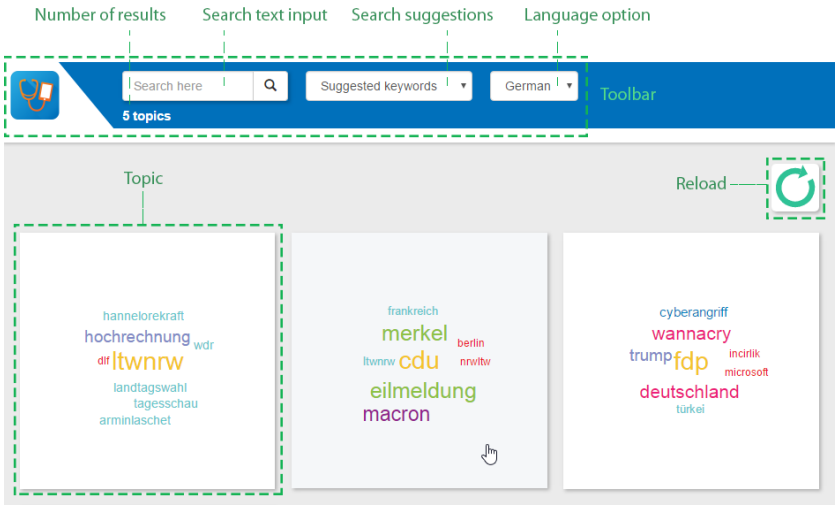

Fig. 3. Screenshot of detected topics in German; from left to right they refer to the elections in North Rhine-Westphalia (NRW), the meeting of Merkel and Macron, and the WannaCry ransomware attack.

dropdown selection in the toolbar, serving as suggested keywords to start a new search. The reload button on the top right corner can be used to call again the service and get new topics, since detection is performed dynamically. Clicking on a topic box reveals the tweets it consists of, in a manner similar to search results.

Finally, as it has already been mentioned above, another module of the user interface is the trending topics. While currently available Twitter trends refer to the most popular topics/hashtags at this given moment without further exploration criteria, the proposed system provides much more room for investigating trends by offering the option to select a target period of time, e.g. the last two days, source streams, such as interesting media and user accounts, and language, useful for trend localization. To perform this topic detection method, a variation of the previous service is invoked, passing no arguments as input (formerly set of tweet ids), but rather pre-configuring filtering options, i.e. language, user accounts and time period, in the system. The trending topics are shown exactly like the topics in Figure 3 and clicking on them navigates to the list of tweets they are composed of, displayed in the same way as previously described, e.g. in Figure 4.

3.2 Interaction Modes

To illustrate the functionality of the proposed interface, we provide a simple usage scenario. Supposing that the user is not interested in a particular topic and desires to be informed about recent news in general, he/she can begin by clicking

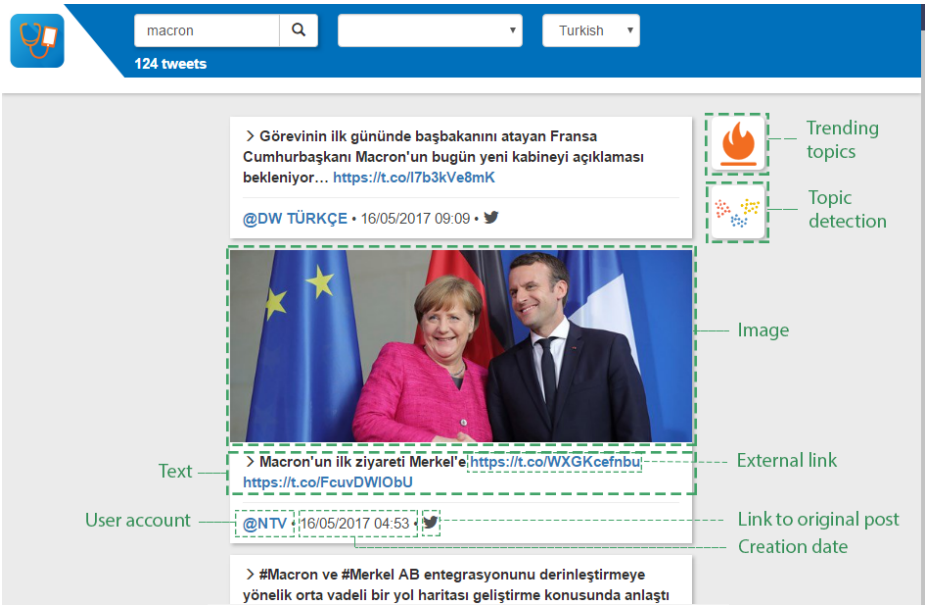

Fig. 4. Screenshot of resulted tweets in Turkish about Macron's visit to Germany

the trending topics button in the front page of Figure 2. A set of most talked-about topics on Twitter appears as in Figure 3 and the user finds interesting the visit of French president Macron to Germany[11]. After clicking on this topic, he/she is able to read all the relative tweets and follow external links for more information. In order to find out more about Macron, the user selects the term "Macron" from the dropdown list of suggested keywords and initiates a search. All tweets in database containing the name of the French president are returned and can be read. The user is curious about the popularity of Macron in Turkey, so he/she changes the language by the available option in the toolbar and clicks the magnifier icon to rerun the search. When new tweets are shown in Turkish (Figure 4), the user prefers to separate them in topics before reading all of them, thus he/she clicks the topic detection button and is presented with topics based on the previous results. A click on the topic with the word "almanya" ("Germany" in Turkish) provides the user with tweets about Macron's visit to Berlin, as posted by Turkish accounts.

4 Conclusion

We have demonstrated a topic detection and visualisation tool, that filters Twitter posts by topic or keyword, in two different languages; German and Turkish. The system offers to the user multiple navigation options to handle large streams

[11] http://edition.cnn.com/2017/05/15/world/macron-merkel-meeting/

of recent news information. The suggested keywords also help the user in his/her attempt to search for content, according to his/her interests or needs. In the future, we plan to advance this system using additional languages and other functionalities for a more user-friendly experience. Furthermore, exploiting user feedback to automatically create and update a training corpus could lead to more personalised information.

Acknowledgements

This work was supported by the EC-funded projects H2020-645012 (KRISTINA) and H2020-700475 (beAWARE).

References

1. Aggarwal, C.C., Zhai, C.: A survey of text clustering algorithms. In: Mining text data. Springer (2012) 77–128
2. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD 1996), AAAI Press (1996) 226–231
3. Ankerst, M., Breunig, M.M., Kriegel, H., Sander, J.: OPTICS: ordering points to identify the clustering structure. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, ACM (1999) 49–60
4. Campello, R.J., Moulavi, D., Sander, J.: Density-based clustering based on hierarchical density estimates. In: Advances in Knowledge Discovery and Data Mining, Proceedings of the 17th Pacific-Asia Conference (PAKDD 2013), Part II. Springer (2013) 160–172
5. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Sharing clusters among related groups: Hierarchical dirichlet processes. In: NIPS. (2004) 1385–1392
6. Kulis, B., Jordan, M.I.: Revisiting k-means: New algorithms via bayesian nonparametrics. arXiv preprint arXiv:1111.0352 (2011)
7. Gialampoukidis, I., Vrochidis, S., Kompatsiaris, I.: A hybrid framework for news clustering based on the DBSCAN-Martingale and LDA. In: Proceedings of the 12th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM 2016). Springer (2016) 170–184
8. Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., Charrad, M.M.: Package nbclust. J. Stat. Soft **61** (2014) 1–36
9. Petkos, G., Schinas, M., Papadopoulos, S., Kompatsiaris, Y.: Graph-based multimodal clustering for social multimedia. Multimedia Tools and Applications (2016) 1–23
10. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan) (2003) 993–1022
11. Gialampoukidis, I., Liparas, D., Vrochidis, S., Kompatsiaris, I.: Query-based topic detection using concepts and named entities. In: Proceedings of the 1st International Workshop on Multimodal Media Data Analytics (MMDA 2016). (2016) 9–13
12. Gialampoukidis, I., Vrochidis, S., Kompatsiaris, I., Antoniou, I.: Topic detection using the dbscan-martingale and the time operator. In: In 17th Conference of the Applied Stochastic Models and Data Analysis (ASMDA2017). (2017)