MigrAnalytics: Entity-based Analytics of Migration Tweets

Mehwish Alam[1,2⋆], Genet Asefa Gesese[1,2], Zahra Rezaie[1,2], and Harald Sack[1,2]

[1] FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany
[2] Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
{firstname.lastname}@fiz-karlsruhe.de

Abstract. This poster focuses on a visual analysis of the tweets related to European migration crisis. It uses TweetsKB as a starting point and then formulates a search criteria for extracting tweets by enriching semantic entities and hashtags starting from the seed word "Refugee". It combines European migration statistics with the information obtained by the tweets and provides visual analysis from different perspectives.

Keywords: Knowledge Graph · Migration · Visual Analytics.

CEUR-WS.org/Vol-2721/paper514.pdf

1 Introduction

Migration related data is one of the most important elements in determining the patterns causing the flow of migration from source to the host country such as poor health care system, war, poverty, etc. Moreover, another important aspect is the sentiments of the citizens living in the host countries. These sentiments, either negative or positive, could influence the prospective migrants' decisions to choose or not to choose the country as a destination. Social media has become one of the most common platforms where users including experts share their opinions. However, processing tweets leads to other kind of challenges, i.e., huge amounts of noisy data are being posted each day which is not processable by humans leading to the necessity of automated processing.

Some of the studies have targeted this problem from different perspectives such as authors in [2] used geo-tagged Twitter data of about 62,000 individuals for 6 years to estimate a set of US internal migration flows. Their findings show the relationship between short-term mobility and long-term migration. Another study [3] focuses on analyzing the social media for cyber hate towards the immigrants in Italy by using geo-tagged tweets as well as the official statistical data of Italy (ISTAT). It uses supervised classification for detecting hateful tweets. Another such resource is TweetsKB [1], a publicly available huge collection of Twitter data in RDF format on any topic. It contains more than 1.5 billion

⋆ First three authors contributed equally to this work.

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

tweets spanning from February 2013 to April 2020. In addition to metadata, the tweets are annotated with semantic entities as well as sentiment polarities. This paper introduces a tool for visual analysis of migration related tweets namely `MigrAnalytics`[3]. It uses TweetsKB as a starting point instead of crawling the whole Twitter data again for the peak migration period, i.e., 2016 and 2017. It then formulates search criteria in TweetsKB by creating and enriching a set of entities and hashtags starting from the single seed word "Refugee" and then further combines European migration statistics with the information obtained via the selected tweets followed by visual analysis from different aspects.

2 MigrAnalytics

`MigrAnalytics` follows a three step approach: **(a)** Extracting Entities and Hashtags, **(b)** Query Formulation and Migration Tweet Filtering, and **(c)** Entity-based Visual Analytics.

2.1 Extracting Entities and Hashtags

Figure 1 shows the overall process followed to extract the hashtags and entities which are used for filtering the tweets. As a starting point, "refugee" was chosen as the seed word. Since in TweetsKB, the entity "Refugee" is linked to the surface form "asylum seeker", entities and hashtags containing the word "refugee", "asylum.*seek", "seek.*asylum" as substrings, and standalone words "asylum(s)" were also extracted. Further filtering led to the extraction of more meaningful entities and hashtags such as `dbr`[4]`:Refugee_camps` and #peopleseekingasylum respectively.

In the second phase, the seed words are enriched with the help of external resources including a lexical database (WordNet[5]), an encyclopedia (Wikipedia[6]), and a knowledge graph (DBpedia [4]). All the obtained keywords are queried as both entities and hashtags against TweetsKB. WordNet is a lexical database which links synsets via relations such as hypernyms, hyponyms, meronyms, etc. `MigrAnalytics` leverages hypernym and hyponym relations to enrich the seed words related to migration. First, synsets for two of the seed words, "refugee" and "asylum", are obtained. Then the hyponyms and hypernyms of these synsets are acquired transitively. The most relevant synsets are chosen by manually analyzing the glosses of each of these synsets. In the next step, WordNet similarity measures [5] such as *path*, *Wu-Palmer's*, *jcn*, and *lin* are used to obtain most similar synsets that are also filtered for relevance based on their glosses. With the help of Wikipedia Categories, related Wikipedia pages under *Category:Refugees* are obtained. All the subcategories and pages under each of these categories are extracted recursively up to depth 8. To filter the most relevant

[3] `https://ise-fizkarlsruhe.github.io/MigrAnalytics/`
[4] `@prefix dbr: <http://dbpedia.org/resource/>.`
[5] `https://wordnet.princeton.edu/`
[6] `https://en.wikipedia.org/`

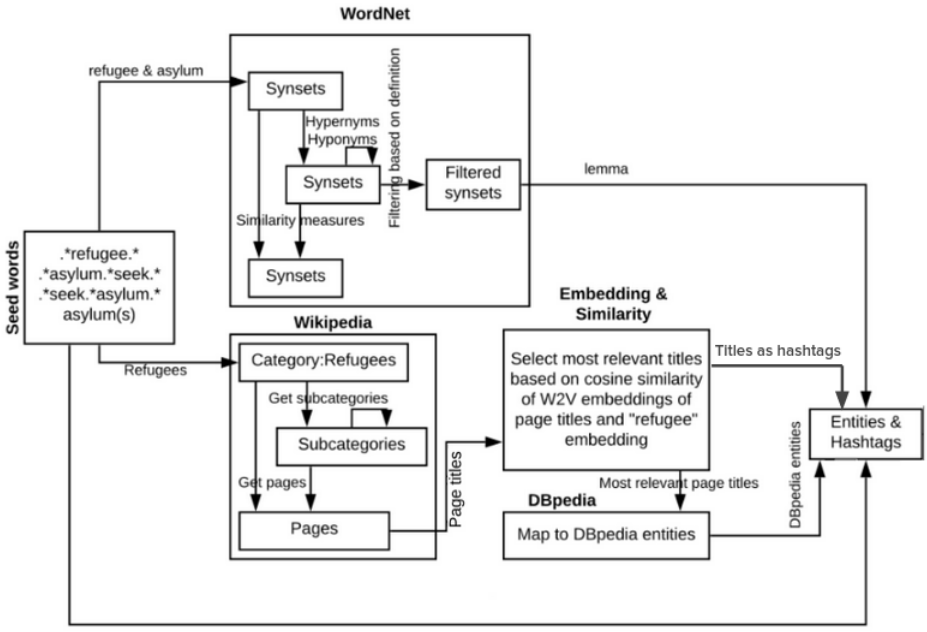

Fig. 1. Entity and Hashtag Extraction

Wikipedia page titles, pre-trained word2vec embeddings are utilized for computing cosine similarity between the seed word "refugee" and Wikipedia page titles. In pre-processing step only alphanumeric characters are kept and then lowercase conversion, stop words removal, and lemmatization are applied to page titles. The similarity threshold was chosen to be 0.5, which led to the selection of 50% of page titles (28 pages out of 56) at depth 1. For depths 2 to 5, percentage of similar page titles are 19%, 7%, 3.6%, and 1% (20 pages) respectively. For depths 6, 7, 8, number of pages with similarity greater than 0.5 is only 2, 2, and 0 respectively. Thus, Wikipedia page titles up to depth 5 has been chosen. Finally, these Wikipedia pages are mapped to corresponding DBpedia entities.

2.2 Query Formulation and Migration Tweet Filtering

Based on the entities and seed words extracted as described previously, SPARQL queries are formulated for extracting the tweets from TweetsKB. Table 1 shows the statistics of the extracted tweets. #tweets is the number of tweets extracted for each year, #entities is the number of entities contained in those tweets as annotated in TweetsKB, and finally #hashtags is the number of hashtags contained in the extracted tweets.

	Total (2016)	Distinct (2016)	Total (2017)	Distinct (2017)
#tweets	197,813	197,813	208,492	208,492
#entities	340,694	23,261	371,944	24,009
#hashtags	238,545	29,756	172,327	28,135

Table 1. Statistics of the information extracted from TweetsKB.

2.3 Entity-based Visual Analytics

Various plots are used to visualize the interactions between the number of tweets regarding refugees along with the hashtags and entities. It also considers the relationship between the tweets extracted in the previous steps and the number of asylum applications during the period of peak migration crisis[7].

The total number of first time asylum applications in EU28 in year 2016 and 2017 were 1,204,280, and 649,855 respectively[8]. Monthly figures for each year were rather steady; however, in 2016 EU received almost twice as many monthly applications as in 2017.

First, the top 20 entities and hashtags in terms of number of occurrences are selected separately for each year. Then, these entities and hashtags are ranked and depicted based on their frequencies on a weekly basis. Among the top 20 entities and hashtags for the year 2016, 7 and 6 of them are terms that co-occurred with the keywords used in the query, respectively. They include relevant countries, politicians, political events, and so on. For example, the term United_Kingdom_withdrawl_from_the_Europen_Union appears as an entity and #brexit as a hashtag. Both of them refer to the same political event during 2016 which could indicate that Brexit has a significant impact on migrant crisis matter. Among the top 20 entities and hashtags for the year 2017, 7 and 9 of them are terms co-occurred with the keywords used in the query, respectively. Several of these co-occurring terms are related to US political issues regarding migrants, e.g., Executive_order, Deferred_Action_for_Childhood_Arrivals or its equivalent hashtag #daca, #nobannowall, and #muslimban. Finally, in order to plot a word cloud of entities and hashtags, top 100 of them (in terms of frequency) were chosen over the course of each week. For example, as shown in the plot, "Immigration" and "Refugee" are some of the words which are among the most frequent entities and hashtags.

3 Discussion and Perspectives

The current study provides an entity-based analysis over the migration related tweets by using European Migration Statistics. As a perspective, the experts related to migrations will be determined on social media and analysis of their views on factors causing migration will be performed. Moreover, the full text of the tweets will also be processed for extraction and analysis purposes.

[7] These visualizations are shown on the associated homepage.

[8] `https://ec.europa.eu/eurostat`

References

1. Fafalios, P., Iosifidis, V., Ntoutsi, E., Dietze, S.: TweetsKB: A public and large-scale rdf corpus of annotated tweets. In: Extended Semantic Web Conference (ESWC'18). Heraklion, Crete, Greece, (2018)
2. Fiorio, L., Abel, G., Cai, J., Zagheni, E., Weber, I., Vinué, G.: Using Twitter data to estimate the relationship between short-term mobility and long-term migration. In: WebSci (2017)
3. Florio, K., Basile, V., Lai, M., Patti, V.: Leveraging Hate Speech Detection to Investigate Immigration-related Phenomena in Italy. In: 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) (2019)
4. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al.: Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. Semantic web **6**(2), 167–195 (2015)
5. Pedersen, T., Patwardhan, S., Michelizzi, J., et al.: Wordnet:: Similarity-measuring the relatedness of concepts. In: AAAI. vol. 4, pp. 25–29 (2004)