Temporal Evolution of the Migration-related Topics on Social Media*

Yiyi Chen, Genet Asefa Gesese, Harald Sack, and Mehwish Alam

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

Abstract. This poster focuses on capturing the temporal evolution of migration-related topics on relevant tweets. It uses Dynamic Embedded Topic Model (DETM) as a learning algorithm to perform a quantitative and qualitative analysis of these emerging topics. TweetsKB is extended with the extracted Twitter dataset along with the results of DETM which considers temporality. These results are then further analyzed and visualized. It reveals that the trajectories of the migration-related topics are in agreement with historical events. The source codes are available online: https://bit.ly/3dN9ICB.

1 Introduction

Social media has become one of the most widely used channels for people to exchange opinions about social events around the globe. It provides one of the most useful resources about social interactions and trending events on important topics such as migration, climate change, political elections, etc. Over the last decade, migration has become one of the most controversial topics in Europe. This poster analyses the Twitter data from the destination countries in Europe, to gain insight into migration-related events. The analysis of the temporal trajectory of the topics in migration-related tweets shows the shifts in the attention of people over time.

For example, people's interest in "Syrian refugees" peaked in 2016 in Europe, while later it decayed. With the emergence of COVID-19 in 2020, the attention shifted to "border control". By analyzing the co-occurring topic words, the driving factors could be identified. As shown in Fig. 2, from 2020 onwards the probability trajectories of co-occurring words "COVID", "migrant", "border" have similar trends, which indicates the pandemic as a causal factor of debates about border control, and further affecting the migration flows in this period.

Several efforts have been made to provide a Knowledge Base (KB) about tweets to make the data more accessible and usable for further research. One

* This work is a part of ITFLOWS project which has received funding from the EU H2020 research and innovation program under grant agreement No 882986.

such effort is TweetsKB [6], which is a KB containing 1.5 billion tweets spanning through 5 years, including entity and sentiment annotations. MigrAnalytics [1] is another such effort which uses TweetsKB as a starting point to analyze migration-related tweets using entity-based approaches. While both provide time-series data, no attempt has been made to analyze the temporal evolution of migration-related topics, which would help us to identify the changing driving factors of migrations in a temporal dimension. This study completes MigrationsKB (MGKB) [3] with the tweets including topics evolving through time using advanced methods based on neural networks.

2 Temporal Evolution of Topics in Migrations

Tweet Extraction. Keyword-based methods are used to extract tweets from Twitter. The words "immigration" and "refugee" are used as the seed words which are then enriched with top-50 most similar words with pre-trained Word2Vec and fastText embeddings. The final set of keywords are then manually verified. The collected tweets are from 11 destination countries, where most refugees in Europe are hosted. These countries are selected by ranking them according to the frequency of the asylum seekers obtained from Eurostat[3]. These countries include the United Kingdom, Germany, Spain, Poland, France, Sweden, Austria, Hungary, Switzerland, Netherlands, and Italy. The final number of extracted tweets spanning over 7 years (2013 - 2020) is 384891 which are then preprocessed by removing user mentions, reserved words (i.e., RT), emojis, smileys, URLs, stop words, punctuations, numerical tokens, HTML tags. Moreover, the tokens except hashtags are lemmatized. Finally, the tweets with at least two words are retained.

Dynamic Topic Modeling. Topic modeling is used to extract hidden semantics in textual documents. The most widely used algorithm for topic modeling is Latent Dirichlet Allocation (LDA) [2]. However, it fails in the presence of large vocabularies. Embedded Topic Model (ETM) [5] aims to solve this problem by employing word embeddings. It defines each topic as a vector on the word embedding space, and then represents the per-topic proportion as joint information from words and topic embeddings. DETM [4] is developed to extend ETM, which uses a probabilistic time series to model the topics varying over time. Similar to ETM, the probability of each word under DETM is a categorical distribution whose parameters depend on joint information from word embeddings and per-topic proportion, however, the topic proportions vary over time.

In DETM, the word and the topic embeddings are trained in parallel on the extracted Twitter data with a time variable on a predefined number of topics. The data is split into the train (85%), validation (5%), and test (10%) (i.e., 306077, 18011, and 35766 tweets respectively) with a vocabulary size of 20865. DETM is optimized on a document completion task [7] and the best model is used to generate topics from the tweets. In [4], the authors choose 50 as the number

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

of topics. For the sake of completion, the experiments are also conducted with 25 and 75 topics in this study. However, after applying the pretrained DETM on the extracted Twitter data, many redundant topics are found, i.e., not being classified to the tweets. Hence, the models with lower numbers of topics are applied. As shown in Table 1, only models trained with 5 and 10 topics have all the topics classified to the tweets, which are used for further analysis.

Table 1: The Results of DETM with Different Number of Topics

Nr. Topics	**5**	**10**	**15**	**20**	**25**	**50**	**75**
Val PPL	2938.2	2757.5	2736.8	2702.1	2698.4	**2638.9**	2741.7
Test PPL	2760.7	2706.1	2638.9	**2594.8**	2647.8	2637.8	2740.8
Topic Diversity	**0.773**	0.7545	0.6953	0.6598	0.6666	0.6951	0.6105
Topic Coherence	**0.3724**	0.2533	0.2453	0.2030	0.1913	0.0619	0.1299
Topic Quality	**0.2878**	0.1911	0.1706	0.1340	0.1275	0.0430	0.0793
Classified Nr. of Topics	5	10	12	14	14	32	37

The trained word embeddings from DETM represent the similarity between the words in the tweets. To select the tweets that are relevant to the topic of migration, the centroid of the word vectors of keywords is clustered with K-Means, and then the cosine similarity between the centroid and the trained topic proportions of tweets is calculated (i.e., the confidence score). The relevant tweets have a confidence score greater than 0. As shown in Fig. 1a, for both models, the scores are normally distributed, while the topic proportions of the tweets are trained under a Gaussian noise [4]. That means, the distance between the tweets and the centroid has the same distribution as the topic proportions. Therefore, "migration" is the central topic of the tweets.

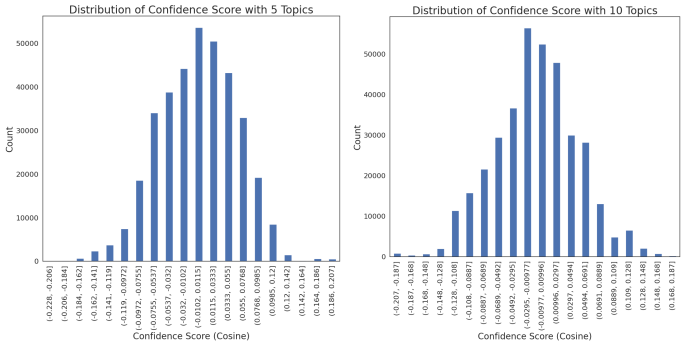
In Fig. 1b, the distribution of the tweets stays almost the same across time and across topics. However, in Fig. 1c, after filtering the non-relevant tweets, the tweets of the 5th topic are filtered out, while the tweets of the 8th topic are more evenly distributed across time. After the automatic selection, since the model trained with 10 topics provide richer topics across the time, the model is selected for further analysis. Fig. 1d shows the distribution of the non-relevant tweets per country.

Fig. 2 shows the evolution of word probability across time for four different specific topics (i.e., the titles for the plots) learned by DETM. For each, a set of words whose probability shift aligns with historical events are presented. For example, for "Syria, Refugee, and EU", the probability of the words shift in the same manner from 2014 to 2017, reflecting the event of a large amount of Syrian refugees entering the European Union, which peaked from 2015 to 2016. In 2018, the Brexit decisions were made, at the same time, the media outlets claimed a strong connection between Brexit and Trumpism, which is also reflected in "Brexit and Refugee". The discussion of the border has a strong correlation with the emergence of COVID-19 since 2020, as shown in "Migrant and Border".

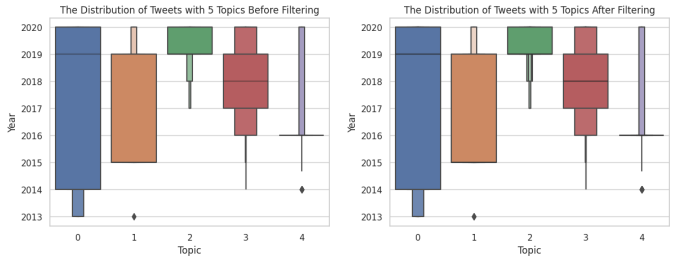
Extending TweetsKB. The tweets along with the assigned topics are stored in an extension of TweetsKB. Moreover, the words associated with each topic

Fig. 1: Tweets assigned with Topics

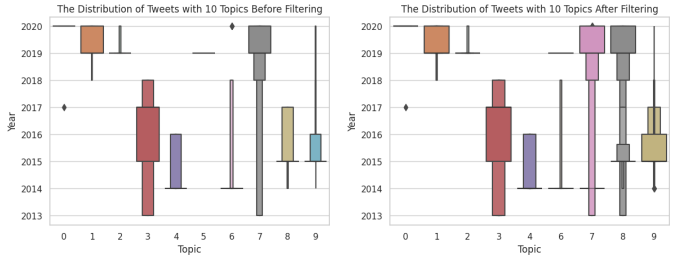
(a) Distribution of Confidence Score for Models with 5 and 10 Topics



(b) The Distribution of Tweets before and after Filtering for Model with 5 Topics



(c) The Distribution of Tweets before and after Filtering for Model with 10 Topics



(d) The Distribution of Non/Relevant Tweets per Country

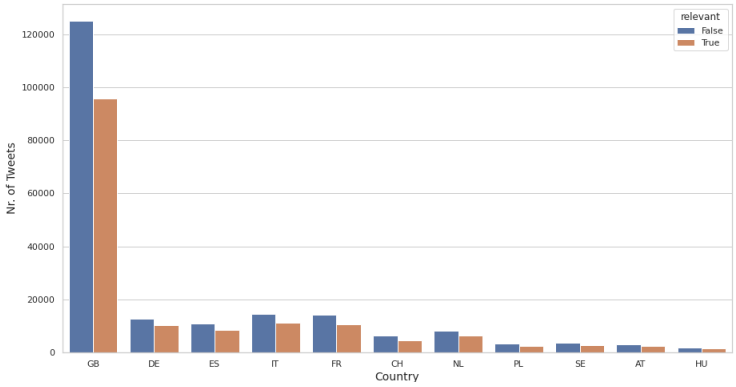
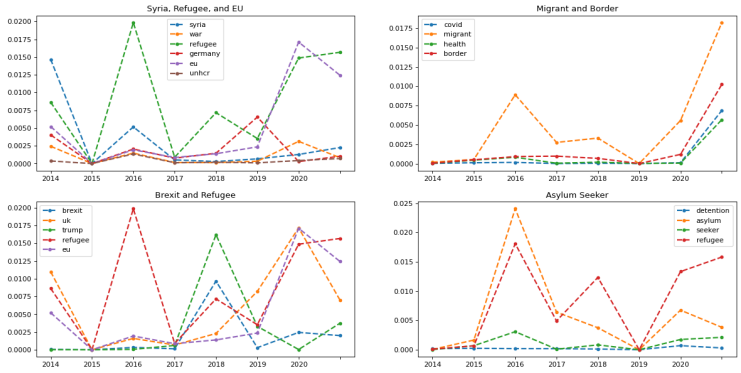


Fig. 2: Evolution of Word Probability Across Time for different Topics



across time are also added using new classes and properties, please refer to the GitHub page for more details. This information can be queried using SPARQL. The following query shows the evolution of words of a given topic.

```
SELECT ?TopicWords ?year WHERE{
?topic sioc:id "7"; mgkb:topicOccur ?spTopic.
?spTopic dc:date ?year. schema:description ?TopicWords.
}ORDER BY DESC(?year)
```

3 Discussion and Future Work

This study focuses on capturing the evolution of topics in migrations-related tweets in the destination countries. It uses the time-aware topic modeling method, DETM, for achieving this goal. The results are then populated in extended TweetsKB where the temporal dimension is represented with the help of literals (date values) for further analysis. As a future study, an extensive manual verification on the chosen topics will be conducted, and the current RDF schema will be extended with RDF-Star.

References

1. Alam, M., Gesese, G.A., Rezaie, Z., Sack, H.: Migranalytics: Entity-based analytics of migration tweets. CEUR workshop proceedings **2721**, 74–78 (2020)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**(null), 993–1022 (Mar 2003)
3. Chen, Y., Sack, H., Alam, M.: Migrationskb: A knowledge base of public attitudes towards migrations and their driving factors (2021)
4. Dieng, A.B., Ruiz, F.J.R., Blei, D.M.: The dynamic embedded topic model. CoRR **abs/1907.05545** (2019), http://arxiv.org/abs/1907.05545
5. Dieng, A.B., Ruiz, F.J.R., Blei, D.M.: Topic modeling in embedding spaces (2019)
6. Fafalios, P., Iosifidis, V., Ntoutsi, E., Dietze, S.: Tweetskb: A public and large-scale RDF corpus of annotated tweets. CoRR **abs/1810.10308** (2018)
7. Wallach, H.M., Murray, I., Salakhutdinov, R., Mimno, D.: Evaluation methods for topic models. In: International Conference on Machine Learning (2009)