CSSM 530: Automated Text Processing for Social Sciences Dr. Ali Hürriyetoğlu, June 2022. (), , 1–14
doi:

RESEARCH NOTE

Anti-refugee discourse on Twitter: GSDMM (Gibbs Sampling Dirichlet Multinomial Mixture), Latent Dirichlet Allocation (LDA) and LSI (Latent Semantic Indexing) for Turkish anti-refugee tweets topic modelling

Işık Topçu*

Computational Social Sciences MA, Koç University, TR
*Corresponding author. Email: itopcu21@ku.edu.tr

Abstract

A dataset of 42.976 Turkish tweets between (May 2021-May 2022) were scraped using a list of keywords related to refugees in order to perform sentiment analysis. A qualitative inspection of 500 preprocessed tweets showed a potential overweighting of negative sentiment in a potential sentiment analysis. Topic modeling was performed to inspect a wider spectrum of topics rather than simply acquiring average polarity scores of tweets. GSDMM (Gibbs Sampling Dirichlet Multinomial Mixture), LDA (Latent Dirichlet Allocation) and LSI (Latent Semantic Indexing) were performed on a subset of 10.000 tweets. The data generating process inspires further study such as emotion detection, to be applied to capture a wider range of emotions in Tweets related to refugees rather than polarity analysis.

Keywords: computational social sciences, natural language processing, anti-refugee speech, twitter API, preprocessing, topic modeling, gibbs sampling dirichlet multinomial mixture, latent dirichlet allocation, latent semantic indexing

1. Topic modeling

Topic modeling is a collection of algorithms used to find hidden topics contained in a document (texts, tweets, emails, books etc.). Linear modeling (e.g., Latent semantic analysis (LSA)) and probabilistic topic modeling are widely used types of topic modeling. LDA(Latent Dirichlet Allocation) and Probabilistic Latent Semantic Analysis (pLSA) are both examples of probabilistic topic modeling. GSDMM (Gibbs Sampling Dirichlet Multinomial Mixture) is a modified LDA technique that assumes that one subject equals one document at the start. The words within a document are generated using the same unique topic, and not from a mixture of topics as it was in the original LDA. Therefore, GSDMM is considered to be a good choice for short text topic modeling for many reasons such as:

1. GSDMM can infer the number of clusters automatically,
2. GSDMM has a clear way to balance the completeness and homogeneity of the clustering results,
3. GSDMM is fast to converge,

4. Unlike the Vector Space Model (VSM)-based approaches, GSDMM can cope with the sparse and highdimensional problem of short texts,

5. Like Topic Models (e.g., pLSA and LDA), GSDMM can also obtain the representative words of each cluster.

LDA and LSI models vary in the fact that while the connection between documents in a corpus is ignored by the LSA model, the LDA may be used to also examine the relationships between documents in a corpus. LDA aims to determine a likelihood of hidden distributions in the input data. LDA is basically a "bag of-words" from a functional standpoint, which implies that the order of the words is irrelevant, but it assumes that:

1. Documents(tweets) with similar terms are frequently about the same thing.

2. Documents(tweets) with groupings of words that appear often together typically have the same subject.

In this research, we focus on the usage of GSDMM, LDA and LSI[1] on gathering latent topics in Turkish refugee-related tweets.

2. Anti-refugee sentiment

Finding latent meanings in tweets with topic modeling can be beneficial for social sciences. Although Twitter does not fully represent the populations; gathering fast and automated information on ever-changing political dynamics can be a principal policy making and sentiment mining tool. Traditional data sources to monitor public opinion are often limited, notably due to slow and expensive collection. Data science appears to provide a faster and more efficient alternative to data collection and processing for social sciences. One of the main research areas of that benefit from data science is refugee studies with the help of the existing metadata especially in host countries. The fast growth in the number of refugees fleeing from African and Middle Eastern nations as a consequence of persecution, war, violence, and human rights violations, as well as the debate about their admittance in Europe and Asia, is considered a refugee crisis by the international community. Although Turkey has become host for many war-ridden countries' citizens such as Iraq, Iran and Afghanistan for decades, the Turkish refugee crisis hit its peak with the Syrian Civil War which started in 2011, resulting in a total of approximately 3.7 millions of refugees and asylum seekers in Turkey by 2022 of whom 46 percent are children. (Sahin et al., UNHCR) Turkey, to this day, remains to be the leading host and a commonly used passageway of Syrian refugees to European countries. Although many refugees seek better life conditions in Europe, many others who are in the labor force and who built their lives and careers in Turkey also plan a future in Turkey. According to a 1.900 participant survey done by Duvell et al. 56% percent of the refugee participants would like to continue living in Turkey. The process of integration of refugees remains an undeniable paradox and the refugee acceptance in the public sphere remains a hot topic. Integration, an already difficult process, can be fueled by social, economic and political dynamics. Analyzing the public opinion on what fuels such sentiments might help governments and non-governmental organizations to take necessary measures such as funding, marketing or other governmental measures for the refugees wishing to rebuild their lives in Turkey. Monitoring the topics the public sphere is discussing whether they are economic, social or racial and connecting them to certain emotions might help governments visualize their future steps.

3. Literature Review

Topic modeling is frequently used in many fields of social sciences to gather an idea of the public sphere. Along with topic modeling, NLP methods such as sentiment analysis were used to capture the public opinion on the European Refugee Crisis. Calderon et al., sought to model and characterize hate speech against immigrants on Twitter in Spain around the appearance of the far-right party

1. LSI and LSA are used synonomously.

Vox. They used unsupervised topic modeling and found that the four underlying topics (control of illegal immigration, economic assistance for immigrants, consequences of illegal immigration, and Spain as an arrival point for African immigrants and Islamist terrorism) were similar to those in the discourse of Vox. Yan et. al. used non-negative matrix factorisation to apply topic modeling on bulks of tweets to discover the hidden patterns within these social media discussions. Inumwa-Dutse et. al., presented an analysis of opinions on migrants and refugees expressed on Twitter using sentiment analysis and topic modeling (Latent Semantic Analysis (LSA) in a cross-country context. Ozturk et al. investigated the public opinions and sentiments towards the Syrian refugee crisis with text mining techniques on related tweets. Erdogan and Guler have analyzed a subset of tweets that included the hashtag *ulkemdesuriyeliistemiyorum* (idontwantsyriansinmycountry) to understand its functions in constructing and proliferating an exclusionary discourse against refugees. Ozturk et al. investigated the public opinions and sentiments towards the Syrian refugee crisis with extensive sentiment analysis technique on related tweets. Rowe et al. measured shifts in public sentiment opinion about migration during early stages of the COVID-19 pandemic via lexicon-based sentiment analysis and topic modeling in Germany, Italy, Spain, the United Kingdom, and the United States. Flores (2017) conducted a quasi-experimental design to assess the impact of a more restrictive immigration policy on rises in anti-immigration attitudes in Arizona. Bartlett and Norrie (2015) studied explored online conversations relating to immigration,the frequency of terms and how they changed over time in online talks about immigration. Previous research looked at how Twitter may be utilized to better understand the experiences of migrants in Europe during the 2015–2017 refugee crisis (e.g., Gualda and Rebollo, 2016; UN Global Pulse, 2017). Recent research (Freire-Vidal and Graells-Garrido, 2019; Freire-Vidal et al., 2021) has looked at how changes in migration sentiment are linked to certain emotions, as well as how positive and negative immigration sentiment communities may be discovered by studying their retweet networks. Several research (Bosco et al., 2017; Sanguinetti et al., 2018; Basile et al., 2019; Comandini and Patti, 2019; Calderón et al., 2020) have used Twitter data to investigate hate speech directed against immigrants.

4. Data

Twitter API between May 2021 and May 2022 was scraped using the SNS TwitterScraper module (https://github.com/JustAnotherArchivist/snscrape). Since the scraper didn't include tweets with different suffixes of the Turkish language, a keyword list containing all the different suffix variations as well as english spelling of "göçmen", "mülteci" and "sığınmacı" was fed into it. Since these words with suffixes were in Turkish, a specification of the language wasn't necessary. A dataset of 42.976 refugee related tweets were acquired in a .json file and saved into a coma separated value file. An initial labeling of 500 tweets showed an overwhelming dominance of negative sentiment. A quick TextBlob sentiment analysis was also performed on a random subset of 2000 tweets to get a grip on the overall sentiment before performing an in depth analysis. This created an intuition that the training data for the sentiment analysis would be unbalanced if used. Although there might be ways to handle this either with a Lexicon based approach or by balancing the training data, an emotion detection where the tweets will be labeled with 7 emotions (Anger, Contempt, Fear, Disgust, Happiness, Sadness and Surprise.) might be much more beneficial for future work. The dataset was scraped using the same keyword list, setting the "maximum tweet" query in the scraping algorithm for every keyword 700 respectively.

5. Preprocessing

The dataset was pre-processed in two steps: first, the initial cleaning consisted of removing capital letters, punctuation, numbers, @users, hashtags, hyperlinks (http:, www.), emojis and tokenization using necessary regular expressions. At this point, tweets were a list of tokens. The second part of the pre-processing consisted of the elimination of Turkish stopwords, normalization and lemmatization.

```
keywords = ["mülteci","mülteciyi","mülteciye","mültecide","mülteciden",
            "mülteciler","mültecileri","mültecilere","mültecilerde", "mültecilerden",
            "multeci", "multeciyi","multeciye","multecide","multeciden",
            "multeciler", "multecileri", "multecilere", "multecilerde","multecilerden",

            "mültecinin","mültecimiz",
            "mültecilerin","mültecilerimiz",
            "multecinin","multecimiz",
            "multecilerin","multecilerimiz",


            "göçmen", "göçmeni", "göçmene","göçmende", "göçmenden",
            "gocmen", "gocmeni", "gocmene", "gocmende", "gocmenden",
            "göçmenler", "göçmenleri","göçmenlere","göçmenlerde","göçmenlerden",
            "gocmenler", "gocmenleri", "gocmenlere", "gocmenlerde","gocmenlerden",

            "göçmenin","göçmenimiz",
            "gocmenin","gocmenimiz",
            "gocmenlerin","gocmenimiz",
            "göçmenlerin","göçmenlerimiz",


            "sığınmacı","sığınmacıyı","sığınmacıya","sığınmacıda","sığınmacıdan",
            "sığınmacılar", "sığınmacıları", "sığınmacıya", "sığınmacıda", "sığınmacıdan",
            "siginmaci", "siginmaciyi","siginmaciya","siginmacida","siginmacidan",
            "siginmacilar", "siginmacilari","siginmacilara","siginmacilarda","siginmacilardan",

            "sığınmacının","sığınmacımız",
            "siginmacinin","siginmacimiz",
            "sığınmacılarının","sığınmacılarımız",
            "siginmacilarinin","siginmacilarimiz"]
```

Figure 1. Keyword List for TwitterScraper

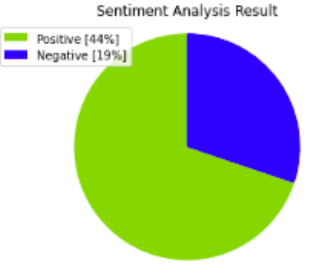

Figure 2. Sentiment Analysis for 2000 random tweets with TextBlob

Zemberek library's TurkishSentenceNormalizer(morphology) and Zeyrek library's MorphAnalyzer() were used. (NLTK Turkish Stemmer didn't work as great.) Turkish stopwords dictionary from NLTK was used to remove all stopwords. The token frequencies were visualized using WordCloud. Top bi and tri-grams were visualized.

```
['uyku', 'ülke', 'tane', 'bile', 'mülteci', 'istemek', 'sayıklamak', 'uyumak', 'bile', 'rahat', 'değil', 'puh']
```

Figure 3. Pre-processed tweet

6. Topic Models

6.1 GSDMM

For the GSDMM, the lemmatized data that was saved in a .csv file was read and inspected. n-grams for the GSDMM were created. Default MovieGroupProcess() algorithm hyperparameters that would work for many short texts were set to:

1. alpha =0.01 1 and beta (eta) = 0.01.
2. *n – iters* = 30. (Number of iteration.)

3. K = 6. This is the number of clusters. We set this value after several experiments in which we started from 15 clusters. As we increased the number of clusters, the empty clusters started to appear.

Eventually, we came up with 6 clusters of topics. We didn't name the exact topics. For the evaluation of the GSDMM model, we used the coherence score metric. The coherence score for the final GSDMM model is 0.32 which needs to be improved. We have also illustrated the dominant words for each topic with WordCloud.

```
Number of documents per topic : [ 538  353  209 2447 2066 4387]

Most important clusters (by number of docs inside): [5 3 4 0 1 2]

Cluster 5 : [('mülteci', 5130), ('bir', 1655), ('ülke', 1600), ('olmak', 1083), ('demek', 846), ('desen', 732), ('va
r', 684), ('yok', 662), ('mi', 654), ('bu', 615)]

Cluster 3 : [('mülteci', 2875), ('ülke', 1083), ('bir', 852), ('olmak', 653), ('mi', 412), ('var', 396), ('değil', 38
1), ('desen', 370), ('demek', 365), ('yok', 336)]

Cluster 4 : [('mülteci', 2568), ('bir', 917), ('ülke', 770), ('olmak', 582), ('demek', 399), ('değil', 377), ('dese
n', 344), ('var', 332), ('bu', 316), ('mi', 316)]

Cluster 0 : [('mülteci', 659), ('ülke', 183), ('bir', 106), ('suriye', 99), ('olmak', 98), ('milyon', 92), ('para', 7
4), ('var', 72), ('türkiye', 72), ('vermek', 71)]

Cluster 1 : [('mülteci', 405), ('bir', 76), ('ülke', 62), ('demek', 61), ('kendi', 52), ('vatandaş', 52), ('olmak', 5
1), ('suriye', 45), ('milyon', 45), ('mi', 44)]

Cluster 2 : [('mülteci', 287), ('ülke', 77), ('bir', 65), ('değil', 56), ('desen', 51), ('olmak', 45), ('olur', 31),
('milyon', 30), ('almak', 28), ('ukrayna', 28)]
```

Figure 4. GSDMM Clusters

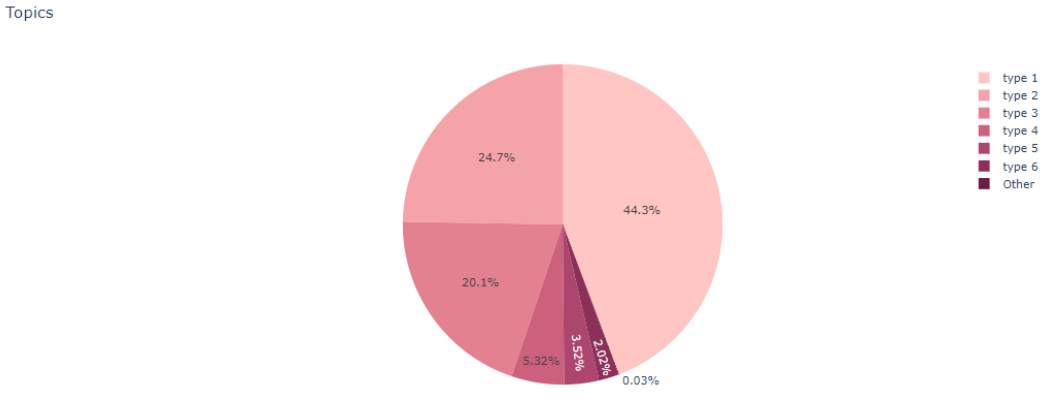

Figure 5. GSDMM Chart

6.2 LDA, LDAMulticore, LSI

For the LDA models, first of we tried implementing the original LDA algorithm without the TF-IDF transformation on the raw lemmatized data. The Gensim *LDAmodule* was imported. While creating the LDA model, we need the dictionary and corpus as input. We created a gensim dictionary of the pre-processed tweets. Unlike a python dictionary, the Gensim library treats words in text data as tokens, assigning each one an index number, which is then stored in the dictionary as a "ID." A dictionary, in other terms, is a structure that holds words and their "ID" numbers. The IDs of these

	Text	Topic	Rating	Lemma-text	len
0	uyku " ülke tane bile mülteci istemek sayıklam...	type 1	NaN	[uyku, ülke, tane, bile, mülteci, istemek, say...	13
1	ülke mülteci istilâ kurtarmak	type 6	NaN	[ülke, mülteci, istilâ, kurtarmak]	4
2	lan mülteci yerleşmek amk ö kadar mi bakmak mü...	type 3	NaN	[lan, mülteci, yerleşmek, amk, kadar, mi, bakm...	35
3	sığınmak göçmenmülteci geçmek yapmak sonra akp...	type 2	NaN	[sığınmak, göçmenmülteci, geçmek, yapmak, sonr...	14
4	biz arkadaş arabistan dönmek ülkemülteci kamp mi	type 2	NaN	[biz, arkadaş, arabistan, dönmek, ülkemülteci,...	7

Figure 6. GSDMM Topics DataFrame

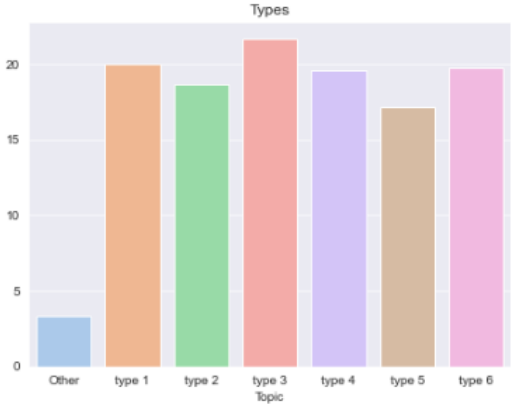

Figure 7. Topic Dispersion with GSDMM

Figure 8. Dominant Keywords: Topic 1

terms and their frequency in the text data make up the corpus. As previously stated, we used the dictionary, num-topics (the total number of topics we constructed for the whole corpus), and corpus as inputs while building our model. With the num-words argument, we printed the words with the greatest rate of frequency for each topic. A "document-topic matrix" and a "topic-word matrix" are

Figure 9. Dominant Keywords: Topic 2

Figure 10. Dominant Keywords: Topic 3

created using the LDA method. The probabilities of the distribution of words into topics are kept in the topic-word matrix.The topic-word matrix has 8 topics and words that best characterize each topic, as seen in the output. We can make assumptions about the terms that rely on it in every area by looking at this. The "document-topic matrix allows for the allocation of documents into different themes. PyLDAvis was created to assist users in interpreting topics in a topic model as they relate to a set of text data. It displays information from the LDA topic model in an interactive web-based visualization. As we can see, in the first LDA model, even though the best coherence topic number and the best parameters were chosen[2] these hypterparameters were fed to the first LDA., there is an

2. the highest coherence score: n : 10 ; alpha : symmetric ; beta : 0.7 ; Score : 0.7011236056887038

Figure 11. Dominant Keywords: Topic 4

Figure 12. Dominant Keywords: Topic 5

overlapping of topics. There might be many reasons for this. Therefore, we tried the LDAMulticore, which is a faster alternativce to LDAModel. The LDA module in gensim is very scalable, robust, well tested by its users and optimized in terms of performance, but it still runs only in single process, without full usage of all the cores of modern CPUs. A multicore implementation can be very useful and might save a lot of time waiting, especially for very large corpora. LDAMulticore basically uses all CPU cores to parallelize and speed up model training.

The LDAMulticore model was initiated, and the models with the best number of topics were created and a chart illustrating the best coherence score for every possible number of topics was

Figure 13. Dominant Keywords: Topic 6

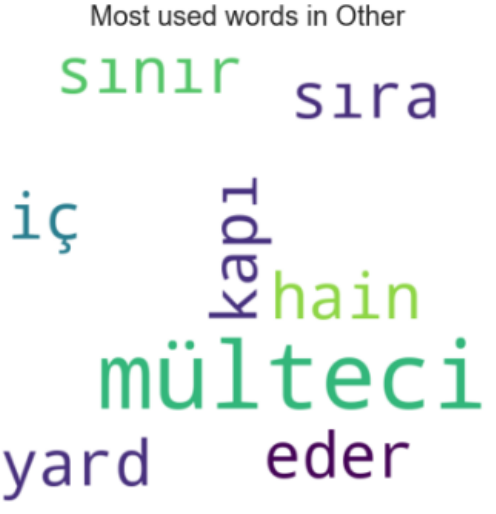

Figure 14. Dominant Keywords: Other

created. Then, by using the GridSearchCV, the hyperparameters of the models with the optimal number of topics (highest coherence score which can be seen in Figure.20) were calculated. These parameters were later fed into the LDAMulticore model with the optimal number of topics to get a coherence score of 0.3 which needs improving. For the LSI model, we found the optimal number of topics with a chart that illustrates the best coherence scores. After setting the model on the highest coherence score, we acquired 13 topics as seen on the chart.

```
pprint(ldamodel.print_topics())
```

```
[(0,
  '0.030*"ırk" + 0.016*"ukrayna" + 0.009*"aşı" + 0.008*"nefret" + '
  '0.007*"başlamak" + 0.006*"eğitim" + 0.005*"faşist" + 0.003*"taliban" + '
  '0.002*"el_kol" + 0.002*"slâm"'),
 (1,
  '0.016*"an" + 0.009*"yerinmek" + 0.007*"hoca" + 0.004*"statü" + 0.003*"bölü" '
  '+ 0.002*"smail" + 0.002*"batı" + 0.002*"twitter" + 0.002*"nema" + '
  '0.001*"the"'),
 (2,
  '0.007*"silâh" + 0.006*"dışarı" + 0.006*"evlât" + 0.005*"saat" + '
  '0.004*"kimlik" + 0.004*"te" + 0.003*"öz" + 0.003*"mil" + 0.002*"dağ" + '
  '0.002*"tıkmak"'),
 (3,
  '0.017*"turist" + 0.008*"pkk" + 0.007*"terörist" + 0.004*"test" + '
  '0.003*"maske" + 0.003*"yaka" + 0.003*"mamoğlu" + 0.002*"virüs" + '
  '0.002*"masa" + 0.002*"bulaşmak"'),
 (4,
  '0.010*"savaş" + 0.004*"büyümek" + 0.002*"bakarım" + 0.002*"matrix" + '
  '0.001*"dinamik" + 0.001*"tv" + 0.001*"geçilmiyor" + 0.001*"youtube" + '
  '0.001*"giriyorlar" + 0.001*"danimarka"'),
 (5,
  '0.011*"stanbul" + 0.010*"müslüman" + 0.008*"pay" + 0.005*"bari" + '
  '0.002*"sayfa" + 0.002*"harcamak" + 0.002*"onarmak" + 0.002*"duran" + '
  '0.001*"homofobik" + 0.001*"solcu"'),
 (6,
  '0.024*"ülke" + 0.022*"bir" + 0.014*"olmak" + 0.012*"demek" + 0.010*"var" + '
  '0.010*"mi" + 0.009*"değil" + 0.008*"se" + 0.008*"bu" + 0.008*"desen"'),
 (7,
  '0.025*"bin" + 0.015*"düşman" + 0.010*"beklemek" + 0.005*"misafir" + '
  '0.002*"muhacir" + 0.002*"azınlık" + 0.002*"gazi" + 0.002*"ayasofya" + '
  '0.002*"gem" + 0.002*"milyar_har"'),
 (8,
  '0.038*"ev" + 0.010*"do" + 0.008*"kira" + 0.008*"değer" + 0.007*"tl" + '
  '0.006*"lira" + 0.006*"enflasyon" + 0.004*"fiyat" + 0.002*"konut" + '
  '0.002*"faiz"'),
 (9,
  '0.013*"seçi" + 0.013*"açık" + 0.011*"vere" + 0.009*"hesap" + 0.004*"bulmak" '
  '+ 0.003*"merhamet" + 0.002*"üç_beş" + 0.001*"aciz" + 0.001*"erke" + '
  '0.001*"gidişini"')]
```

Figure 15. Dominant Topics for the first LDA model

7. Assesment

The GSDMM was successful in creating clusters by itself. But the coherence score of the GSDMM model was 0.3 which means that it needs to be improved. For further research, one might seek better evaluation metrics for the GSDMM algorithm: how to tune it, how to decide the hyperparameters etc. After the GridSearchCV tuning and multiple coherence score calculations with different topic numbers and after we fed the best hyperparameters and best number of topics to the original LDA model, we got a very impressive coherence metric of 0.5. Although, there was a severe overlapping of almost every topic. With the LDAMulticore and LSI algorithms, we got 0.3 coherence scores despite the fact that we chose the best hyperparameters and best number of topics with the elbow method.

8. Discussion

After inspecting our research we have come to notice a few setbacks that would threaten the quality of our research. These setbacks were;

1. Number of Turkish stopwords should be expanded. The NLTK Turkish stopwords did not contain enough stopwords so therefore not enough were eliminated in the pre-processing stage, causing the GSDMM and LDA to build their clusters containing these stop words such as "mi", "yani","bile" etc.
2. GSDMM parameters could be better examined, tested with many other parameters. After a literature review, the hyper parameters used seemed fit for short text analysis. But further experimenting could increase the coherence score.

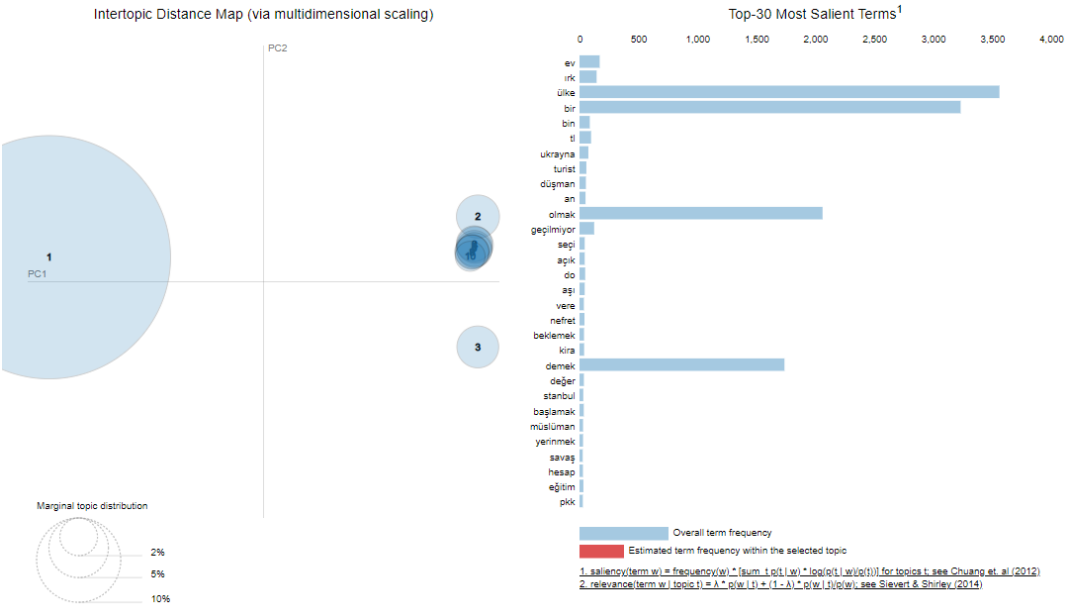

Figure 16. Dominant Topics for the first LDA model Coherence Score: 0.5, Hyperparameters and topic numbers tuned.

	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	6.0	0.9244	ülke, bir, olmak, demek, var, mi, değil, se, b...	uyku " ülke tane bile mülteci istemek sayıklam...
1	1	6.0	0.7749	ülke, bir, olmak, demek, var, mi, değil, se, b...	ülke mülteci istilâ kurtarmak
2	2	6.0	0.9727	ülke, bir, olmak, demek, var, mi, değil, se, b...	lan mülteci yerleşmek amk ö kadar mi bakmak mü...
3	3	6.0	0.8170	ülke, bir, olmak, demek, var, mi, değil, se, b...	sığınmak göçmenmülteci geçmek yapmak sonra akp...
4	4	6.0	0.8864	ülke, bir, olmak, demek, var, mi, değil, se, b...	biz arkadaş arabistan dönmek ülkemülteci kamp mi
...	...	...	...	...	...
9995	9995	6.0	0.7996	ülke, bir, olmak, demek, var, mi, değil, se, b...	işçi bulmak konu zor yaşamak iş var eleman yok...
9996	9996	6.0	0.8075	ülke, bir, olmak, demek, var, mi, değil, se, b...	biri desen allah kitap demek oy toplamak nas v...
9997	9997	6.0	0.7464	ülke, bir, olmak, demek, var, mi, değil, se, b...	nidal zaten sorumak ora isveç kura başka kita...
9998	9998	6.0	0.8183	ülke, bir, olmak, demek, var, mi, değil, se, b...	yıldız bugün gün mülteci ülke dönmek günü se g...
9999	9999	5.0	0.8200	stanbul, müslüman, pay, bari, sayfa, harcamak,...	of bildirim okuyunca mülteci san kdiodhohdoejd...

Figure 17. Dominant Topics for the first LDA model

3. The length of the data might have been increased. In our case, because of the RAM capacity of our computer, we decided to perform the topic modeling on a random subset of 10.000 refugee related tweets although we acquired via the SNS TweetScraper over 40.000 tweets over a span of a year. The study can be replicated on a bigger number of tweets.
4. Although we tuned the LDA model with GridSearchCV and checked related evaluation scores (perplexity etc.) and found the most efficient topic number, the LDA model created a pathological clustering /overlapping of topics. In LDA, the topics are expected to be dispersed somewhat equally. Although the coherence score was somewhat good (.5), further research might examine this clustering in more detail.
5. Although the LDAMulticore topics were dispersed somewhat equally and the hyperparameters were tuned and the number of topics were chosen with the elbow method, the coherence score remains to be .3. Even though LDA algorithms perform low on short texts, this coherence score migiht have been between .4 or .5. A better tuning of the algorithm might be necessary.

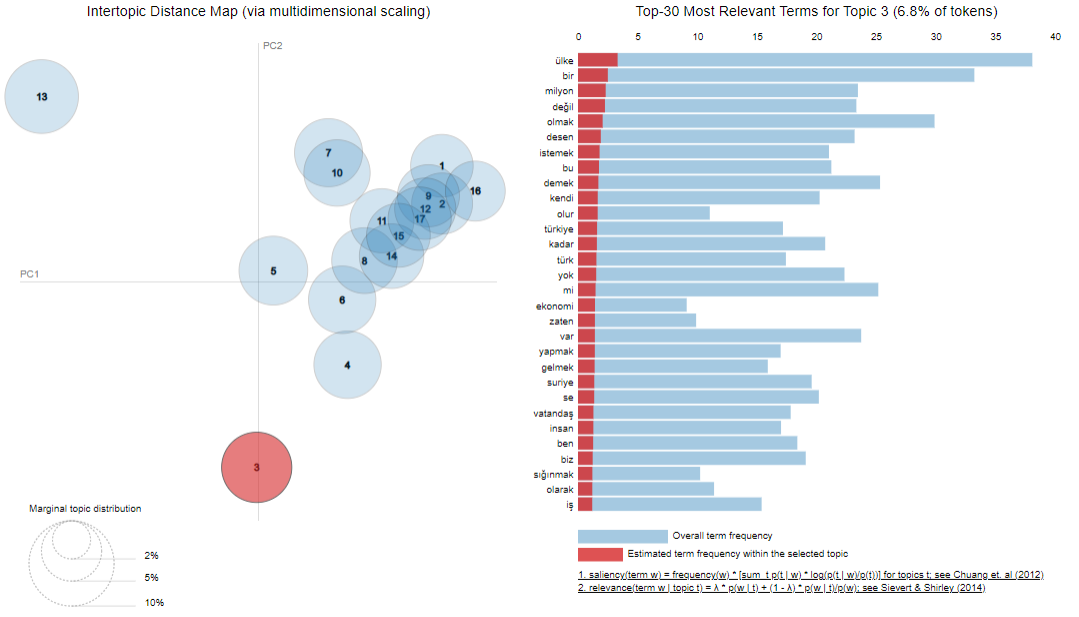

Figure 18. LDAMulticore, Coherence Score 0.3, Hyperparameters and topic numbers tuned.

	Topic_Num	Topic_Perc_Contrib	Keywords	Text
0	0.0	0.5189	ülke, bir, sınır, desen, olmak, kadar, mi, ist...	[cinsel, yönelim, hetero, yaş, ad, ece, burç, ...
1	1.0	0.4909	ülke, yok, demek, olmak, suriye, bir, istemek,...	[chp, mersi, millet, vekil, ali, mahir, başarı...
2	2.0	0.4947	ülke, bir, milyon, değil, olmak, desen, isteme...	[bura, bahis, geçen, mevzu, tam, olarak, değil...
3	3.0	0.5023	bir, ülke, olmak, istemek, değil, ben, bu, var...	[sırf, mülteci, rahatınızı, bozmak, demokrasi,...
4	4.0	0.5008	ülke, bir, olmak, desen, var, mi, değil, demek...	[ongu, vallaha, mi, ata, yapmak, fabrika, sata...
5	5.0	0.4920	ülke, suriye, bir, gelmek, milyon, türkiye, va...	[ergi, kuzey, irak, uçuşmak, yasak, bölge, ilâ...
6	6.0	0.4973	ülke, olmak, demek, bir, var, biz, vatandaş, m...	[duyuru, dışarı, mülteci, dolay, kendi, güven,...
7	7.0	0.4945	ülke, mi, bir, olmak, bu, yok, demek, gelmek, ...	[abi, cezaevi, ev, kira, çocuk, okumak, bir, g...
8	8.0	0.4903	ülke, bir, var, olmak, demek, biz, ben, desen,...	[mem, davut, ben, gazi, antep, şehit, kâmil, i...
9	9.0	0.4905	ülke, olmak, bir, desen, değil, kendi, mi, ist...	[ney, kuzey, irak, pkk, pençe, harekât, düzenl...
10	10.0	0.4983	bir, desen, yok, ülke, değil, demek, mi, olmak...	[esat, tam, bam, telin, dok, infaz, görüntü, ü...
11	11.0	0.4880	mi, demek, ülke, bir, olmak, değil, milyon, ke...	[esad, öksüz, yetim, biz, iyi, bakıyor, bura, ...
12	12.0	0.5112	ülke, bir, olmak, se, milyon, var, mi, kendi, ...	[shadow, orta, zekâ, yok, ahlak, yalan, varadu...
13	13.0	0.4941	ülke, suriye, bir, istemek, demek, yok, değil,...	[hulya, sübliminal, ukrayna, mülteci, biri, ta...
14	14.0	0.4952	biz, ülke, bir, olmak, değil, se, yok, oy, dem...	[bm, milyon, mülteci, göçünde, çark, et, mülte...
15	15.0	0.4842	var, ülke, bir, olmak, mi, demek, yok, desen, ...	[savaşmak, var, banane, ayçiçeği, ekmek, buğda...
16	16.0	0.5011	ülke, bir, demek, olmak, milyon, var, mi, yok,...	[bahar, evet, muktedir, yeni, oluyor, icraatla...

Figure 19. LDAMulticore

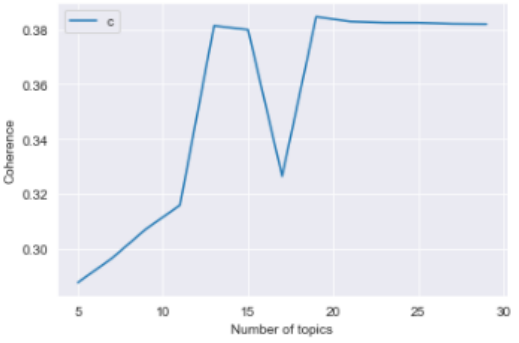

Figure 20. LDAMulticore coherence chart, optimal number of topics is 17.

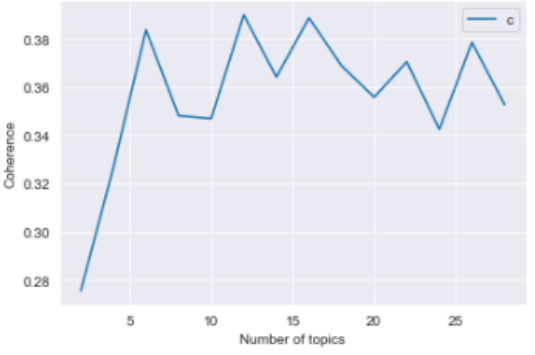

Figure 21. LSI coherence chart, optimal number of topics is 12.

```
[(0,
  '0.249*"ulke" + 0.209*"bir" + 0.190*"olmak" + 0.164*"demek" + 0.156*"mi" + 0.155*"istemek" + 0.150*"degil" + 0.149*"var" + 0.147*"desen" + 0.142*"yok"'),
 (1,
  '-0.854*"istemek" + -0.333*"ulke" + 0.110*"var" + 0.101*"mi" + 0.097*"yok" + 0.094*"milyon" + -0.082*"hicbir" + -0.067*"kendi" + 0.067*"demek" + -0.056*"gitmek"'),
 (2,
  '0.545*"kendi" + 0.317*"ulke" + 0.294*"vatandas" + 0.273*"olmak" + -0.240*"istemek" + -0.217*"demek" + -0.205*"mi" + 0.176*"beter" + -0.128*"desen" + -0.124*"degil"'),
 (3,
  '0.564*"milyon" + 0.456*"mi" + 0.190*"para" + -0.159*"bir" + 0.145*"var" + -0.137*"ben" + -0.134*"turk" + -0.127*"gitmek" + 0.126*"emek" + 0.116*"ukrayna"'),
 (4,
  '-0.460*"mi" + 0.433*"suriye" + 0.270*"turkiye" + 0.224*"milyon" + 0.205*"afgan" + -0.200*"se" + 0.189*"ukrayna" + -0.153*"kendi" + -0.142*"ben" + 0.140*"turk"'),
 (5,
  '-0.437*"yok" + 0.421*"mi" + 0.285*"degil" + 0.263*"suriye" + -0.241*"milyon" + -0.210*"para" + -0.198*"var" + 0.182*"ukrayna" + -0.175*"biz" + -0.136*"emek"'),
 (6,
  '0.451*"yok" + -0.310*"milyon" + 0.246*"var" + -0.237*"vatandas" + -0.217*"se" + 0.200*"suriye" + -0.200*"vermek" + -0.191*"oy" + -0.190*"gitmek" + 0.158*"afgan"'),
 (7,
  '0.416*"turk" + 0.253*"para" + -0.234*"milyon" + 0.226*"vatandas" + 0.217*"yok" + 0.206*"gitmek" + 0.202*"mi" + -0.200*"sorun" + -0.176*"ulke" + -0.133*"sorumak"'),
 (8,
  '-0.377*"se" + -0.375*"demek" + 0.332*"degil" + 0.205*"vatandas" + -0.168*"ben" + -0.167*"suriye" + 0.163*"oy" + -0.156*"afgan" + 0.155*"sığınmak" + 0.147*"sorumak"'),
 (9,
  '-0.347*"gitmek" + 0.293*"degil" + -0.225*"kadar" + -0.216*"mi" + 0.208*"vatandas" + -0.201*"ulke" + 0.198*"sığınmak" + 0.190*"ben" + 0.184*"olmak" + 0.157*"yok"'),
 (10,
  '0.350*"vatandas" + 0.307*"oy" + -0.290*"milyon" + 0.232*"vermek" + -0.229*"turk" + -0.210*"olmak" + 0.202*"suriye" + 0.194*"demek" + 0.180*"var" + -0.166*"degil"'),
 (11,
  '-0.440*"demek" + 0.356*"kadar" + -0.269*"kendi" + 0.258*"se" + -0.207*"milyon" + 0.207*"para" + -0.152*"yok" + -0.132*"irk" + 0.129*"ukrayna" + -0.128*"bahsetmek"')]
```

Figure 22. LSI topics

9. References

Bartlett, Jamie, and Richard Norrie. "Immigration on Twitter: understanding public attitudes online." (2015).

Bosco, Cristina, Viviana Patti, Marcello Bogetti, Michelangelo Conoscenti, Giancarlo Francesco Ruffo, Rossano Schifanella, and Marco Stranisci. "Tools and resources for detecting hate and prejudice against immigrants in social media." In SYMPOSIUM III. SOCIAL INTERACTIONS IN COMPLEX INTELLIGENT SYSTEMS (SICIS) at AISB 2017, pp. 79-84. AISB, 2017.

Calderón, C. A., de la Vega, G., Herrero, D. B. (2020). Topic modeling and characterization of hate

speech against immigrants on Twitter around the emergence of a far-right party in Spain. Social Sciences, 9(11), 188.
Erdogan-Ozturk, Y., Isik-Guler, H. (2020). Discourses of exclusion on Twitter in the Turkish Context: ülkemdesuriyeliistemiyorum (idontwantsyriansinmycountry). Discourse, Context Media, 36, 100400.
Flores, René D. "Do anti-immigrant laws shape public sentiment? A study of Arizona's SB 1070 using Twitter data." American Journal of Sociology 123, no. 2 (2017): 333-384.
Franck Düvell, David Schiefer, Ali Zafer Sagıroglu and Lena Mann. (2021) How Many Syrian Refugees in Turkey Want to Migrate to Europe and Can Actually Do So?Results of a Survey Among 1,900 Syrians. DRN 05 | 21 Berlin
Freire-Vidal, Yerka, Eduardo Graells-Garrido, and Francisco Rowe. "A framework to understand attitudes towards immigration through Twitter." Applied Sciences 11, no. 20 (2021): 9689.
Freire-Vidal, Yerka, and Eduardo Graells-Garrido. "Characterization of local attitudes toward immigration using social media." In Companion Proceedings of The 2019 World Wide Web Conference, pp. 783-790. 2019.
Gualda, Estrella, and Carolina Rebollo. "The refugee crisis on Twitter: A diversity of discourses at a European crossroads." Journal of Spatial and Organizational Dynamics 4, no. 3 (2016): 199-212.
Inuwa-Dutse, Isa, Mark Liptrott, and Ioannis Korkontzelos. "Migration and Refugee Crisis: a Critical Analysis of Online Public Perception." arXiv preprint arXiv:2007.09834 (2020).
Öztürk, Nazan, and Serkan Ayvaz. "Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis." Telematics and Informatics 35, no. 1 (2018): 136-147.
Rowe, F., Mahony, M., Graells-Garrido, E., Rango, M., Sievers, N. (2021). Using Twitter to track immigration sentiment during early stages of the COVID-19 pandemic. Data Policy, 3, E36. doi:10.1017/dap.2021.38
Sahin, Ecem, Tolga E. Dagli, Ceren Acarturk, Figen Sahin Dagli. (2021) Vulnerabilities of Syrian refugee children in Turkey and actions taken for prevention and management in terms of health and wellbeing, Child Abuse Neglect, Volume 119
Yan, Xiaohui, Jiafeng Guo, Shenghua Liu, Xue-qi Cheng, and Yanfeng Wang. "Clustering short text using ncut-weighted non-negative matrix factorization." In Proceedings of the 21st ACM international conference on Information and knowledge management, pp. 2259-2262. 2012.