Online Topic Modeling: Keeping Track of News Topics for Social Good

Zahra Ahmadi, Sophie Burkhardt, and Stefan Kramer

Institut für Informatik, Johannes Gutenberg-Universität,
Staudingerweg 9, 55128, Mainz, Germany
zaahmadi@uni-mainz.de, burkhardt,kramer@informatik.uni-mainz.de

Abstract. The *refugee crisis* has become an important, albeit controversial, issue for European countries. There have been many debates in favor or against accepting refugees in social media; however, there is little work on the interpretation of data in this regard. In this paper, we propose an online topic modeling approach which is able to evolve over time and finds the most important topics at each time slot. Our study shows that outside events have a visible impact on the media and this perception can be changed or evolving over time.

1 Introduction

Europe has witnessed a large movement of migrants and refugees from Africa and the Middle East in recent years. The arrival wave started in August 2015, and since then it has been in the spotlight of the media by reporting an increasing number of events and heated and polarized debates relevant to this phenomenon. The implications of this crisis are complex and wide, however, data mining experts just recently considered the interpretation of the media: Coletto *et al.* [2] proposed an adaptive framework to analyze the spatial, temporal and sentiment aspects of a polarized topic discussed in online social media; the GDELT data[1] was used to answer the question of whether the Arab Spring sparked a wave of global protests[2]; and Data For Democracy built a tool capable of tracking and analyzing refugees and other people forced to evacuate their homes[3].

In this work, our goal is to analyze media from Germany as one of the highly affected European countries to address the following questions: “What are the main concerns of each party or news source? How does the perception evolve over time? How is the perception influenced by events? How similar are different parties and sources in this aspect?”. As a result, we propose an online topic modeling method to keep track of the topics appearing over time and evaluate the results on a relatively large dataset from German media.

[1] http://gdeltproject.org/data.html#rawdatafiles

[2] https://foreignpolicy.com/2014/05/30/did-the-arab-spring-really-spark-a-wave-of-global-protests/

[3] https://www.un.org/press/en/2017/pi2207.doc.htm

2 Online Topic Modeling

The generative process for Latent Dirichlet Allocation (LDA) is given as follows:

$$\phi \sim \mathrm{Dir}(\beta), \theta \sim \mathrm{Dir}(\alpha), z \sim \mathrm{Mult}(\theta), w \sim \mathrm{Mult}(\phi_z) \tag{1}$$

For each topic k, a multinomial distribution ϕ_k over words is drawn from a Dirichlet distribution with parameter β. For each document d, a distribution over topics θ is drawn from a Dirichlet with parameter α. For each word w_{di} in document d a topic indicator z_{di} is drawn from the multinomial distribution θ. Finally, the word w is drawn from the multinomial distribution $\phi_{z_{di}}$ associated with the chosen topic.

To track the topics online, we separate the data into different time slices $D = \{D^1, \ldots, D^{t-1}, D^t\}$. Following the method proposed by AlSumait *et al.* [1], for each time slot t our method learns a topic model by Gibbs sampling [3] where the parameters β are a weighted mixture of the matrices $\phi^1, \ldots, \phi^{t-1}$ from the previous time slots:

$$\beta_k^t = \sum_{t'=1}^{t-1} \omega^{t'} \phi_k^{t'}, \tag{2}$$

where ω^t is the weight associated with time slot t.

In practice, this means that one has to keep all matrices ϕ^t associated with all time slots in memory to compute the weighted sum for the current time slot. This is inefficient in terms of memory and runtime and not in the spirit of a true online method. In their experiments, AlSumait *et al.* [1] therefore only use the previous time slot, meaning ω^t is zero for all other time slots. This makes the method more practically relevant; however, it introduces a problem: Consider the case where a certain topic occurs in one time slot, is absent in the next time slot, and reoccurs in the next. In this case, the model will forget everything from the previous occurrence of the topic since it only takes the previous time slot into account. This makes the results highly dependent on the size of the data slices and the content of the data.

Our solution is based on the definition of variational Bayes online topic models [4]. In online variational Bayes, instead of taking samples, a natural gradient is calculated. After each batch, the model is then updated as

$$\phi^t = (1 - \rho)\phi^{t-1} + \rho\hat{\phi}^t, \tag{3}$$

where ρ is a real-valued update parameter in the $[0, 1]$ interval and $\hat{\phi}$ is the estimate for ϕ based on the current batch. In our model, we adopt this strategy and only use it for updating the parameter β:

$$\beta^t = (1 - \rho)\beta^{t-1} + \rho\phi^{t-1}. \tag{4}$$

This means that we let the prior parameter β converge to a stationary point, whereas ϕ^t is specific to a certain time slot t. Thus, we can analyze the data in a certain time slot while inducing the model to keep the topics stable over time without having to save any of the previous matrices ϕ. In contrast to the online method by Hoffman *et al.* [4], our method learns topics that are only based on the data from the current time slot, making it easier to track changes or detect specific events.

3 Experiments

We extracted a set of news articles with a keyword related to refugees from January 2016 to May 2017 from German media. We preprocessed the data by removing numbers, stop words and words with one letter and made all letters lower case. Those instances which become empty by preprocessing are removed; hence, the dataset is reduced to *208, 683* articles with *71, 633* features. To run offline LDA and the online topic modeling method, we set the number of topics to *100*. The time slots contain *10, 000* instances, and the method is repeated *100* times for each slot. The update parameter ρ is set to $\frac{1}{100^{0.9}}$, according to the instructions provided by Hoffman *et al.* [4].

Figure 1 illustrates the results of offline LDA and the topic evolution of the online topic modeling method as a word cloud for one of the topics which is mainly about the AFD party (a right-wing populist political party in Germany) and their news related to refugees. Each box presents the English translation of the top *20* most frequent words of the current topic. The dates show the start point of the time slot, and we represent every third slot in this figure. We can see that although the topic changes over time, it is interestingly all about the AFD party news. The advantage of this model over the previous online LDA models (e.g. [4]) is that it puts more emphasis on the current batch than updating the previous topic model incrementally with a marginal effect of the new batch.

Looking into the most frequent words of each temporal topic, we can observe that in each period, based on the upcoming events, some topics are highlighted: e.g., in the Landtag election of the state North Rhine Westphalia, which was held on 14 May 2017, Helmut Seifen (the AFD politician) was elected. We can already see him on top of the news related to refugee politics on $2017 - 03 - 31$; or because of the importance of the Bundestag election for a young party like AFD, we can observe many discussions related to that since the beginning of 2017, although the election is held only in September 2017. This topic is one of the 100 resulting topics by our method. We observe some other topics about other parties (e.g., SPD and CDU), some topics related to refugee integration, some about job markets or even about women and children. However, not all of the topics are that well-defined.

4 Conclusion, Challenges, and Future Work

We proposed an online topic modeling method to find the topics related to refugees in German media. During our experiments, we faced several challenges. Our first goal was to find a categorization of the reasons for being against or in favor of accepting the refugees among different opinions. Although the model

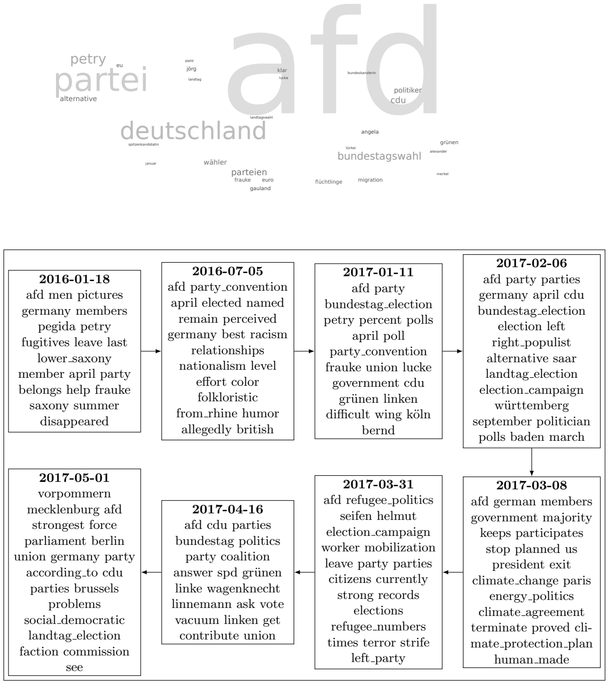
petry
eu
partei
alternative
steht
jörg
landtag
afd
klar
lucke
bundeskanzlerin
politiker
cdu
landtagswahl
deutschland
spitzenkandidatin
angela
grünen
türkei
alexander
bundestagswahl
januar
wähler
parteien
frauke
euro
gauland
flüchtlinge
migration
merkel
2016-01-18 afd men pictures germany members pegida petry fugitives leave last lower_saxony member april party belongs help frauke saxony summer disappeared
2016-07-05 afd party_convention april elected named remain perceived germany best racism relationships nationalism level effort color folkloristic from_rhine humor allegedly british
2017-01-11 afd party bundestag_election petry percent polls april poll party_convention frauke union lucke government cdu grünen linken difficult wing köln bernd
2017-02-06 afd party parties germany april cdu bundestag_election election left right_populist alternative saar landtag_election election_campaign württemberg september politician polls baden march
2017-03-08 afd german members government majority keeps participates stop planned us president exit climate_change paris energy_politics climate_agreement terminate proved climate_protection_plan human_made
2017-03-31 afd refugee_politics seifen helmut election_campaign worker mobilization leave party parties citizens currently strong records elections refugee_numbers times terror strife left_party
2017-04-16 afd cdu parties bundestag politics party coalition answer spd grünen linke wagenknecht linnemann ask vote vacuum linken get contribute union
2017-05-01 vorpommern mecklenburg afd strongest force parliament berlin union germany party according_to cdu parties brussels problems social_democratic landtag_election faction commission see

Fig. 1: An example of topic evolution on a topic relevant to AFD party. The word cloud illustrates the output of a batch LDA on the data while the text boxes show the output of our proposed method.

finds interesting topics in the data, this goal remains unsolved. Another unsuccessful attempt was to find an unsupervised method to cluster different sources based on their opinions expressed in their articles with the hope of finding their political view. As a future work one could develop a semi-supervised approach to build a topic model which can reach these goals.

References

1. AlSumait, L., Barbará, D., Domeniconi, C.: On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In: Eighth IEEE International Conference on Data Mining. pp. 3–12 (2008)
2. Coletto, M., Esuli, A., Lucchese, C., Muntean, C.I., Nardini, F.M., Perego, R., Renso, C.: Sentiment-enhanced multidimensional analysis of online social networks: Perception of the mediterranean refugees crisis. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp. 1270–1277 (2016)
3. Griffiths, T.L., Steyvers, M.: Finding scientific topics. In: Proceedings of the National Academy of Sciences of the United States of America. vol. 101, pp. 5228–5235. National Academy of Sciences (2004)
4. Hoffman, M., Bach, F.R., Blei, D.M.: Online learning for latent dirichlet allocation. In: Advances in neural information processing systems. pp. 856–864 (2010)