MultiTM: A Multilingual Topic Modeling approach based on Clustering

Shashank Subramanya and Gerasimos Spanakis

Department of Advanced Computing Sciences, Maastricht University

Many real-world datasets contain text in multiple languages generated from responses to open-ended questions. As data is collected from domain experts writing in their native language, it is usually rich in information. Topic models can be used to analyse textual information as they identify the main themes in data and there are various techniques to extract topics from multilingual text. However, research shows that the performance of topic modeling techniques differs across datasets and varies based on the choice of evaluation metric. Moreover, there is a lack of consensus on when a certain technique performs better and why. This thesis addresses the issue in two steps. First, extensive analyses of various topic modeling techniques is performed and insights are provided on their ability to generate multilingual topics. On a text corpus containing unstructured data in Dutch, German, English, French, Polish, Czech, and Italian, topics are generated from three different approaches. Latent Dirichlet allocation (LDA) - a traditional bag-of-words method [1], Combined TM - a neural topic model [2], and BERTopic - a clustering-based method [3] are evaluated using quantitative and qualitative metrics. Subsequently, a modified clustering-based approach for multilingual topic modeling, MultiTM is presented and shown to produce more interpretable topics.

As topic modeling is an unsupervised problem, evaluation is conducted using various metrics that include NPMI topic coherence (TC) [4], topic similarity (TS) using pre-trained multilingual word embeddings aligned in the same vector space [5], and topic diversity (TD). For multilingual documents, language can adversely influence the topic assigned to it. Thus, a new metric, language concentration (TL) is introduced to measure the impact of linguistic differences on topic generation. Finally, topics are qualitatively evaluated by 34 domain experts who rated 10 randomly selected topics for each technique individually on a 1-3 scale on how useful they are in identifying the main themes in data (HE) [6].

Experiments were conducted using the quantitative and qualitative evaluation metrics to identify the best topic model for each technique. Table 1 contains the results comparing them. LDA generated topics were highly skewed by language as it uses a simple bag-of-words representation of a document. CombinedTM, through the use of contextual multilingual sentence embeddings [7], performed better with a higher topic similarity and a significantly higher HE than LDA. However, as it used a bag-of-words layer as input, document language impacted topic creation and 15 of the 20 topics had top 10 words of a single language. BERTopic with multilingual embeddings [7] performed the best among the three. It had the highest TS and was least affected by linguistic differences in documents. Moreover, its topics received the highest average rating

of 2.32 from evaluators. Another key finding was that, unlike monolingual topic modeling, topic coherence was not correlated with human interpretability.

While results from BERTopic were promising, it had a few drawbacks. TD was a low 0.93 with many topics sharing the same topic words. This was due to imperfect clustering that grouped semantically similar documents separately. Additionally, it was affected by low quality multilingual sentence embeddings of low-resource languages. All Polish documents were part of a single cluster despite having similar content as the others. Also, topics included many frequently occurring words that hampered interpretability. The dataset was from the finance domain and there was a repeated occurrence of words such as "credit", "invoice", and "customer" in different languages.

All the issues are due to multilinguality of data and to resolve it MultiTM is proposed with three main modifications to BERTopic. First, the data is split into training and test sets prior to clustering. The training set includes documents in Dutch, German, and English that comprise 94% of the documents and are used for creating clusters and topics. The test set is then assigned to the existing clusters. This reduces the impact of low-quality multilingual embeddings on clustering. Polish documents are now marked as outliers by the density clustering algorithm and do not interfere with topic generation. The choice of the training set is subjective and the aim is to choose the least number of languages that capture the most information in data. Second, the most frequently repeated words are removed post clustering before topic generation. Third, an additional term is added to the class-based TFIDF score used to extract topic words from each cluster. All terms that occur only in a few clusters are upweighted in a form similar to the IDF score.

Table 1. Topic evaluation scores

Topic Model	TC↑	TS↑	TD↑	TL↓	HE↑
MultiTM	0.08	0.13	**0.98**	**1.68**	-
BERTopic	**0.37**	**0.20**	0.93	5.10	**2.32**
CombinedTM	0.20	**0.20**	0.97	6.08	2.23
LDA	0.27	0.18	**0.98**	9.15	2.05

Following these updates, MultiTM generated topics with a TD of 0.98, a 5.4% increase compared to BERTopic. Also, it had 67.1% lesser impact of multilinguality on topics with a TL of 1.68. The gains were achieved by sacrificing TC which was shown to not be correlated with human evaluation of relevance and lower TS due to the presence of rare topic words with no multilingual embeddings. Moreover, a comparison of similar topics from MultiTM and BERTopic showed that MultiTM had more topic-specific rare words with a higher importance in the top 10 words. The proposed modifications make MultiTM topics easier to interpret. Finally, MultiTM can be used to generate topics that summarize information in any multilingual text corpus irrespective of the domain or choice of dataset.

References

1. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022
2. Bianchi, Federico, Silvia Terragni, and Dirk Hovy. "Pre-training is a hot topic: Contextualized document embeddings improve topic coherence." arXiv preprint arXiv:2004.03974 (2020).
3. Grootendorst, Maarten. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure." arXiv preprint arXiv:2203.05794 (2022).
4. Lau, Jey Han, David Newman, and Timothy Baldwin. "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality." Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. 2014.
5. Multilingual Word Embeddings, https://github.com/facebookresearch/MUSE
6. Newman, David, et al. "Automatic evaluation of topic coherence." Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics. 2010.
7. Multilingual Sentence Transformer, https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2