**Topic Extraction BERTopic’s Insight into the 117th Congress’s Twitterverse**

**Goal**: Comparing BERTopic’s to other methods in twitter data and tuning BERTopic for this type of data

Social media plays a big role in politics, but it also spreads a lot of noise and disinformation. In twitter, the language used it short, direct and in real time which makes the data chaotic and challenging. Therefore, the classic models like LDA struggle to deal with it.

There were two parts to this article: the first was comparing BERTopic’s to the classic methods in tweet data and the second to tune BERTopic to this type of data since it is very customizable.

The evaluation of BERTopic’s against other methods (like LDA and LSA) was done according to coherence, perplexity, diversity, and stability. The results showed the BERTopic having a higher performance over the other methods. It also has the big advantage of its flexibility since every stage of the topic modeling pipeline can be customized and the fact that it doesn’t require a pre-defined number of topics.

In terms of customization, it advises the use of Word2Vec in the embedding phase since it is less unstable than SBERT in terms of diversity. In the dimensionality reduction stage, not using any model was best and for the clustering stage HDBSCAN was advised as it showed significantly lower average perplexity than K-means.

Best parameters in Vectorizer:

* ngram\_range: (1, 1) - was the better one because of the parameter coherence (0.79)
* min\_df: 5 - perform better on coherence (0.79), diversity (1), and stability (0.66) score
* max\_features: 170,000 - value with the highest stability (0.674)

Best parameters in Topic representation:

* bm25\_weighting: False – it does not improve the model as both values behave the same way in all metrics
* reduce\_frequent\_words: True – the optimized outperforms the default on average coherence and stability

Compared to the default setup, the optimized BERTopic model achieved higher average coherence and stability, while maintaining similar diversity and nearly identical perplexity.