Topic Modeling by Clustering Language Model Embeddings: Human Validation on an Industry Dataset

Anton Eklund
Umeå University
Adlede AB
Umeå, Sweden
anton.eklund@cs.umu.se

Mona Forsman
Adlede AB
Umeå, Sweden
mona.forsman@adlede.com

Abstract

Topic models are powerful tools to get an overview of large collections of text data, a situation that is prevalent in industry applications. A rising trend within topic modeling is to directly cluster dimension-reduced embeddings created with pretrained language models. It is difficult to evaluate these models because there is no ground truth and automatic measurements may not mimic human judgment. To address this problem, we created a tool called STELLAR for interactive topic browsing which we used for human evaluation of topics created from a real-world dataset used in industry. Embeddings created with BERT were used together with UMAP and HDBSCAN to model the topics. The human evaluation found that our topic model creates coherent topics. The following discussion revolves around the requirements of industry and what research is needed for production-ready systems.

1 Introduction

Contextual advertising is a rising solution for ad placement on the Internet, which avoids the need for user data and cookies. However, to find good contexts for a placement, the content of a page needs to be known and classified as a useful advertising context. News media are dependent on advertising for funding their work and is therefore an important market for contextual advertising. The news is constantly changing, which makes it difficult to create classifiers that can catch and categorize new articles. A possible way to solve this is to use unsupervised topic models, which are powerful tools to structure large collections of text data.

Traditional approaches to do topic modeling are stochastic, the most well-known one being Latent Dirichlet Allocation (LDA) (Blei et al., 2003). The problem with stochastic approaches is that they are slow and getting increasingly more difficult to integrate with modern language models (Zhao et al., 2021; Vayansky and Kumar, 2020). To tackle this, Neural Topic Models (NTMs), which leverage the power of neural networks to create topic models, are becoming increasingly popular. The techniques of our particular interest are Neural Topic Modeling by Clustering Embeddings (NTM-CE). We define NTM-CE as models that use a distance-based clustering algorithm on the document embeddings created with a Pretrained Language Model (PLM). Performing topic modeling by directly clustering embeddings has been shown to perform comparably to LDA (Sia et al., 2020), or better (Meng et al., 2022), and is claimed to create more coherent topics than other types of NTMs (Zhang et al., 2022). These results make the models attractive for use within industry as analytical tools, and hopefully as part of an automatic classification pipeline. Here, we evaluate a modified BERTopic (Grootendorst, 2022) on an industry dataset consisting of unstructured news articles from a brief period of time.

The evaluation of topic models is not trivial since the lack of annotated datasets makes methods like the F1 score not applicable. It also is counterintuitive to our purpose of having flexible topic models if they are evaluated through static dataset categories. Instead, the field has gravitated to automatic measurements which do not require a ground truth like the Normalized Pointwise Mutual Information (NPMI) (Bouma, 2009). NPMI is a popular way to evaluate topic models, as it is claimed to emulate human judgment of topic coherence (Lau et al., 2014). However, an alarming study by Hoyle et al. (2021) argues that automatic evaluation with NPMI cannot emulate human judgment, a fact which topic modeling papers usually rely on to make their claims. From an industry perspective, it is important to be able to trust and validate topic models before they can be used in a production system. Additionally, to the best of our knowledge, there are no studies that use human evaluation for NTM-CE. Therefore, this paper presents a human expert evaluation using our novel

Proceedings of EMNLP 2022 Industry Track, pages 645–653
December 9–11, 2020.

tool Systematic Topic Evaluation Leveraging Lists of ARticles (STELLAR), which is described further in Section 4. The human expert evaluation is described in Section 5.

A problem that remains for topic models, including NTM-CE, is the interpretability of the resulting topics. This is usually addressed by selecting a set of keywords deemed to be the most descriptive of the topic. The words closest to the centroid of a cluster can be used as descriptors as seen in Bianchi et al. (2021). Another solution by Grootendorst (2022) is to use a class-based term weighting to extract keywords. The question remains if, and to what extent, human evaluators would find the keywords descriptive enough for the overall topic. Hence, we add an assessment of the topic description using a simple four-point scale.

In this paper, we demonstrate NTM-CE in the industry setting of contextual advertisement and do a human expert evaluation of the topic model using our new STELLAR tool. The NTM-CE is an implementation of BERTopic, described in Section 3, that has been applied to a news data set, described in Section 2. The STELLAR tool for topic evaluation is described in Section 4, with further explanation of the human expert evaluation in Section 5. The results of the human evaluation are presented in Section 6, with further discussions of the process, the results, and future improvements in Section 7.

2 Data

The dataset used for this study is a unique collection of publicly available English online news articles. The collection consists of 10000 articles from 58 publishers collected between 2022-05-29 and 2022-06-22. The lengths of the articles range from 501 to 99000 characters with a mean length of 3052. 9753 articles are shorter than 10000 characters. Except for removing articles shorter than 500 characters, no other filtering of the articles was applied. This makes the dataset have the same characteristics as the news from the sampled weeks, with topics such as the *Queen's jubilee*, *Cancelled flights*, and *Formula 1 racing* taking up a disproportionally large part of the content. These are examples of large but brief news topics that will be irrelevant in a few weeks, illustrating the dynamic nature of the news cycle and why static classifiers are of limited use.

3 Topic Modeling Pipeline

Our NTM-CE approach adopts the pipeline of BERTopic (Grootendorst, 2022) and CETopic (Zhang et al., 2022) by using the sequence presented in Figure 1. The class-based TF-IDF of Grootendorst (2022) is used to create keywords for the topics. The components are described in more detail in the following section.

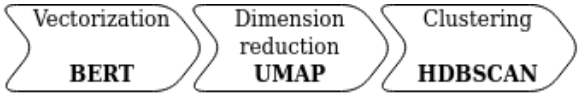

Figure 1: The topic modeling pipeline starts with vectorization using BERT, followed by dimension reduction using UMAP, and ends with clustering using HBDSCAN.

Vectorization was performed with Transformer-based (Vaswani et al., 2017) model BERT (Devlin et al., 2019). It was chosen as it has been widely used in neural topic models and shown to perform well (Grootendorst, 2022; Zhang et al., 2022; Bianchi et al., 2021). Those models used Sentence-BERT from Reimers and Gurevych (2019). However, to have the results less tied to a specific BERT model, the model in this project used the HuggingFace base model[1] $(768D, 12A, 12L)$ which was fine-tuned with Masked Language Modeling for the task. As the final document embedding we used the averaged token embeddings of *hidden_state 1*.[2]

Dimension reduction of high dimensional vectors is used to, among other things, reveal patterns in the data and reduce vector space noise. Techniques for dimension reduction come in two main categories: dimensionality reduction based on matrix factorization and based on neighbor graphs. In this study, we used the neighbor graph method UMAP (McInnes et al., 2018) because it was reported to be both faster and have better clustering quality than the popular t-SNE (Maaten and Hinton, 2008) in the original article.

Clustering has a plethora of techniques but we settled for HDBSCAN (Campello et al., 2013;

[1] https://huggingface.co/docs/transformers/model_doc/bert

[2] Using the unconventional hidden_state 1 as the embeddings was due to a bug in the code which was found after the human evaluation was completed. However, the vector space is similar to the embeddings from the more conventional *last_hidden_state*. Therefore, for showcasing STELLAR, and exploring NTM-CE, we deemed using the embeddings from hidden_state 1 sufficient as the topic model still follows our definition of NTM-CE.

McInnes and Healy, 2017) because of its successful use in Grootendorst (2022) and its ability to dynamically choose the number of topics and their size. We used soft clustering for HDBSCAN, meaning that all points in the vector space get assigned to a cluster, which in turn means that no points were considered outliers.

4 The STELLAR Topic Browser

Systematic Topic Evaluation Leveraging Lists of ARticles (STELLAR) is a tool developed to simplify the in-depth exploration of a topic model into what constitutes the topics rather than just considering the top keywords. The user wants to: 1) get a visual overview of what topics exist and how they are related to each other, 2) be able to quickly identify articles that do not fit into the topic, and 3) go beyond keywords to validate a topic. To solve 1), there is a list of topics with their description alongside a 3D vector space visualization. For 2) and 3), the proposed solution is to allow the user to read the title and keywords of the article and, if needed, to read the text body. The tool needs to be dynamic and interactive, as the user needs the flexibility to study topics freely and investigate different aspects without recreating the topic model and plots.

STELLAR, shown in Figure 2, was created as an application that can run directly in an Internet browser. Its purpose was to aid the user to perform activities 1–3 as described above. It was implemented using the Python Flask[3] library. The core functionalities are a topic list, an article list from the chosen topic, a box for the article text body, and a 3D visualization made in Plotly[4]. For each topic, the articles can be marked as not belonging to the topic and thus assist with the evaluation of the topics. The 3D visualization shows the individual articles as points in the vector space reduced to three dimensions, which are color-coded to the cluster to which they belong. Hovering over the articles shows the title and the keywords of the article which helps the user to get a better understanding of the cluster and search in different sections of a cluster. The repository[5] for STELLAR was released.

[3]https://flask.palletsprojects.com/en/2.1.x/

[4]https://plotly.com/python/

[5]https://github.com/antoneklund/STELLAR

5 Human Expert Evaluation

The first human evaluation of the topic model and STELLAR was made by three experts (including the first author) in the field of news space analysis. From now we call them *evaluators*. An evaluator is distinguished from an annotator in this work, by having the more complex task of contextualizing a set of articles, finding patterns, and drawing the line for what is considered a topic by excluding articles. In contrast, an annotator selects topics from a list of choices and makes decisions for individual articles. We deem the evaluation task too complex to easily be crowd-sourced. We acknowledge that three evaluators may be too few to make strong claims about NTM-CE in general. However, for our purpose of demonstrating one NTM-CE model on an industry dataset, as well as collecting suggested improvements for STELLAR, we deemed the small expert group adequate.

We make a distinction between the terms *cluster*, *topic*, and *focus topic*. A *cluster* is a set of points that are grouped by the clustering algorithm representing a group of article embeddings. A *topic* is a cluster of article embeddings combined with the descriptive topic keywords. This is the output of the topic model. A *focus topic* is the topic of a cluster that the evaluators decide that most of the article supports.

Each evaluator received an individual dataset with 20 randomly sampled articles per topic. Five of the articles per topic overlapped between the evaluators to calculate inter-rater agreement. The task given was to systematically analyze each topic by reading the article titles and keywords, and reading the article body if needed. Then, record their evaluation by 1) deciding on a focus topic with the help of suggestions[6], 2) record the id of articles that did not belong to the focus topic, and 3) give a score between 1 and 4 on how well they thought the keywords given by the topic model reflected the focus topic. The scores correspond to: 1=very bad, 2=bad, 3=good, and 4=very good. We chose a four-point scale to force the evaluators to decide if the keywords are good or bad. The instructions given to the evaluators are specified in Appendix A.

Inter-rater agreement AG was used to assess the

[6]The list of suggested topics was compiled by the first author which will introduce biases. However, evaluators were encouraged to record their custom focus topic if none of the items on the list was satisfactory.

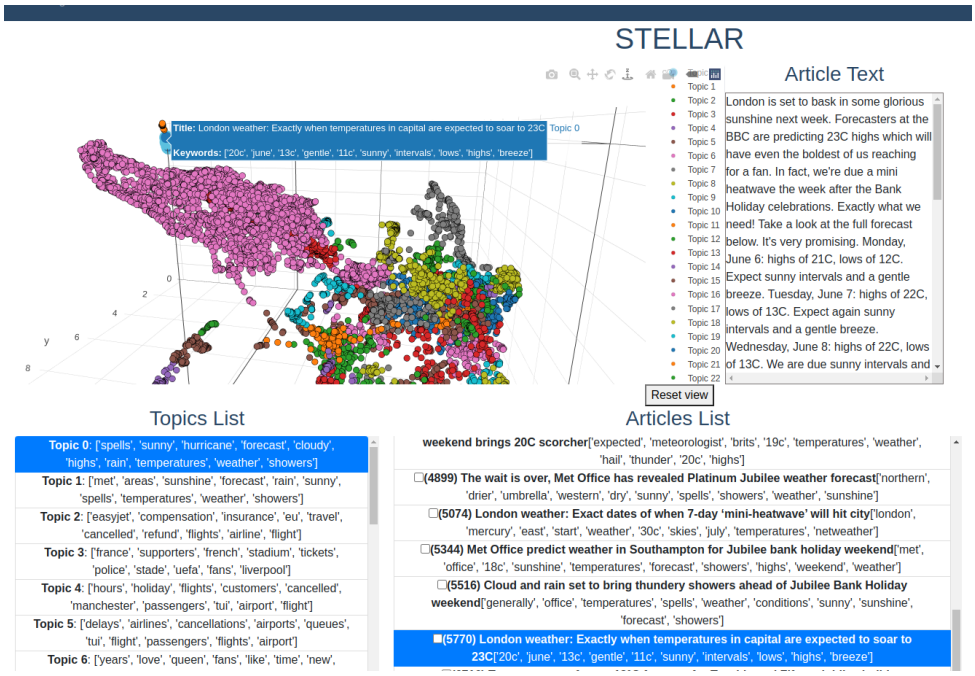

Figure 2: The user interface of STELLAR. The four core components are the 3D visualization, a list of topics with keywords, a list of articles from chosen topics with keywords, and the article text of a chosen topic.

reliability of the evaluators. It was calculated as:

$$AG = \frac{n_{agree}}{N} \quad (1)$$

where n_{agree} is the number of the agreeing decisions, and N is the number of possible decisions. Two different types of decisions are aggregated. The first type of decision is the focus topic of the clusters, called *Agreement focus topic*. The second type of decision is for each overlapping article. The evaluator decides whether they belong to the focus topic or not, called *Agreement overlapping articles*.

To assess whether the topic model produced coherent topics. Our definition of evaluator-determined coherence score (Coh) is:

$$Coh = 1 - \frac{n_{misplaced}}{n_{articles}} \quad (2)$$

where $n_{articles}$ is the number of articles evaluated in the topic and $n_{misplaced}$ is the number of articles that the evaluators found was misplaced into that topic. We call a topic *coherent* if $Coh \geq 0.8$. This threshold at 80% is where we consider a topic coherent enough for being useful in an industrial application. Further, a topic that has at least one evaluator labeling it *Incoherent* will be considered incoherent, regardless of the opinions of other evaluators.

We are aware that the judgment of how coherent a topic is will depend on the individual experiences and interests of every evaluator. However, the purpose of the evaluation is not to find a ground truth of what is the focus topic, but rather to determine if the articles presented by the model form a coherent topic. If the evaluators draw the line on what constitutes a topic differently, we see that as a limitation of the model, and the reduction in coherence score is justified.

Nr topics found	63
Nr coherent topics	52
Average Coh for coherent topics	96%
Articles in coherent topics	57%
Agreement focus topic (including incoherent topics)	87%
Agreement focus topic for coherent topics	98%
Agreement overlapping articles	95%
Keywords describing topic	2.8

Table 1: Statistics of the topic modeling and the human evaluation.

6 Results

The topic modeling pipeline resulted in 63 clusters with varying sizes as seen in Figure 3. The largest one contains 3347 articles, and the smallest ones are around 20. In total, 2367 of the 10000 articles were manually analyzed. The evaluators agreed on 95% of their decision on overlapping

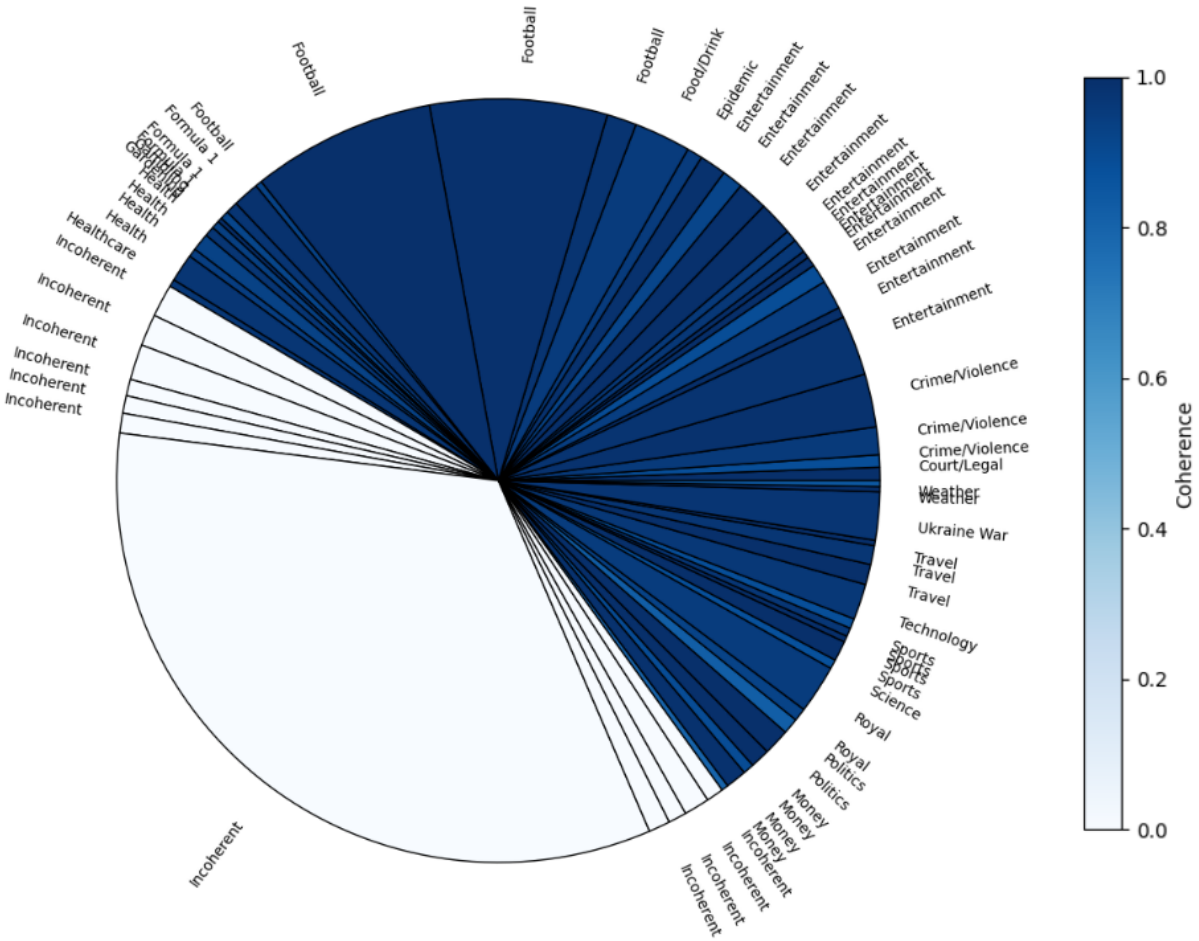

Figure 3: A pie chart illustrating the sizes and coherence of all topics. Light blue means that the topic is incoherent. Different shades of blue indicate how strong the coherence is based on the human evaluation. In the north-west part of the graph, the mixed wedges are Formula 1, Gambling, and Gardening.

articles. The focus topics identified by the evaluators were [*Court/Legal, Crime/Violence, Entertainment, Epidemic, Food/Drink, Football, Formula 1, Gambling, Gardening, Health, Incoherent, Money, Politics, Royal, Science, Sports, Technology, Travel, Ukraine War, Weather*], a total of 20 focus topics. A focus topic can be assigned to multiple clusters. Broader focus topics such as *Entertainment* contain articles about gossip, celebrities, movies, and TV-series. The *Sports* topic is all sports excluding *Football* and *Formula 1*. It is made up of tennis, combat sports, golf, cricket, and rugby among the larger ones.

The data from the human evaluation in Table 1 shows that 52 out of 63 topics were over 80% coherent. Topics that were determined coherent had an average coherence score of 96%. Those topics contain 5653 of the articles, which is 57% of the dataset. A little less than half of the articles ended up in incoherent topics. The largest *Incoherent* cluster (see Figure 3) consists of shorter articles, with an average length of 1500 characters. The features of *Incoherent* clusters will be further explored in Section 7.

The evaluators agree on the focus topic for 87% of topics. In topics where coherence is high, the evaluators agree on the focus topic for 98% of the topics, that is, all topics except one. The disagreements between the evaluators usually came from when one evaluator had chosen *Incoherent* while the others had specified a topic.

Another common disagreement, important for understanding the difficulty of the topic modeling problem is shown in Table 2. The focus topic was about *Weather*, but one can find that the topic consisted of two subtopics, which we can call *Forecasts* and *Hurricanes*. One evaluator decided the focus topic to be *Forecasts* and then continued to mark the articles about *Hurricanes* as misplaced. However, the other evaluators considered the focus topic as *Weather* and thus fully coherent. Cases like these, where one evaluator creates a more nar-

row focus topic than the others, make up for much of the reduction in the total coherence score.

7 Discussion

The topic model found 63 topics in which 20 focus topics were identified. We interpret it as a good partitioning of the corpus, except for the fact that the largest topic was labeled as *Incoherent*. The optimal number of topics found may vary greatly between corpora and the aim should not be to find one cluster per focus topic from the focus topic list, e.g. finding 20 clusters for this dataset. However, for an analytical application in the industry, it would be advantageous to have a way to collect clusters with a similar focus topic into a larger collection. Whether that should be done with the vector space distance or with other methods remains to be studied.

The human evaluation determined 52 out of 63 topics to be coherent, with an average coherency of 96%. However, only 57% of the dataset ended up in coherent topics. One reason for this is the strict requirement that all evaluators should agree on the focus topic for a topic to be considered coherent. However, the foremost reason is that the largest topic, which contained 3347 articles, was labeled *Incoherent*.

A deeper inspection of the clusters of incoherent topics gave interesting insights. The largest cluster contained short articles with half the average length as the rest of the dataset. We assume they have been padded before the BERT vectorization and are clustered on the padding artifacts. An informal test to re-cluster this particular cluster was done to see if dividing it into smaller partitions would create coherent clusters. However, these new clusters did not create coherent topics either. Since we found large coherent clusters such as *Football*, as well as small *Incoherent* clusters, we believe that the clusters should not necessarily be as small nor as balanced as possible, balancing being something Meng et al. (2022) emphasized. Rather, we think that the dynamic properties of HDBSCAN could suffice if guided by inputs from suitable vector space statistics, and applied to a well-formed vector space.

Further analysis of the *Incoherent* topics revealed patterns among the articles but not enough to make them coherent. Some of the topics contain articles on multiple focus topics. An example is a topic with the combination *Real estate*, *Home styling*, and *Tourist attractions* that all describe nice environments but not in one coherent focus topic. One topic has locally anchored articles, but about different focus topics and locations. One of the topics is dominated by first-person stories, however, the focus topics differ, and hence it was incoherent in the evaluation. The same can be said for a topic with very emotional content. These topics deserve a deeper examination and understanding, as they have themes that have a stylistic or emotional character. Studying this remains future work since it was not included in the evaluator instructions.

According to the keyword evaluation, the keywords describing the topics were on average positive (> 2.5), yet not good (2.8). The overall positive assessment was still slightly unexpected as the perception before the study was that the keywords did not describe the focus topics well. One factor affecting the results might be that the evaluators had a better understanding of what the keywords mean after reading the topic articles and therefore thought the keywords described the articles well. A more focused study on keywords for topic descriptions needs to be done to investigate this. Nevertheless, since the description was not close to very good (4), ways forward might be to find better keyword extractors or other methods to describe the topics. A preferred scenario for our industry purposes is a topic model where we trust that all topics have $Coh \geq 80\%$ and that the topic descriptions are clear enough for a human to determine the focus topic without looking deeply into what articles are in the topic. An ideal scenario would be that we can trust a system to automatically decide the focus topic similar to human judgment.

The evaluators agreed on the focus topic for 87% of the topics and also had an agreement of 95% for overlapping articles. The agreement on the focus topic for coherent topics was almost perfect at 98%, which means that it was almost always recognizable. However, as shown in Table 2, there are difficulties even for humans to determine where to limit a focus topic. Then, can we expect topic models to do that for us? We expect topic models to be able to divide the articles into topics with a focus reasonably well. However, for the contextual advertisement vital finesse of correctly finding narrow or trendy focus topics, a human will still be needed. An important addition for managing contextual campaigns would be the possibility to analyze topics over time in the style of Blei and

Topic 0	[spells, sunny, hurricane, forecast, cloudy, highs, rain, temperatures, weather,...]
id: 777	**1st of 2022, Hurricane Agatha heads for Mexico tourist towns**
	[landfall, millimeters, mazunte, mexico, kph, storm, puerto, oaxaca, center, ...]
id: 2510	**Hurricane Agatha is first named storm of Atlantic season after hitting Mexico...**
	[maximum, hurricanes, noaa, atlantic, storm, inches, southern, mexico, hurricane, ...]
id: 2197	**Met Office gives Scotland weather update for Queen's Platinum Jubilee weekend**
	[forecast, places, warmer, rain, unsettled, dry, drier, weather, spells, showers]
id: 4899	**The wait is over, Met Office has revealed Platinum Jubilee weather forecast**
	[northern, drier, umbrella, western, dry, sunny, spells, showers, weather, sunshine]
id: 5770	**London weather: Exactly when temperatures in capital are expected to soar to 23C**
	[20c, june, 13c, gentle, 11c, sunny, intervals, lows, highs, breeze]

Table 2: Example of when the evaluators disagreed on the focus topic. One evaluator decided the focus topic to be *Forecasts* and then continued to mark the articles about *Hurricanes* (on the top) as errors. However, the other evaluators decided that the whole topic was coherent as *Weather*.

Lafferty (2006) or Wang and McCallum (2006). This is something that Grootendorst (2022) has been working on with BERTopic. Another observation was that it would be beneficial if the time-consuming analysis was only required to be done once and then have systems detecting new topics emerging and disappearing.

The tool STELLAR, presented in this paper, was created to allow an evaluator to systematically evaluate a topic model by reading the articles that make up a topic. The main purpose was to be able to apply a credible coherence score on a topic model while using as little evaluator time as possible. We believe that STELLAR aids this purpose reasonably well. Since this is the first evaluation with STELLAR, naturally, there are improvements to be made both to the evaluation process and the tool itself. When doing human evaluation of topic models, usually the concept of an intrusion task is used to identify how coherent a topic is (Chang et al., 2009; Hoyle et al., 2021). This task is not fully transferable to our concept of coherence. However, we believe the incorporation of ideas around the intrusion task would make STELLAR better.

Finally, we believe that using the PLM not only allows for the topic models to stay relevant when new language models are released but also creates a more interpretable vector space for analysis since one can observe what topics are related to each other with visual inspection. This human expert evaluation of NTM-CE concludes that the technique is viable and has many attractive benefits. However, it has some limitations that need to be addressed before being used to its full potential as an automatic classifier.

8 Conclusions

In this study, we applied Neural Topic Modeling by Clustering Embeddings (NTM-CE) made with BERT on an industry dataset of news articles. Our human evaluation of NTM-CE, done with our novel STELLAR tool, agrees with previous studies of the technique: coherent topics can be created by clustering embeddings from a pretrained language model. However, only 57% of the articles ended up in coherent topics. Inspection of incoherent topics revealed them to consist of multiple focus topics, or have some other emotional or stylistic characteristic. Unraveling the workings of incoherent topics to increase the number of articles in coherent topics shows great opportunity for industry application. With the STELLAR tool, we hope to keep improving on NTM-CE as a promising technique for the future.

Ethical Considerations

The dataset was scraped from public news sites of established publishers. No personal blogs were used. The names of the people who are written in the articles are mentioned as public persons. We do not view this as a privacy infringement. The articles are not redistributed.

We identified that our personal biases have an impact on the outcome of the results. Examples are choosing the list of focus topics or determining when to limit a topic. In practice, those choices in turn might have an effect on what type of topic model is deployed in the end. A consideration could be to include a more diverse group of expert evaluators. A model might be too generalizing and fail to identify topics that are associated with marginalized groups or cultures, leading to technology being catered to a homogeneous majority.

Acknowledgements

We thank Adrian Andreasson and Dusan Mitic for being expert evaluators, Frank Drewes and anonymous reviewers for their insightful input, and the rest of the university and Adlede teams. This Ph.D. student is funded by the Swedish Foundation for Strategic Research, project id ID19-0055.

References

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021. Cross-lingual contextualized topic models with zero-shot learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683, Online. Association for Computational Linguistics.

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 113–120.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of the Biennial GSCL Conference*.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? the incoherence of coherence. In *Advances in Neural Information Processing Systems*, volume 34, pages 2018–2033. Curran Associates, Inc.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden. Association for Computational Linguistics.

Laurens Van Der Maaten and Geoffrey Hinton. 2008. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Leland McInnes and John Healy. 2017. Accelerated hierarchical density based clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Topic discovery via latent space clustering of pretrained language model representations. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 3143–3152, New York, NY, USA. Association for Computing Machinery.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ike Vayansky and Sathish Kumar. 2020. A review of topic modeling methods. *Information Systems*, 94:101582.

Xuerui Wang and Andrew McCallum. 2006. Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 424–433, New York, NY, USA. Association for Computing Machinery.

Zihan Zhang, Meng Fang, Ling Chen, and Mohammad Reza Namazi Rad. 2022. Is neural topic modelling better than clustering? an empirical study on clustering with contextual embeddings for topics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3886–3893, Seattle, United States. Association for Computational Linguistics.

He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray Buntine. 2021. Topic modelling meets deep neural networks: A survey. *arXiv preprint arXiv:2103.00498*.

A Human Evaluation Protocol

Instructions for filling in Table

For each Topic in Topics List:

1. Click on the Topic in the Topic list.
2. Notice the keywords describing the Topic.
3. Read the article titles and keywords.
4. Click on the article to read the body if not clear what the topic is about.
5. Choose the main topic from the list. A topic in the list can be chosen multiple times.
   (a) If you can't find a topic that includes 50% of the articles, then choose "no topic"[7].
   (b) If you don't agree with any of the topics in the list, write "custom" and then in the notes write your custom topic. Please make suggestions so that it is not only my biases determining the categories.
6. Write down the article id for articles that do not belong to the topic.
7. Write on a scale (1-4) if you think the keywords are a good representation of the topic. 1=bad, 2=sort of bad, 3=sort of good, 4=good.

While doing the task. Write notes of interesting things that you reflect over. Also, make general notes about what improvements that can be made to the tool.

[7] We have translated 'No topic' to 'Incoherent' when writing the article.