Added Value of Semantic Language Models in Topic Clustering

Anouk van Wingerden
Data Science & AI, CTO
NATO C & I Agency
The Hague, The Netherlands
anouk.vanwingerden@ncia.nato.int

Arvid Kok
Data Science & AI, CTO
NATO C & I Agency
The Hague, The Netherlands
arvid.kok@ncia.nato.int

Michael Street
Data Science & AI, CTO
NATO C & I Agency
The Hague, The Netherlands
michael.street@ncia.nato.int

Ivana Ilic Mestric
Data Science & AI, CTO
NATO C & I Agency
The Hague, The Netherlands
ivana.ilicmestric@ncia.nato.int

***Abstract*— This paper addresses the added value of semantic language models in topic clustering. The benefits from LDA, getting a dirichlet distribution of documents and topics, is extended with BERT embeddings which adds semantic information. In this paper, different clustering algorithms and algorithms for extraction of topic words are compared. The clusters themselves and the representation of clusters by topic words are evaluated by manual assessments and evaluation metrics: coherence score and topic diversity. The results show that taking semantic meaning into account using BERT embeddings significantly improves results of clustering articles.**

Keywords—Topic Extraction, topic words, semantic topic extraction, LDA, BERT, K-means, HDBSCAN, TF-IDF, Natural Language Processing, NLP, Machine Learning, ML

I. INTRODUCTION

Natural Language Processing (NLP) applied in Media Analysis (MA) is gaining interest for collecting valuable insights, otherwise difficult to get and often labor intensive. The key challenge is the ever growing volume of content, which is too big to handle for human analysts. This is where Natural Language Processing techniques can help. Using publicly available sources for assessment demonstrates the value of detecting specific topics across certain geographic areas and tracking trending topics. Analyzing articles one by one would take vast human analyst resources, but more recent algorithms can rapidly provide overviews at a glance.

The research conducted and described in this paper addresses Topic Modeling and, more specifically, the effect of adding semantic language models to the equation. A Topic Model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents [18]. We demonstrate the capabilities of clustering methods and underpin them with evaluation metrics and manual assessments. The following two research questions are leading:

1. Are articles assigned to the appropriate cluster, sharing a common topic?
2. Are the clusters labeled with appropriate words, covering the topics of (all) associated articles?

In our experiments we compare five clustering methods, where each method combines a series of techniques; with or without utilizing a semantic language model. The dataset and tools used, methods applied, and experiments conducted are described in detail in respective chapters.

II. DATASETS

For this research a collection of news articles is used. The dataset consists of news articles coming from The Guardian, NewsAPI, Google News and local news websites, using the (random) search term 'Northampton'. The articles were published between 12 October 2020 and 3 December 2020. In total the set contains 1,254 unique articles.

Figure 1 shows the length (characters) of the articles included in the dataset. Well noticeable is the number of documents with a length between 200-300 characters. This can be explained by NewsAPI truncating articles to a maximum of 260 characters for Developer plan users.

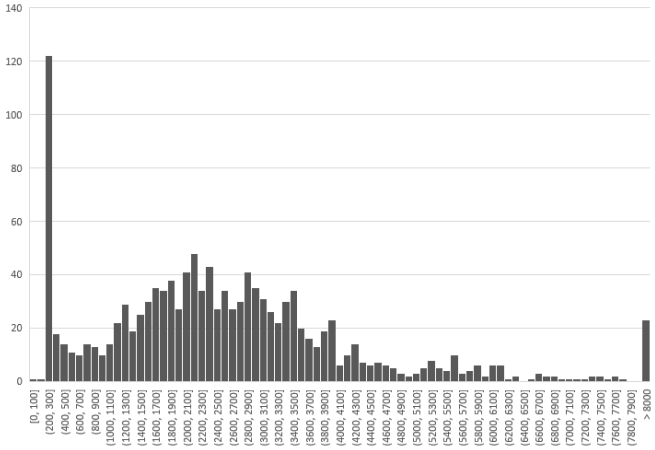


Figure 1 Histogram news article lengths (characters)

I. TOOLS

Anaconda Jupyter Notebook and Python is used to prepare and analyze data as well as to create and evaluate models. The visualizations are made in Microsoft Power BI which gives sufficient flexibility and simplicity to manually evaluate results.

II. METHODS

For the experiments conducted we combined several techniques into five unique methods for clustering the news articles in the used dataset. An overview of the structure of the

five methods is given in Figure 2. Next paragraphs introduce the techniques in the context of the methods where they are applied.

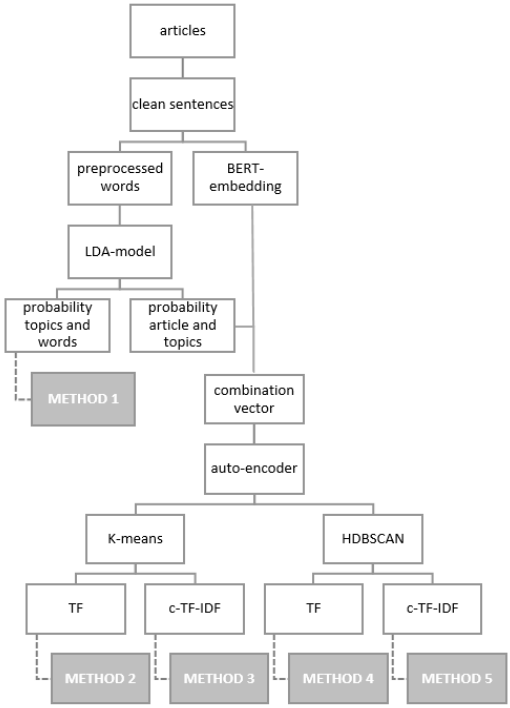


Figure 2 Schematic overview of the establishment of the five methods

A. *Latent Dirichlet Allocation (LDA)*

LDA is one of the well-known algorithms in Topic Modeling. This algorithm of Blei [2] is a generative probabilistic model of a corpus. It aims to bring latent articles in the clusters to which topic words can be attached, depending on how well that word fits in the topic. Each article can be described by a distribution of topics and each topic can be described by a distribution of words. LDA specifically looks at co-occurrence of words, not to their contextual meaning.

For method 1 the dataset is prepared for the Gensim LDA model [13]. Article text is cleaned, words are lemmatized and stop words are removed. Resulting in a bag of words for each article. The LDA model then uses a fixed number of topics and assigns each article; no (unassigned) outliers remain. On top the LDA model returns the highest probability topic words for each topic.

B. *Bidirectional Encoder Representations from Transformers (BERT)*

BERT is a language model, first presented by Vaswani et. al in “Attention is all you need” [17]. Research has shown relevant and state-of-the-art performance on various NLP tasks [4]. BERT embeds articles in a vector space where capturing the contextual meaning, semantics of sentences.

Methods 2, 3, 4 and 5 combine LDA (also used in method 1) with the semantic language model BERT. The document topic probabilities from LDA are extended with the semantic embedding from BERT. For each article the two are concatenated, resulting in an 868 dimensional representation.

C. *Auto-encoder for dimensional reduction*

An auto-encoder is a type of artificial neural network used to learn efficient data encodings in an unsupervised manner [10]. Auto-encoders are a common alternative for dimensionality reduction.

In our experiments an auto-encoder is trained to reduce the LDA+BERT representations from 868 to 32 dimensions. Using the auto-encoder significantly brings down the computational requirements of clustering algorithms in methods 2, 3, 4 and 5.

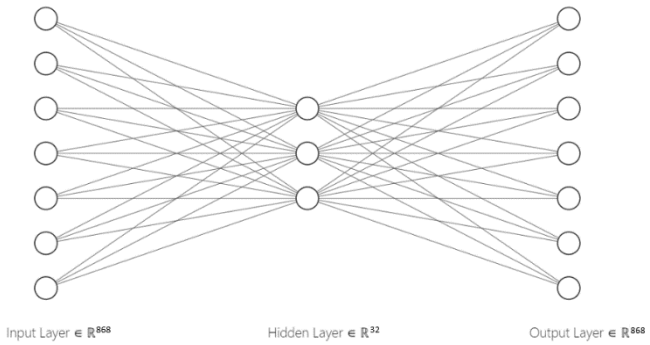


Figure 3 Structure of the auto-encoder used for the clustering of the combination vector of LDA and BERT embeddings

D. *K-means*

For the clustering method, we consider the widely used K-means algorithm, which is a relatively straightforward computation. The algorithm attempts to create K clusters of M articles in remaining N dimensions, by reducing sum of squares (Euclidean distance) within a cluster to a local minimum [7]. It forces every document to belong to a cluster.

As shown in Figure 2, the K-means algorithm has been applied to complete clustering on (dimensionally reduced) LDA+BERT representations in methods 2 and 3. K-means requires the number of clusters to be fixed. For our experiments the number of clusters is fixed to 100. The high number allow for articles not covering a main topic to be parked into spare “outlier” clusters.

E. *HDBSCAN*

Another clustering method is the density-based algorithm HDBSCAN [12]. It does not force all data points to belong to a cluster by considering outliers. The used method for cluster selection is Excess of Mass (EOM), in which the tendency is to choose one or two large clusters and only then re-clusters the large clusters to smaller ones.

Methods 4 and 5 use HDBSCAN to cluster, with a minimum cluster size of 3 articles; smaller clusters to be treated as outliers. Same as with K-means clustering, also here the dimensionally reduced LDA+BERT representations are used for computational performance reasons.

F. *Term Frequency (TF)*

To identify the topic words Term Frequency (TF) can be used. TF counts the frequency of each unique word occurring in a set of documents. [19]

Methods 2 and 4 obtain topic words using TF. All articles from a cluster are grouped together. The frequency for each unique word is counted. The 20 words with highest frequency for each topic are selected as topic words.

G. *Context Term Frequency-Inverse Document Frequency (c-TF-IDF)*

To identify the topic words, combination of Term Frequency (TF) and Inverse Document Frequency (IDF) – together TF-IDF – will be used. TF-IDF assigns each unique word a weight according to the uniqueness compared to other documents. In other words, TF-IDF captures the relevance among words, texts and particular clusters [18]. A variant to this is context-Term Frequency-Inverse Document Frequency (c-TF-IDF) [9], which makes it possible to extract what words make each cluster unique. This function includes the intuition that (i) the more often a term appears in a document, the more representative it is of its content, and (ii) the more documents a term appears in, the less discriminating it is [16].

Methods 3 and 5 obtain topic words using c-TF-IDF. The 20 words with highest scores for each topic are selected as topic words.

III. EXPERIMENTS

We the split assessment of the different approaches in two: a) how well are articles clustered, b) how representative are topic words per cluster. Former is assessed using distribution number of articles in clusters and a manual review; latter is assessed using five complimenting measures.

A. *Distribution number of articles in clusters*

Figures below show distributions of articles across identified clusters. For LDA and K-means based methods limited to largest 40 out of the fixed 100 clusters; for HDBSCAN all clusters, excluding outliers. Each method shows on the right how many articles are outside biggest 40 clusters.

The LDA model has one cluster with most articles, and numbers decrease immediately. K-means has smallest curve of all three clustering algorithms. Here the number of articles in the 40 largest clusters varies between 11.6% and 0.9%. HDBSCAN starts with a fast drop after first cluster, but steadily decreases after.

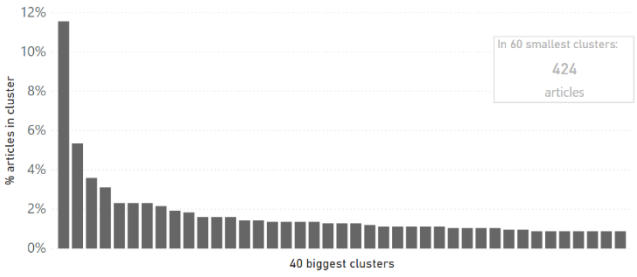


Figure 4 Distribution of the number of articles assigned to a particular cluster with LDA

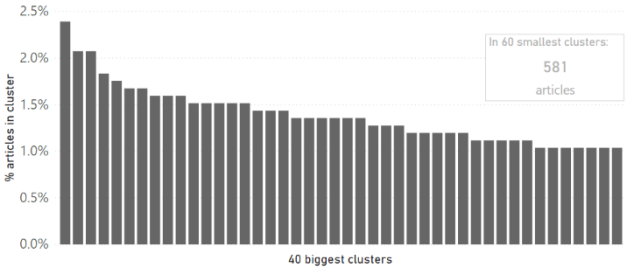


Figure 5 Distribution of the number of articles assigned to a particular cluster with the LDA+BERT vector clustered by K-means

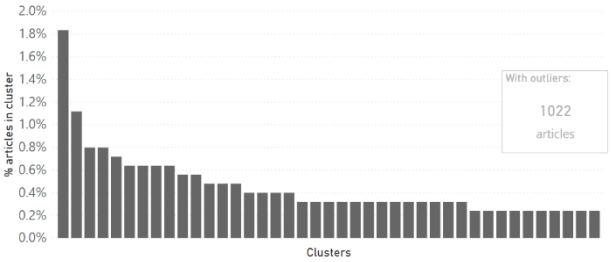


Figure 6 Distribution of the number of articles assigned to a particular cluster with the LDA+BERT vector clustered by HDBSCAN

Articles in an extremely large cluster can represent many generic topics, or indicate a single topic. HDBSCAN considering many articles as outliers can be an indicator that the smaller clusters are formed more accurately than in other clustering algorithms. From these results, the following two questions emerge: Will topic words of large clusters be able to represent the common topic? Do HDBSCAN clusters contain more specific topics than other clustering algorithms?

B. *Manual review clusters*

Besides using evaluation metrics to assess the performance of each method, we find that a human assessment of clusters is equally important. Each method is assessed in following manner. The five largest clusters are selected. For each cluster, five randomly chosen articles are read and major subjects are determined. The score (last column of Table 1) is assigned according to number of articles fitting the major topic. By this, the following question is answered: Does the article in the cluster actually belong to that cluster from a human perspective? So, did the model get it right?

TABLE 1 FIVE BIGGEST TOPICS FOR EACH CLUSTERING ALGORITHM, MANUALLY DEFINED THE MAJOR TOPIC AND THE NUMBER OF ARTICLES WHICH FITS IN TOPIC

Method	***Topic id***	***Major defined topic***	***Number of articles fits in defined topic***
LDA	1	arrest	4/5
	11	online events	3/5
	25	impact on media business	5/5
	43	charity/social	2/5
	45	economic effects of covid	5/5
K-means	10	protest	5/5
	24	crash	5/5
	4	incidents on the road	5/5
	48	engineering works	4/5
	77	covid and schools	5/5

Method	*Topic id*	*Major defined topic*	*Number of articles fits in defined topic*
HDBSCAN	1	rugby	5/5
	19	covid cases	5/5
	23	vehicle crash	5/5
	24	covid increase	5/5
	27	on the road with alcohol and drugs	5/5

HDBSCAN outperforms K-means and largely outperforms LDA. For the biggest clusters, all samples fits in major defined subjects. K-means only has one topic that does not fit all samples.

Interesting to see is the influence of COVID-19 related footer occurring a larger subset of articles. Figure 7 shows the cluster assignment of these articles. LDA puts a large portion of these articles into a single cluster, where BERT-based clustering with K-means and HDBSCAN show more variance.

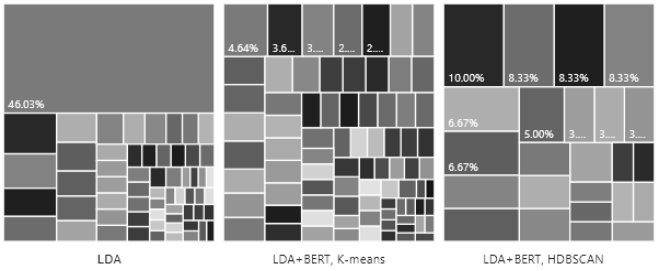


Figure 7 Number of articles in a topic filtered by articles with the footer statement for each clustering algorithm

C. *Article coverage by topic words*

The percentage of articles that contains topic words from the cluster. The overall result are presented by method in Table 2, followed by tables on the results zooming in on the top 5 clusters per method.

TABLE 2 COVERAGE OF TOPIC WORDS IN ARTICLES FOR EACH METHOD

Method	Number of articles with topic words (avg)	Number of articles in cluster (avg)	Percentage covering
LDA	7.46	12.69	59%
LDA+BERT, K-means, TF	7.13	12.54	57%
LDA+BERT, K-means, c-TF-IDF	3.37	12.54	27%
LDA+BERT, HDBSCAN, TF	16.22	29.86	54%
LDA+BERT, HDBSCAN, c-TF-IDF	7.28	29.86	24%

The slightly higher coverage percentage of LDA can be explained by the way the technique a probability matrix of topics and words to get to the cluster split.

TABLE 3 COVERAGE OF TOPIC WORDS IN ARTICLES WITH LDA

Topic id	Number of articles in cluster	Percentage covering
25	145	96%
11	45	64%
1	39	49%
43	29	62%
45	29	64%

TABLE 4 COVERAGE OF TOPIC WORDS IN ARTICLES WITH BERT+K-MEANS

Topic id	Number of articles in cluster	Percentage covering TF	Percentage covering c-TF-IDF
10	30	39%	12%
4	26	54%	36%
24	26	53%	41%
77	23	68%	53%
48	22	47%	9%

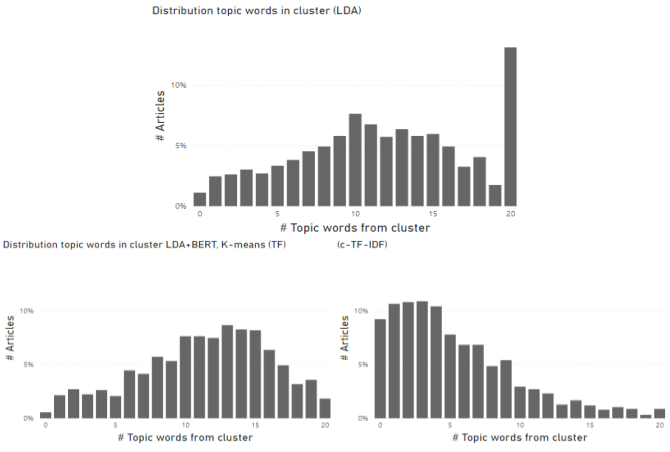
TABLE 5 COVERAGE OF TOPIC WORDS IN ARTICLES WITH BERT+HDBSCAN

Topic id	Number of articles in cluster	Percentage covering TF	Percentage covering c-TF-IDF
24	23	83%	73%
19	14	83%	64%
1	10	44%	30%
27	10	73%	65%
23	9	69%	63%

The increasing covering percentage when only filtering on biggest clusters of HDBSCAN c-TF-IDF is notable. Possibly topic words in smaller clusters are too specific to a subset of articles.

D. *Coverage in articles*

Figure 8 show histograms of the number of topic words found in articles. From left to right, occurrence of topic words present in articles associated with the cluster increases from 0 (none) to 20 (all). LDA shows a relative high number of articles that contain all topic words. LDA is based on term counts and document counts but does not take into account unique words for each cluster when assigning the weights.



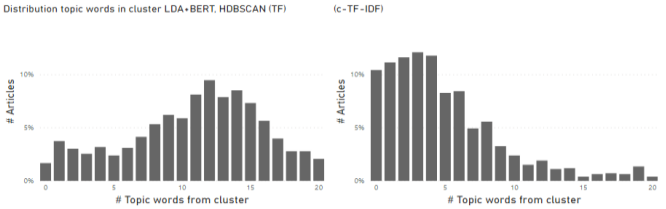


Figure 8 Distribution number of topic words from cluster in articles

The distribution of K-means TF is skewed towards the right, where K-means c-TF-IDF has many articles on left side of the distribution. Latter means that the number of topic words found in articles is often very low. Due to BERT embeddings, we expect similar articles being clustered together, but c-TF-IDF might not always reflect this.

HDBSCAN, compared to K-means and LDA has higher but steep peaks and is more distributed to left. Especially by c-TF-IDF, many articles contain a few topic words.

We narrow and examine number of topic words for each article in each cluster. The figures below will provide an indication of the variance of each cluster. For each of 40 biggest clusters, number of topic words consist in article is shown. We assume an optimal situation to have all article covered with some topic word.

We see LDA is wide spread. The uncertainty of wide spread is whether articles are in the right cluster and whether the topic words are relevant.

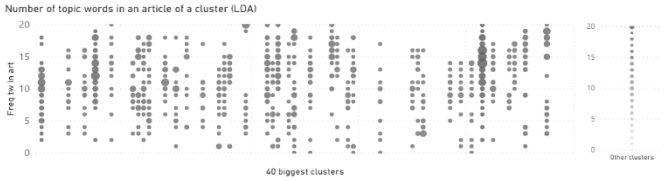


Figure 9 Number of articles (buble size) with number of topic words (y-axis) per cluster (x-axis) for LDA method

For 40 biggest clusters, K-means TF shows many points in upper half. Number of topic words in articles for the smaller clusters is high. We believe K-means c-TF-IDF has a small number of clusters with high scores. This can possible be because of taking semantics into account in the method.

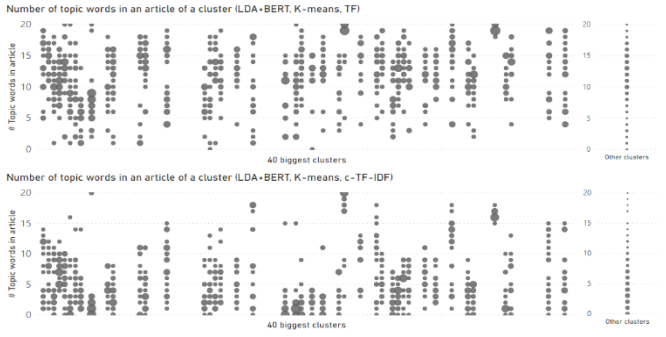


Figure 10 Number of articles (buble size) with number of topic words (y-axis) per cluster (x-axis) for K-means with TF (top) and c-TF-IDF (buttom)

For HDBSCAN we see many articles are plotted in upper half using TF; using c-TF-IDF the largest cluster has most of points at top, where others have small positive correlation between the cluster size and number of topic words in articles.

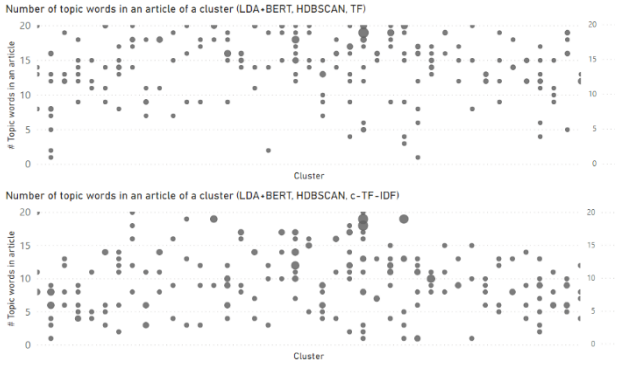


Figure 11 Number of articles (buble size) with number of topic words (y-axis) per cluster (x-axis) for HDBSCAN with TF (top) and c-TF-IDF (buttom)

E. Interpretability and representation of topic words

It is important to measure whether topic words interpretable by a human. For this purpose, a subject is manually derived for the five largest clusters. Five random articles of each cluster are read to see if chosen subject fits the five read articles. Example results are shown in Table 6 and Table 7.

TABLE 6 LDA – MANUALLY REVIEWING TOPIC WORDS OF THE FIVE BIGGEST CLUSTERS WITH 5 RANDOM ARTICLES OF EACH CLUSTER. FOR THE LAST THREE COLUMNS, THE HIGHER IS BETTER.

Id	Topic words	Defined subject by topic words	Match of articles	Ease (1=diffi-cult, 5=easy)	Specificity (1=gene-ral, 5=specific)
25	also, local, provide, u, thank, copy, newspaper, news, help, now, make, event, order, please, continue, website, receive, ask, site, important	Event	1/5	1	1
11	council, christmas, will, s, child, year, northamptonshire, support, say, help, can, county, people, need, community, local, charity, make, new, give	Charity	3/5	4	3
1	police, man, officer, incident, northamptonshire, call, anyone, arrest, say, northampton, road, drug, stop, information, black, men, witness, offender, two, old	Incident with police	4/5	5	4
43	say, s, people, will, can, get, one, go, t, vaccine, service, make, work, need, help, health, time, call, come, day	Health	1/5	3	2
45	say, s, uk, government, year, lockdown, will, economy, business, job, pandemic, sunak, covid, cut, support, month, people, economic, rise, scheme	Effects of Covid-19	4/5	4	3

TABLE 7 HDBSCAN + C-TF-IDF – MANUALLY REVIEWING TOPIC WORDS OF THE FIVE BIGGEST CLUSTERS WITH 5 RANDOM ARTICLES OF EACH CLUSTER. FOR THE LAST THREE COLUMNS, THE HIGHER THE BETTER.

Id	Topic words	Defined subject by topic words	Match of articles	Ease (1=diffi-cult, 5=easy)	Specificity (1=gene-ral, 5=specific)
24	death, patient, hospital, nh, general, positive, reveal, coronavirus, care, among, confirm, link, bulletin, previous, rise, certificate, kettering, acute, commission, weekly	Covid care in hospital	5/5	4	5
19	student, isolate, self, bubble, parent, letter, positive, school, stopford, staff, test, pupil, inform, symptom, close, learn, silverthorne, wrenn, birkett, bishop	Effects and tests covid	5/5	3	5
1	reuters, leinster, lange, char, rugby, botham, cup, file, marchant, glamorgan, exeter, lloyd, williams, autumn, debut, flanker, barbeary, saint, lineup, scrumhalf	Rugby	4/5	5	5
27	guilty, plea, surcharge, fin, breath, victim, namely, chmielewski, pay, age, alcohol, cost, exceed, drove, ban, limit, drive, prescribed, consume, sentence	Drunk driving	5/5	4	5
23	junction, delay, traffic, luton, southbound, lane, keynes, highway, milton, morning, ques, near,	Traffic delay morning	5/5	3	5

Id	Topic words	Defined subject by topic words	Match of articles	Ease (1=difficult, 5=easy)	Specificity (1=general, 5=specific)
	crawl, broken, rush, warn, tailback, driver, mile, congestion				

The defined subject scores high in HDBSCAN c-TF-IDF. LDA has least number of articles matching. The other results from this table are explained in next paragraph.

It is important to generate topic words which are easy to interpret into a subject and which are not too general. Difficulty in generating topic words and the degree of specificity is shown in the last two columns of Table 6 to Table 7.

Ease refers to what it takes to manually derive a single topic with given topic words. Specificity refers to the abstract-level of the topics words. In both measurements topic words of K-means c-TF-IDF and HDBSCAN c-TF-IDF are easiest to bend to a single topic and are also most specific. LDA turns out to result in more general topic words.

F. Coherence score and Topic Diversity

Table 8 shows the coherence scores and topic diversity. Higher coherence is in method LDA with BERT, K-means c-TF-IDF. Clustering with HDBSCAN yield better topic diversity but not a significant improvement in coherence score.

TABLE 8 COHERENCE SCORE AND TOPIC DIVERSITY OF TOPIC WORDS IN THE METHODS. FOR BOTH, THE HIGHER THE BETTER.

Method	Coherence score	Topic diversity
LDA	0.35290	0.38350
LDA+BERT, K-means, TF	0.38738	0.26000
LDA+BERT, K-means, c-TF-IDF	0.45942	0.76450
LDA+BERT, HDBSCAN, TF	0.42686	0.44405
LDA+BERT, HDBSCAN, c-TF-IDF	0.45096	0.87024

IV. CONCLUSION

This paper addresses the added value of semantic language models in topic clustering. Five methods were evaluated with derived sub-questions: "Which method is most suitable for creating clusters?" and "Which method is most suitable for representing clusters with topic words?" From the results we can conclude that taking semantic meaning into account by using BERT embeddings significantly improves clustering of articles.

Results show that HDBSCAN best generates clusters of articles. TF returns (too) generic topic words in both K-means and HDBSCAN; LDA returns topic words that are difficult to interpret.

V. RECOMMENDATIONS

Further research is recommended to investigate techniques for selecting topic words that take into account the strength of similar words.

Effects related to length of articles were excluded in this research. This might be a subject for future work.

Due to the current situation, many articles focus on COVID and topics are often linked to this. It will be interesting to measure the impact of pandemic on the overall information environment and compare it to pre/post pandemic time.

References

[1] Aletras, N., & Stevenson, M. (2013, March). Evaluating topic coherence using distributional semantics. In Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers (pp. 13-22).

[2] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[3] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. Advances in neural information processing systems, 22, 288-296.

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[5] Dieng, A. B., Ruiz, F. J., & Blei, D. M. (2020). Topic modeling in embedding spaces. Transactions of the Association for Computational Linguistics, 8, 439-453.

[6] Hahnloser, R. H., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J., & Seung, H. S. (2000). Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. Nature, 405(6789), 947-951.

[7] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. Journal of the royal statistical society. series c (applied statistics), 28(1), 100-108.

[8] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[9] Li, X., Wang, D., Li, J., & Zhang, B. (2007, July). Video search in concept subspace: a text-like paradigm. In Proceedings of the 6th ACM international conference on Image and video retrieval (pp. 603-610).

[10] Liou, Cheng-Yuan; Cheng, Wei-Chen; Liou, Jiun-Wei; Liou, Daw-Ran, "Autoencoder for words", https://doi.org/10.1016%2Fj.neucom.2013.09.055, 2014

[11] Malzer, C., & Baum, M. (2020, September). A Hybrid Approach To Hierarchical Density-based Cluster Selection. In 2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI) (pp. 223-228). IEEE.

[12] McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. Journal of Open Source Software, 2(11), 205.

[13] Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks.

[14] Röder, M., Both, A., & Hinneburg, A. (2015, February). Exploring the space of topic coherence measures. In Proceedings of the eighth ACM international conference on Web search and data mining (pp. 399-408).

[15] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5), 513-523.

[16] Sebastiani, F. (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), 1-47.

[17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

[18] Wikipedia: https://en.wikipedia.org/wiki/Topic_model

[19] Yun-tao, Z., Ling, G., & Yong-cheng, W. (2005). An improved TF-IDF approach for text classification. Journal of Zhejiang University-Science A, 6(1), 49-55.