Proceedings of the 21st International Conference on Modeling & Applied Simulation, 009
18th International Multidisciplinary Modeling & Simulation Multiconference

2724-0037 © 2022 The Authors.
doi: 10.46354/i3m.2022.mas.009

Modelling shifting trends over time via topic analysis of text documents

Oliver Krauss[1,*], Andrea Aschauer[2] and Andreas Stöckl[3]

[1]Advanced Information Systems and Technology, University of Applied Sciences Upper Austria, Softwarepark 13, Hagenberg, 4232, Austria
[2]Playful Interactive Environments, University of Applied Sciences Upper Austria, Softwarepark 13, Hagenberg, 4232, Austria
[3]Digital Media Department, University of Applied Sciences Upper Austria, Softwarepark 11, Hagenberg, 4232, Austria

[*]Corresponding author. Email address: oliver.krauss@fh-hagenberg.at

Abstract

Generating new business concepts is an important part of founding new start-ups as well as innovation in existing companies. We identify rising and falling trends in different domains, via topic modelling of text data published over time. Topic modelling is an important tool to classify and cluster documents. Based on the top2vec method, which uses common embeddings of documents and words to not only find but also describe clusters, we have implemented an incremental variant, tracing of growth and decline of these clusters over time. Identification of these trends over large text data collections enables a decision support for innovators, by identifying rising trends, and declining opportunities in different domains. The method was tested and evaluated on the example of arXiv articles. Visualizations of the clusters and descriptions serve to provide people with an interface to identify the trends. In the future, this method can build a foundation of decision support systems that generate innovation ideas based on upcoming trends in research.

Keywords: Decision Support; Topic modelling; NLP; Visualization

1. Introduction

We model the rise and decline of topics over time, mined from large scopuses of text documents. Our work is based on the top2vec (Angelov, 2020) method, which enables clustering of text bodies into topics.

The change over time of the topics occurring in text documents reflects the trends developing in society or science, depending on the scopus under analysis. The identification of topics and trends is not possible, or only possible with enormous effort for humans, due to the large amounts of available data. To identify topics and trace their development over time, algorithms must be used for analysis, and appropriate visualizations must be created to support decision making by stakeholders.

The aim of this work is to develop a suitable method to identify trends in the change of topics and to present them by means of visualizations. Figure 1 shows the current state of the art (top) with an added trend analysis (bottom). The goal being the support of innovation management to either propose interesting Startup ideas in topics that are on the rise, or to penetrate larger markets with new products of companies. The use of artificial intelligence to support expert decisions in this area is an up-and-coming trend in research (Mühlroth, 2020; Nazemi et al., 2022).

Over the next sections, we present a summary of the top2vec algorithm in section 2, as well a comparison to related work in this domain in section 3. In section 4 we show the extension of top2vec to enable analysis of trends over time. We then present preliminary results in section 5,

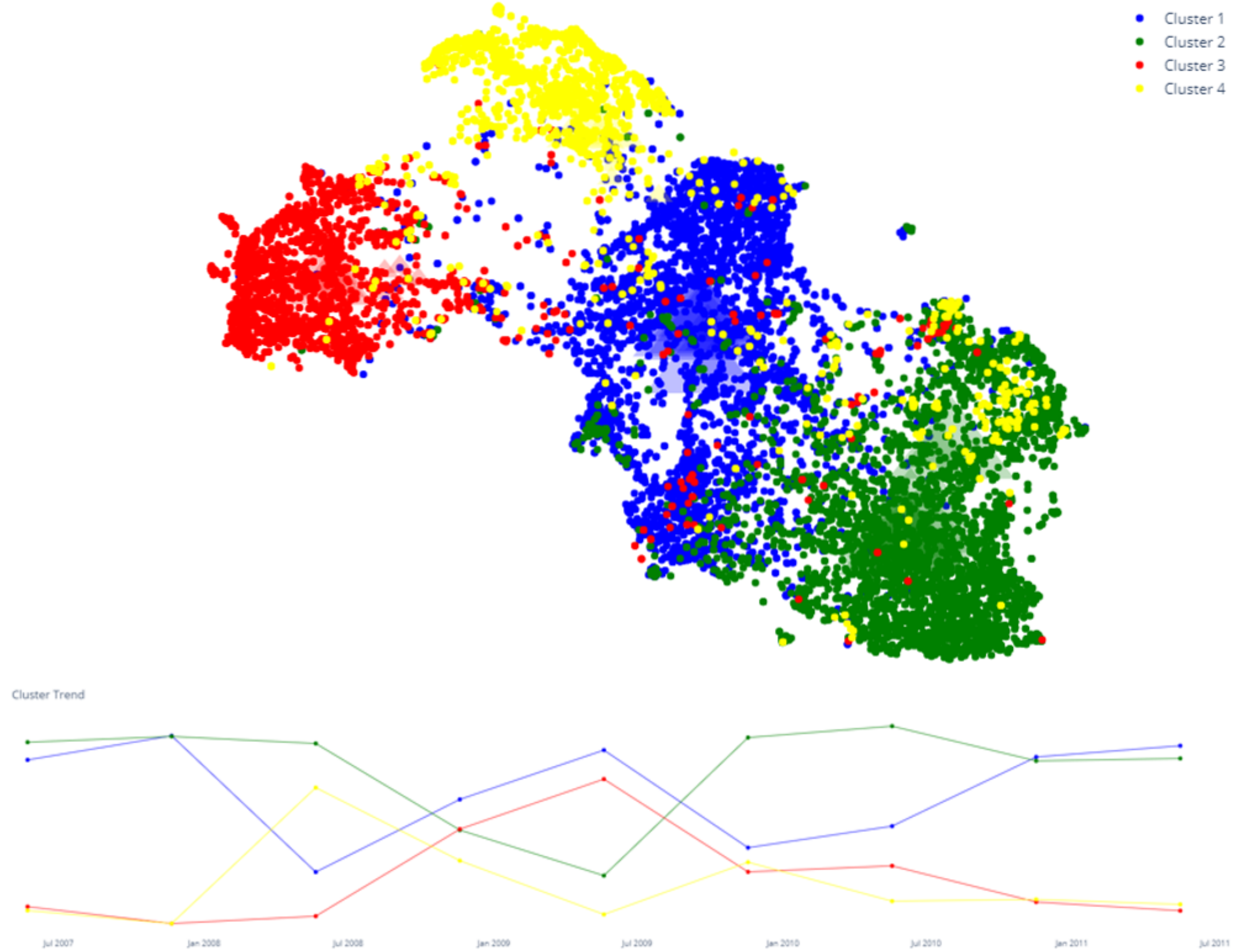

Figure 1. Topic modelling (top) with trend analysis (bottom).

including visualizations via different information graphics, such as bubble charts of projections of the document representations or Sankey diagram. Finally in section 6 we summarize our findings and outline future work.

2. Background

We base our work on top2vec (Angelov, 2020), a novel algorithm for mining topics via mining documents. The top2vec algorithm has multiple advantages over comparable algorithms in the domain, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) or Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999). In the original publication, Angelov shows that top2vec finds topics that are more informative and understandable than LDA or PLSA. In addition the top2vec algorithm automatically detects the number of topics.

top2vec achieves these results by taking into consideration the ordering and semantics of words, thus not removing stop words, or applying generic preprocessing usual in text mining and clustering such as stemming or lemmatization. Instead top2vec uses the doc2vec (Le and Mikolov, 2014) algorithm to featurize documents into vectors, which in turn is based on word2vec (Mikolov et al., 2013).

After featurization, the document vectors are reduced in dimension via Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) (McInnes et al., 2018). After this dimension reduction, the documents are clustered via Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (Campello et al., 2013).

After the clustering process has finished, the centroids of the topics are calculated, vectorized and then deduplicated to remove overly similar topics. Finally, the most relevant words for each topic are selected and documents are assigned to a topic.

This process allows top2vec to cluster semantically related documents into larger clusters, while also automatically selecting the amount of clusters via HDBSCAN (Angelov, 2020).

What top2vec does not support is the modification or filtering of the source data, which is necessary to extend the process of identifying topics, to identifying trends in topics. We extend top2vec into an incremental version by adding an option to add additional documents, as well as filtering and recalculating the topics based on time constraints.

3. Related Work

The discussion of the related work serves to answer two questions. First why we base our approach of top2vec and not another topic mining algorithm. Second to compare to other approaches for identifying trends.

Topic retrieval in text documents is an extensively studied subfield of NLP with many established methods such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and newer approaches such as (Angelov, 2020), (Grootendorst, 2020). The newer approaches use text representations via word and document embeddings (Mikolov et al., 2013), (Pennington et al., 2014), (Le and Mikolov, 2014) and (Devlin et al., 2018). A good overview of applications of the LDA process is provided by (Jelodar et al., 2019). top2vec is already shown to outperform these methods in (Angelov, 2020).

Another comparable approach to top2vec is BERTopic (Grootendorst, 2020). BERTopic is rather similar to top2vec, and primarily differs in the use of Term Frequency - Inverse Document Frequency (TF-IDF) (Joachims, 1996) as embedding. Instead of assuming that documents generally lie around a centroid concerning their words, BERTopic attempts to categorize the words and use their assigned classes in TF-IDF. In the publication BERTopic is shown to outperform top2vec on mining topics. The primary reason we do not use BERTopic, is the capability of doc2vec to add additional data after initial embedding to the embedding model.

(Egger and Yu, 2022) gives an additional comparison of topic modelling algorithms including BERTopic, Top2Vev, LDA and non-negative matrix factorization (NFM) (Hoyer, 2004). They analyze these algorithms on a dataset of twitter posts. In this use case they show that BERTopic also outperforms top2vec, and give an overview of advantages and disadvantages of the approaches. They explicitly note that top2vec works on larger datasets and that BERTopic has the disadvantage of only assigning a single topic to each document, which often is not the case. This single label may also be the reason why BERTopic does not result in as many topics, and topics being less similar in appearance than top2vec tends to generate. As we work on the arXiv dataset top2vec is better suited to our approach.

To be able to interpret the topics, visualizations of the results of these methods are important. A basic foundation for this kind of visualizations has been laid by (Sievert and Shirley, 2014), (Zou and Hou, 2014), for example.

These methods have been refined and applied in different areas. A common application area is text data from social networks (Liu and Jansson, 2017), (Stöckl et al., 2020), (Krstić et al., 2019). Another frequently considered area is documents from the health sector (Di Corso et al., 2019).

We extend this research by investigating how topic modelling methods and visualizations can be used to find and display trends. A similar goal was pursued by (Chen et al., 2017) with data from the medical field.

4. Methods

We base our work on the top2vec algorithm. top2vec is an unsupervised topic modelling approach based on word and document embeddings. It automatically detects the topics present in the text without the need to specify the number of clusters and generates jointly embedded topic, document and word vectors. top2vec assigns documents to multiple clusters by match percentage, enabling the subsequent analysis of documents in multiple topic contexts.

We extend top2vecs algorithm in two ways, to enable identifying shifts in trends over time. First, we modify the algorithm to consider the entire dataset in frames, e.g. subsets of the data, dependent on their creation time, and analyze clusters via a sliding window approach. Sliding windows are regularly used in data science to mine or learn subsets of data, and in topic modelling have been previously used to analyze topics by text lenght (He et al., 2017). Secondly we enable the option to add data to an already learned topic model. This is primarily to consider additional data sources at a later date, or to add new data to the existing model, e.g. training it at one point and adding information for a month before retraining.

Figure 2 shows the general approach of topic modelling, on the left, that most, if not all, topic modelling algorithms such as top2vec and BERTopic take (Angelov, 2020; Egger and Yu, 2022). This entails the embedding of the entire source data, then conducting a dimension reduction from the word (or concept) embeddings, after which an clustering algorithm identifies topics. Algorithms in this domain tend to differ more in the final step, the derivation of topics. top2vec (second column in Figure 2) conducts a deduplication of overly similar topics, then derives the words defining the topic, and finally assigns documents to the topics.

To enable mining trends in these topics we did not change any of these steps, but rather extended the existing approach (rightmost column in Figure 2). This may also allow us to apply this approach to other topic modelling algorithms in the future, as the core concepts of the topic modelling algorithms are not modified.

4.1. Detecting Trends With top2vec

To detect trends with top2vec, the data has to be ordered on a time axis, and split into frames along it (step "split data over timeframes" in Figure 2). As we analyze arXiv

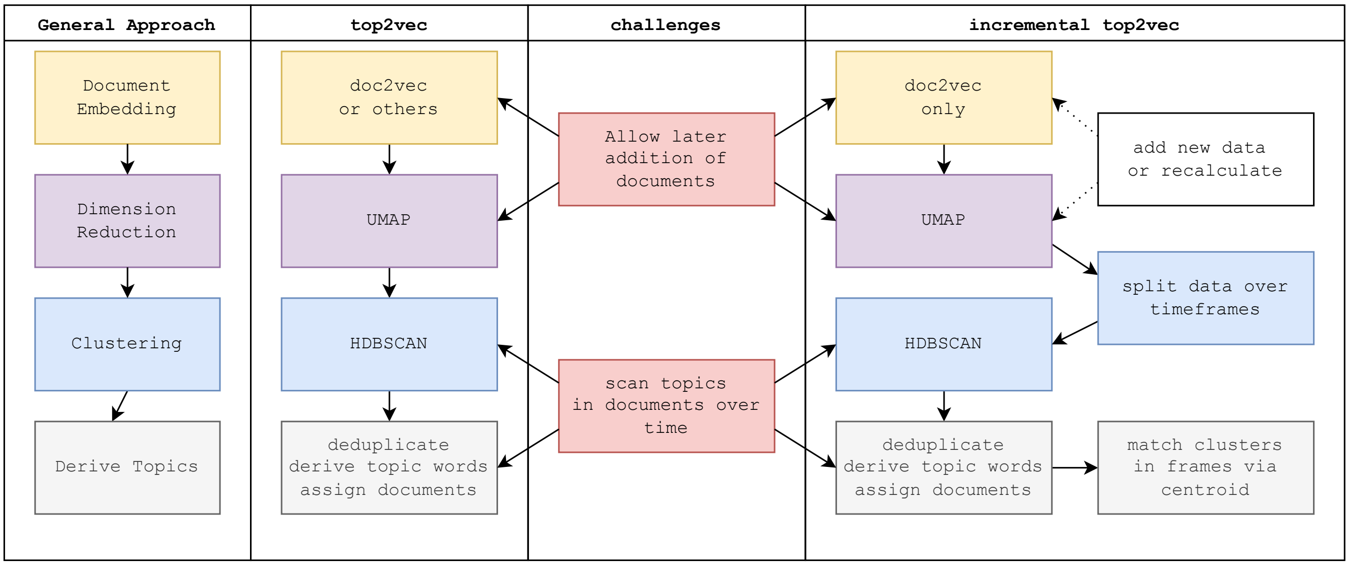
General Approach
Document Embedding
Dimension Reduction
Clustering
Derive Topics
top2vec
doc2vec or others
UMAP
HDBSCAN
deduplicate derive topic words assign documents
challenges
Allow later addition of documents
scan topics in documents over time
incremental top2vec
doc2vec only
add new data or recalculate
UMAP
split data over timeframes
HDBSCAN
deduplicate derive topic words assign documents
match clusters in frames via centroid

Figure 2. General approach of the incremental top2vec algorithm, to enable identification of changes in topics over time based on the dataset.

data, we use the date the source document was last modified. We then create data frames along this timeline, in the concept of a sliding window. This requires the choice of two variables. The increment size, and the frame size. The increment size determines how far the window is moved with every increment. The frame size determines how much data is viewed per frame. As per usual in a sliding window approach, a single data point, i.e. document, can occur in multiple frames.

As an example, if one were to analyze documents from 2014 to 2022, with an increment size of 1 month, and a frame size of 6 months, the first slide would entail all documents from January 2014 to June 2014. The second slide would entail all documents from February 2014 until July 2014 and so on. Except for the first and last few slides, each document is contained in 6 slices.

After these frames are created, the clustering and topic modelling parts of the top2vec algorithm (Clustering and Derive Topics steps in Figure 2) are run individually for each frame, identifying topics individually over time.

Finally we match these clusters via their centroids as defined by top2vec in the UMAP space. top2vec will usually generate between 2 and 100 clusters, primarily depending on settings and the amount of documents provided. As this is a fairly low amount of points to validate, we calculate the distances between all points of frame A to all points of frame B in a brute-force approach and match the clusters between the frames by their minimal distance, in order of size (largest cluster first, down to smallest).

Of course, each frame may have more or fewer clusters than the preceding or following frame. In this case we currently assume that new topics have emerged, or that a topic has died out entirely. This, however, is a point for future work, as in reality topics often separate into new groups, or merge because of their overlap. Modelling this information in the future, is an interesting area of research, as this could allow us to model the emergence and interaction of different research domains based on arXiv in the future.

The question arises, if the core assumption of each cluster being matched to another cluster by matching the minimal distance of centroids holds true, as there could be such severe skips in distance that an entirely new topic should be assumed. At least in the arXiv data this does not occur, possibly because topic modelling on large data sets always allows forming some sort of connex to another cluster over time. Another possibility might be the nature of the documents. As arXiv primarily hosts scientific information, which is based on a host of publications building upon each other to extend the state of the art, no large leaps in topics exist. This may be different for other domains in which one might want to conduct topic analysis.

Figure 3, shows how this extension of top2vec works, on a subsample of the arXiv data set. The darker colors in the image represent older arXiv publications, with lighter colors representing newer publications. The large circles show the centroid of the cluster which moves from the middle-right of the space, towards the lower right over time. This is the core reason why we introduce our modifications to top2vec between the dimension reduction, and clustering steps. Since all documents have been embedded together, we can map each cluster over time in the same dimensional space, allowing us to present understandable visualizations for decision support.

4.2. Adding Additional Data Points

Topic modelling algorithms are cost intensive, due to the embedding and clustering of thousands or millions of documents. Since our method requires repeated clustering in addition to this already extensive process, we also take into consideration how to handle smaller changes in the data

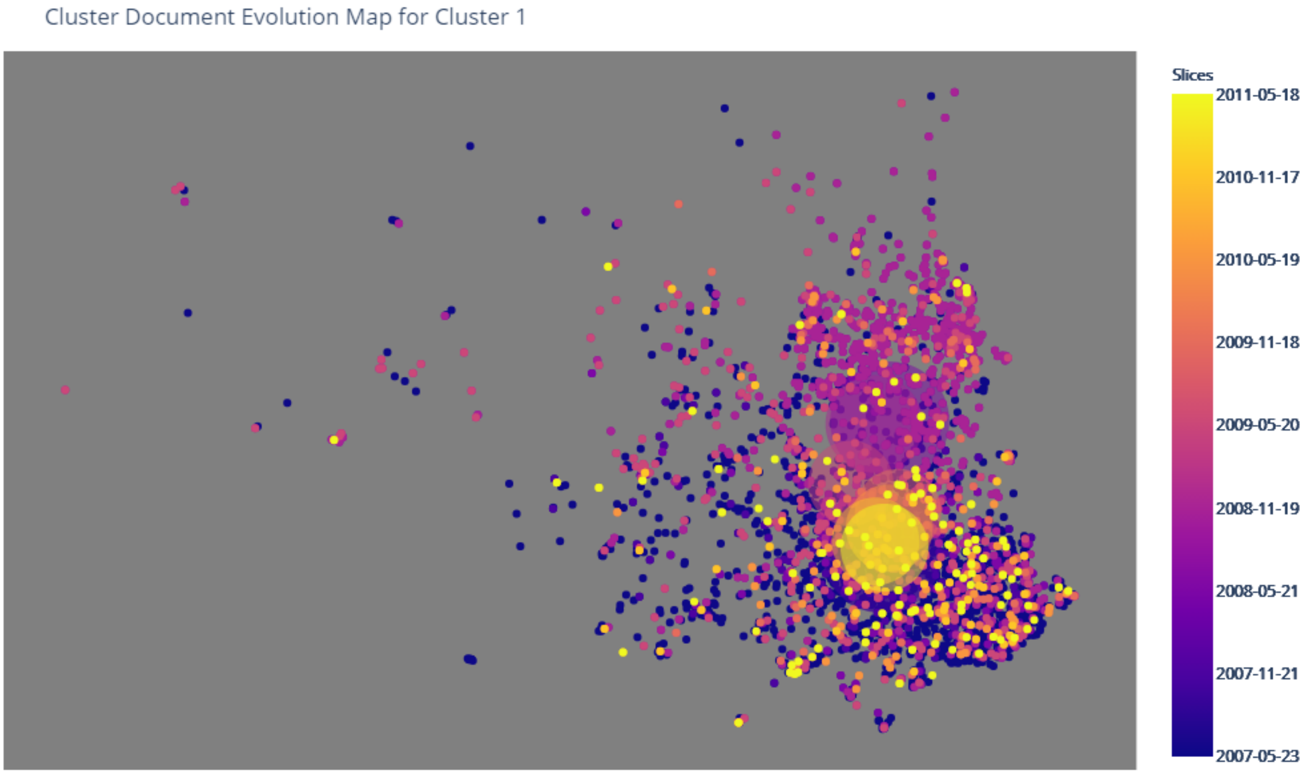

Figure 3. Evolution of a single topic over time. Larger circles indicate the centroids, while the coloring of the dots indicates the age.

set. The purpose is to train the topic model on the entire available data, and then expand in monthly intervals.

We use the existing capabilities of doc2vec and UMAP to add new data to an embedding. In this case doc2vec provides the embedded information, that we add to UMAP for dimension reduction, enabling the representation of the new data points within the existing framework. If the new documents are not more than 10% of the data overall, we do not recalculate the clustering, but instead assign the new data to existing clusters, in each frame, possibly adding a new frame if the documents occur outside the currently existing frame. In case the datapoints exceed 10% of the data overall, we recalculate the entire topic model, beginning from the embedding. This is another point for future improvement, as suggested by literature outlier analyis or other methods could be chosen to dynamically make the decision to recalculate Hassani (2015, 2019).

5. Results

The novel extension of top2vec with a sliding window mechanism allows us to trace shifting trends in topics over time. To enable use of this method for innovation management, we model the following information in a decision support system:

- Trends of topics over time. I.e. Which topic is becoming more prevalent, and which topic is becoming less relevant.
- Word evolution, e.g. how the key topic words evolve over time.
- Cluster connections, of key overlaps between the word vectors of documents in different clusters.

For the trend of a topic over time (see bottom Figure 1), we simply use the amount of documents primarily assigned to a cluster per time frame. In the future, we hope to extend this approach with a more valuable metric. In the case of arXiv data, this may be the amount of times a publication was cited, although this may result in tail end scarcity, as newer publications will have fewer citations. Other domains may use the engagement factor of web-content, or other performance indicators.

To help give deciders an understanding of what a topic cluster is about, we provide an analysis of word vectors in the clusters. Figure 5, shows the keywords per cluster over all time. For decision-making these words can also be visualized in the embedding space over time (not shown as this is too convoluted without possibility of interaction). Corresponding flow diagrams (see Figure 4), allow the analysis of a topic vector over time. Via the words making up the cluster deciders can also gain an understanding of more subtle shifts in focus of a large research domain.

To visualize the overlap of clusters a sankey diagram shows the overlap of key information. In the case of Fig-

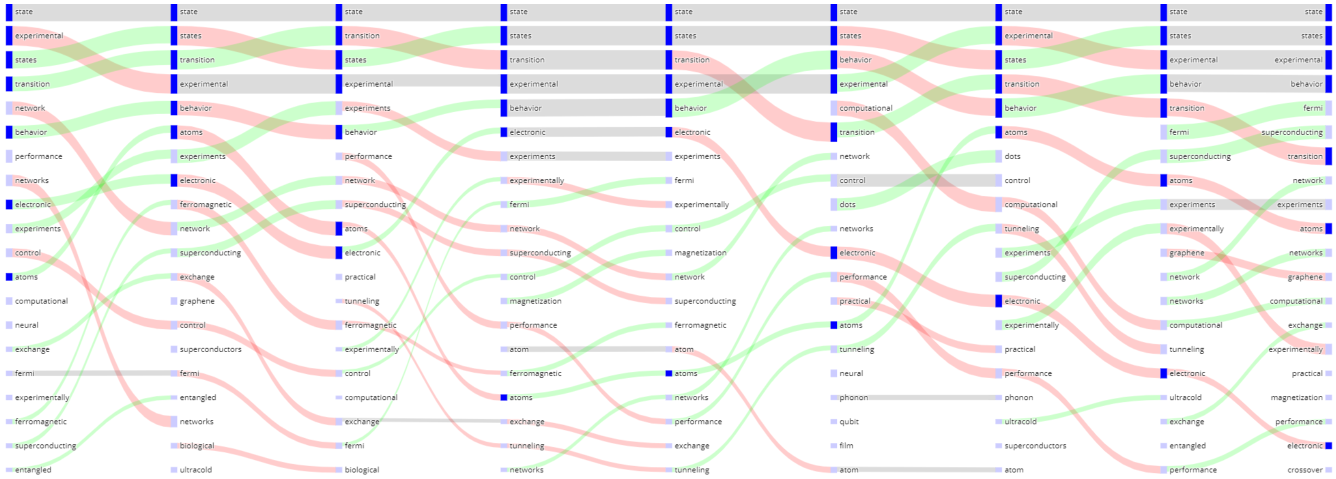

Figure 4. Evolution of key words in a cluster over time. Green lines indicate a word becoming more prevalent, red indicates it becoming less relevant.

ure 6 we have chosen to show the deciders the manually labelled categories from arXiv, as assigned by the uploader, on the left, to the topics identified by our incremental top2vec algorithm version on the right. Generally the documents from a manual category are assigned to the same topic, with only a few documents being assigned to others. Top2vec generally tends to identify fewer categories than arx-iv provides, and is rather successful in categorizing similar categories together. In the figure, the blue cluster 1 categorizes math related topics, the green cluster 2 summarizes pyhsics related topics, the red cluster 3 primarily contains astro physics, and the final yellow cluster four contains primarily high energy and nuclear physics.

6. Conclusion and Outlook

We show an approach to extend the topic modelling algorithm top2vec to enable tracing of topic trends by extending the algorithm with a sliding window approach to

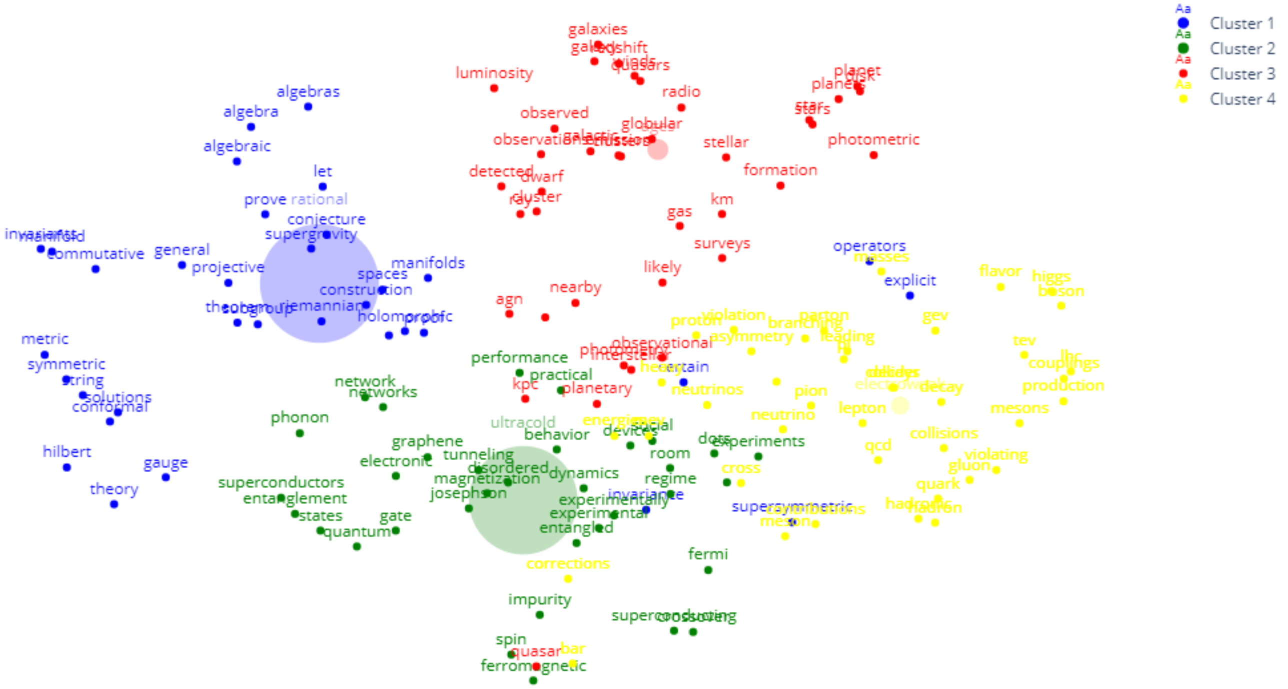

Figure 5. Key word embedding of words over all given clusters

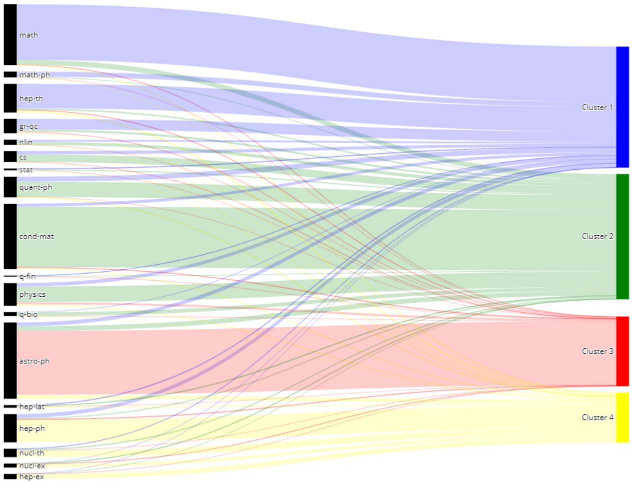

Figure 6. Overlap between different words in multiple clusters in the form of a sankey diagram.

cluster the embedded documents.

Based on this method, we used visualization techniques to create a decision support system for analyzing arXiv documents to show trends in research topics over time, and also enabled analysis of the evolution of a topic over time.

In general, the presented approach can be used for any given domain, and is possibly extensible to other topic modelling algorithms, as they all follow a similar logic (Egger and Yu, 2022). This could be used for example, to not only model shifitng trends in science, but analyze large bodies of text in general. For example shifting topics in newspapers over time, or the evolution of social media platforms could be traced this way. An interesting topic for future research would be to analyze the evolution of fake news which has been on the rise lately Kalsnes (2018).

The primary caveat of our approach is its run-time as it requires repeated clustering, and topic / document matching as opposed to the base topic modelling approaches. Steps were taken to reduce this by allowing adding data without re-clustering.

In the future we hope to improve our approach by using performance metrics, such as citation count of publications, to calculate the rise and fall of trends as opposed to simply using the amount of publications over time. In addition, the approach could benefit from an additional matching when topics split into multiple topics, or when topic domains begin to merge over time.

7. Funding

Our thanks to the Austrian Research Promotion Agency FFG for facilitating the project *Predictive Innovation* in the General Programme funding program, with research budget provided by the Federal Republic of Austria.

8. Acknowledgements

The authors thank their project partner mobile agreements GmbH, for the close cooperation and their efforts.

References

Angelov, D. (2020). Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Campello, R. J. G. B., Moulavi, D., and Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In Pei, J., Tseng, V. S., Cao, L., Motoda, H., and Xu, G., editors, *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Chen, J., Wei, W., Guo, C., Tang, L., and Sun, L. (2017). Textual analysis and visualization of research trends in data mining for electronic health records. *Health Policy and Technology*, 6(4):389–400.

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

Di Corso, E., Proto, S., Cerquitelli, T., and Chiusano, S. (2019). Towards automated visualisation of scientific literature. In *European Conference on Advances in Databases and Information Systems*, pages 28–36. Springer.

Egger, R. and Yu, J. (2022). A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in Sociology*, 7.

Grootendorst, M. (2020). Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics.

Hassani, M. (2015). *Efficient clustering of big data streams*. Apprimus Wissenschaftsverlag.

Hassani, M. (2019). *Overview of Efficient Clustering Methods for High-Dimensional Big Data Streams*, pages 25–42. Springer International Publishing, Cham.

He, J., Li, L., and Wu, X. (2017). A self-adaptive sliding window based topic model for non-uniform texts. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 147–156.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, page 50–57, New York, NY, USA. Association for Computing Machinery.

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.*, 5:1457–1469.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. (2019). Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211.

Joachims, T. (1996). A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. *Carnegie-*

mellon university of pittsburgh.

Kalsnes, B. (2018). Fake news. In *Oxford Research Encyclopedia of Communication*.

Krstić, Ž., Seljan, S., and Zoroja, J. (2019). Visualization of big data text analytics in financial industry: A case study of topic extraction for italian banks. In *Proceedings of the ENTRENOVA-ENTerprise REsearch InNOVAtion Conference*, volume 5, pages 35–43.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Liu, S. and Jansson, P. (2017). City event detection from social media with neural embeddings and topic model visualization. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 4111–4116. IEEE.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mühlroth, C. (2020). Artificial intelligence as innovation accelerator. In *Proceedings of the 2020 on Computers and People Research Conference*, SIGMIS-CPR'20, page 6–7, New York, NY, USA. Association for Computing Machinery.

Nazemi, K., Burkhardt, D., and Kock, A. (2022). Visual analytics for technology and innovation management. *Multimed. Tools Appl.*, 81(11):14803–14830.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Sievert, C. and Shirley, K. (2014). Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70.

Stöckl, A., Diephuis, J., and Aschauer, A. (2020). Instavis: Visualizing clusters of instagram message feeds. In *2020 24th International Conference Information Visualisation (IV)*, pages 435–439. IEEE.

Zou, C. and Hou, D. (2014). Lda analyzer: A tool for exploring topic models. In *2014 IEEE International Conference on Software Maintenance and Evolution*, pages 593–596. IEEE.