Dynamic Topic Modeling of Video and Audio Contributions

A. Stöckl[1]

[1]Department Digital Media, University of Applied Sciences, Upper Austria

DOI 10.3217/978-3-85125-976-6-18

Abstract. This paper shows how topics and their temporal evolution in audio and video broadcasts can be analyzed and visualized automated. For this purpose, Deep Learning systems such as "OpenAI Whisper" and "GPT3" are used to transcribe the audio data and extract the essential content per broadcast. The "BERTopic" method (Grootendorst, M. 2022) is used for dynamic topic modeling. The result is clusters of content ("topics") that are described and visualized using scatter plots, word clouds, and line charts. The method solves problems of topic modeling and enables the automated analysis of large amounts of data. A software prototype was developed that combines the sub-models and enables the analysis. The method is demonstrated using the example of Austrian TV channel ServusTV's weekly commentary "Der Wegscheider" over a period of more than four years (2018-2022). It is shown that migration, "mainstream media," and the Covid-19 pandemic are dominant topics. The time trend analysis illustrates how the COVID-19 pandemic increasingly crowded out the other topics from mid-2019. This method demonstrates how AI can be applied to journalistic work to enable the analysis and visualization of large data sets.

1 Introduction

In our digitally connected world, vast information is expanding exponentially (Settembre, M. 2012). From social media platforms to online publications, blogs, and academic literature, textual data has become integral to our daily lives. However, making sense of this massive unstructured text has proven daunting (Masood et al., A. 2019). How can we efficiently analyze and extract meaningful information from this abundance of words?

Enter topic modeling – a powerful computational technique that enables us to discover hidden patterns, uncover thematic structures, and extract valuable insights from unstructured textual data. Topic modeling (Albalawi et al., 2020) has emerged as a transformative field within natural language processing (NLP) and machine learning, providing researchers, businesses, and organizations with a remarkable toolkit to make sense of vast text collections.

At its core, topic modeling is a statistical and machine-learning approach that aims to identify topics or themes within a given corpus of documents. Using sophisticated algorithms, topic modeling techniques can automatically group similar documents, identify key themes, and even uncover latent patterns that may not be immediately apparent to human readers. These extracted topics act as clusters of related words, which encapsulate the central ideas and concepts present within the text.

Topic modeling applications (Boyd-Graber et al., D. 2017) are extensive and span various industries and domains. In academia, researchers employ topic modeling to explore large document collections, identify research trends, and gain a comprehensive understanding of a specific research area. In business and marketing, topic modeling helps uncover customer sentiment, detect emerging trends, and analyze online reviews or social media conversations. Governments and policy-makers can leverage topic modeling to monitor public opinion (Ma, B. et al. 2016), analyze public debates, and make data-driven decisions.

This article shows how topics and their development can be automatically analyzed and visualized in audio and video contributions. These can be YouTube channels, podcasts, or recurring TV shows. Suppose the volume of posts is so large and spread over extended periods that it would only be possible to consume or analyze them with much effort. In that case, the automated solution presented can help to achieve results.

The programs usually address different topics, which also change over time. If you follow them for a more extended period, you have a particular impression that topics keep coming up, but often, this appearance is deceptive and does not match the facts. It is desirable to automatically process and analyze the content to make objective statements, often available in large volumes. Subsequently, derivations can be formed and presented by suitable information visualizations.

In this paper, we would like to show a procedure that enables a data journalist to effectively process this large amount of audio/video material and draw conclusions from it with a fully automated approach. In contrast to earlier work, we use current technologies to make the process as automated as possible.

2 Related Work

In this section, on the one hand, we give a brief overview of research works dealing with the analysis of video content for topic extraction. On the other hand, we refer to different ways of identifying topics in text documents available after the transcription step.

For information extraction for data journalists, different approaches are available that use Natural Language Processing (Wiedemann et al., 2018) or analyze multimedia data (Salvador et al., 2017) to enable partial automation of the analysis.

In Stappen, Baird, Cambria et al. (2021), a lexical knowledge-based extraction method based on "SenticNet" (Cambria, E. et al. 2010) was used to analyze the video transcriptions of the "MuSe-CAR" dataset (Stappen et al., 2021). The "MuSe-CAR" dataset is a large multimodal dataset that contains video, audio, and text. It was developed specifically for researching multimodal sentiment analysis to understand it better. A concrete application example for multimodal sentiment analysis is emotional engagement in product reviews, such as automobile reviews. Here, a specific sentiment is linked to a particular topic or entity. Analyzing the video transcripts of the "MuSe-CAR" dataset, meaningful information can be obtained about the emotional engagement and sentimental connection in such reviews.

The method of lexical knowledge-based extraction is based on "SenticNet," (Cambria, E. et al. 2010) a knowledge resource that provides an extensive collection of terms, emotions, and concepts with corresponding sentiments. By matching the video transcripts with “SenticNet,” the terms they contain and their sentimental meaning can be extracted. This allows for a more in-depth analysis of emotional engagement and sentimental connections in the ratings of the "MuSe-CAR" dataset.

In Raaijmakers, den Hartog, and Baan, J. (2002), a model for multimodal topic segmentation and classification was developed using Dutch news videos as an example. The focus was on investigating the interaction between three modalities - visual, auditory, and textual information - within an integrated model for video analysis. The presented model is based on a fully automated sequential approach, where linguistic analysis and visual information are equally used for segmentation and classification. By combining linguistic analysis and visual information, the model allows for a more comprehensive and accurate analysis of news videos. It allows the identification and categorization of different topics and content in the videos by considering both the linguistic elements and the visual features.

In Zhu, Shyu, and Wang, H. (2013), a content-based video recommendation system called Video Topic is proposed, which utilizes a topic model in the context of recommendation systems. Its objective is to capture user interest in videos by employing a topic model to represent them and then generating recommendations by finding the videos that best align with the topic distribution of user interests. The Video Topic system introduces a novel approach to address the challenges of video recommendation by leveraging videos' inherent structure and content. Unlike traditional collaborative filtering techniques relying on user-item interactions, Video Topic focuses on the content. Using a topic model, the system can extract latent topics from videos, enabling a more profound

understanding of their content. Video Topic presents an innovative content-based recommendation system that employs a topic model to capture user interest in videos. By utilizing the thematic composition of videos and matching it with user preferences, the system generates personalized recommendations, aiming to enhance the relevance and satisfaction of users in their video consumption. The method is demonstrated and validated on the Movie Lens dataset[45] (Harper and Konstan, 2015).

For topic modeling of text content after transcription, several approaches are available, of which the "LDA method" (Blei, Ng, and Jordan 2003) can be considered the classical method, which, together with the visualization from Sievert and Shirley (2014), set the standard for a long time. The main limitation of the method is that it is based on a "Bag of Words" representation of the text contents. These consider the frequency of words in a document without regard to the order in which the words appear and thus cannot adequately capture the meaning of the documents.

With the advent of "word embeddings" such as "Word2Vec" (Mikolov et al., 2013) and "Glove" (Pennington et al. 2014), a more potent form of representing words in the form of high-dimensional vectors was available. Further developments also brought representations of entire text documents (Le and Mikolov 2014). These form a starting point for recent methods for topic modeling, such as "Top2Vec" (Angelov, 2020) and "BERTopic" (Grootendorst 2022), which are based on the representations of Devlin et al. (2018). We will use BERTopic in our method, representing state-of-the-art technology for topic modelling.

The detection of trends by Topic Modeling with "Top2Vec" over time was analyzed in Krauss, Aschauer, and Stöckl, A. (2022) using the Top2Vec method, and appropriate visualizations were given. Hall, Jurafsky, and Manning (2008) use topic modeling to analyze the history of ideas in computational linguistics, each based on text documents. An overview of different topic modeling techniques is given by Vayansky and Kumar (2020).

4 The Method and the Data

To demonstrate the method, it was necessary to find a suitable use case that is typical for the task and for which we have the data. We chose the series "Der Wegscheider"[46] of the Austrian TV channel "ServusTV", which provides weekly commentaries on topics from politics and society.

[45] https://grouplens.org/datasets/movielens/

[46] https://www.servustv.com/aktuelles/b/der-wegscheider/aa-1q66uk71n1w11/

ServusTV played a controversial role during the COVID-19 pandemic. The channel was criticized for disseminating partly false information about the pandemic[47]. Some of this information questioned the virus's dangerousness and the vaccines' effectiveness. For example, Corona down player Sucharit Bhakdi was regularly invited as a guest[48], and station chief Ferdinand Wegscheider disseminated questionable information in his weekly commentaries, such as claims that the Corona vaccine was a poorly tested, genetically modified substance or that the deworming agent ivermectin was successfully used against Covid-19[49].

ServusTV's role was particularly highlighted in the anti-Corona measures demonstrations in Vienna. Many demonstrators who protested against the measures and considered the vaccines dangerous obtained their information from ServusTV or the Telegram platform.

These practices led to controversy and criticism. For example, the Viennese press club Concordia filed a complaint with the media authority KommAustria[50], arguing that the station was not fulfilling its duty of journalistic care. It should be noted. However, the station boss, Wegscheider, defines his show as commentary and personal opinion, not reporting.

"Der Wegscheider" has been broadcast since 2018 at prime time on Saturdays at 19:30, and also reaches many viewers and listeners via social media and the podcast, making a not inconsiderable contribution to shaping opinion in Austria. To demonstrate our method, the series is well suited because this period of about four years has produced an immense amount of audio/video material, which is no longer easy to overview without technical tools.

In the following, we show, based on the presented method, which topics the broadcaster has covered and how these have changed over time. We needed the transcribed texts of the podcast versions available across different platforms to perform the analysis. In the analyzed period from November 3, 2018, to December 3, 2022, 161 broadcasts were available to us. The length of each broadcast varied between 6 and 8 minutes, resulting in about 19 hours of material.

[47] https://www.tagesschau.de/ausland/europa/servus-tv-corona-101.html

[48] https://www.servustv.com/aktuelles/a/corona-nur-fehlalarm-talk-spezial-mit-prof-dr-sucharit-bhakdi/118407/

[49] https://www.rnd.de/medien/servus-tv-verschwoerungserzaehlungen-und-tierdokus-WNRQT64KDND7VNPQ7XLWQ4CYYI.html

[50] https://www.rtr.at/medien/wer_wir_sind/KommAustria/KommAustria.de.html

To perform a topic analysis, the content must be in text form. Manual transcription would be both time-consuming and tedious for humans. Therefore, we need a mechanical method to create the transcriptions. This also allows us to scale to even larger data sets than the example shown.

A machine learning process for transcribing audio content can be implemented using speech recognition technologies. Such technologies use complex algorithms to convert spoken language into written text. Several commercial vendors and open-source solutions based on machine learning enable automated transcription.

When using a machine transcription process, it is essential to note that the accuracy of the result depends on several factors. The quality of the audio recording, the speakers' voice quality, background noise, and the transcription service used can all affect accuracy. The manual post-processing of the transcripts may be required to correct errors and improve quality.

After the transcriptions have been created, the texts are ready for further analysis. Natural Language Processing (NLP) methods can perform various analyses, such as extracting keywords, classifying topics, or detecting moods and emotions. These analyses can provide valuable insights into the podcast content and contribute to developing insights.

A recent paper (Radford et al., 2022) presented a new method, including open-source software, that produces a meager error rate comparable to human transcription. The model "OpenAI Whisper" is a robust multilingual speech recognizer consisting of a neural network according to the Transformer architecture (Vaswani et al., 2017) trained on 680,000 hours of audio. We use the "large" version, which promises the most accurate results but requires powerful hardware with GPU support to obtain usable transcription times.

This provides us with 161 text files with an average length of 7002 characters (maximum, 8466 and minimum, 5298). Together with the broadcast date, these form the basis for dynamic topic modeling over time.

5 Analysis and Results

Another problem in topic modeling is the correct setting of the hyperparameters that influence the model's performance. The choice of the number of topics is of particular importance. Too few topics can lead to missing essential topics, while too many can lead to overlap and ambiguity.

Various evaluation metrics can be used to improve the quality of Topics, such as the coherence score. This score evaluates the coherence of the words within a topic and allows the selection of the topics with the highest coherence values.

Another challenge is to consider the context and semantics of words. Often, words can have different meanings depending on the context in which they are used. Here, contextual word vectors, such as those used in BERT (Devlin et al. 2018), can help better capture the meaning of words.

Furthermore, the visualization of the topics is essential to make the results interpretable and understandable. In addition to the information visualizations already mentioned, word clouds or heat maps can show the most important words or the distribution of words in the topics.

Overall, topic modeling is a complex process that requires careful data preparation and an accurate selection of methods. It is essential to consider the specifics of the use case and make iterative improvements to obtain meaningful and interpretable Topics.

Here, we use the large language models that have recently become fashionable to extract the most important keywords from the individual transcripts, allowing us to obtain more interpretable topics. Current representatives of this genre are "GPT-3" (Brown et al., 2020), "Palm" (Chowdhery et al. 2022), and "Lambda" (Thoppilan et al., 2022). The training of these generative language models is that they are optimized for predicting the next word in an existing text corpus. We use the GPT3 language model with 175 billion parameters ("davinci-003") for our analysis[51], publicly available via a paid API. It has the advantage over the other representatives of being available in a better-optimized version by being enhanced with additional steps consisting of supervised learning and reinforcement learning (Ouyang et al., 2022).

This then provides a list of key terms for a transcript, such as in the example below:

Wegscheider, Delta-Mutation, Impfkampagne, Kündigungen, Abschiebestopp, Dolm der Woche, Nationales Impfgremium, Hans Maher, Zwangsmaßnahme, 3G-Regel, Testpflicht, Wiener Gemeinde-Wohnung, Afghane, Vergewaltigung, Ermordung, ORF-Reporter, Bundesregierung, Virologen, WHO, Birgit Hebein, Asylaktivisten, Abschiebung, SPÖ-Parteitag, Pamela Rendi-Wagner, Alma Zadic, Grünen

Translated to English:

Wegscheider, Delta Mutation, Vaccination Campaign, Terminations, Deportation Stop, Dolm of the Week, National Vaccination Panel, Hans Maher, Coercive Measure, 3G Rule, Mandatory Testing, Vienna Municipal Apartment, Afghan, Rape, Murder, ORF Reporter,

[51] https://beta.openai.com/docs/models/gpt-3

Federal Government, Virologists, WHO, Birgit Hebein, Asylum Activists, Deportation, SPÖ Party Conference, Pamela Rendi-Wagner, Alma Zadic, Greens

This representation is then subjected to a cleaning process that normalizes all words to lower case and removes numbers and special characters. This gives us short descriptions per transcript, as in the example:

delta mutation impfkampagne kündigungen abschiebelstopp dolm woche nationales impfgremium hans maher zwangsmaßnahme 3g regel testpflicht wiener gemeinde wohnung afghane vergewaltigung ermordung orf reporter bundesregierung virologen who selbsttests ferien gewalttäter asylwerber autobahnblockade birgit hebein asylaktivisten

Translated to English:

delta mutation vaccination campaign announcements deportation stop dolm week national vaccination board hans maher coercive measure 3g rule mandatory testing vienna community apartment afghans rape murder orf reporter federal government virologists who self-tests vacations violence perpetrators asylum seekers highway blockade birgit hebein asylum activists

With the 161 short descriptions, we now start topic modeling with BERTopic, which yields three topics, shown in **Fig 1**. For the 2-dimensional representation, the document vectors were dimensionally reduced using the UMAP method (McInnes, Healy and Melville 2018).

UMAP is a dimensionality reduction technique commonly used for visualizing high-dimensional data in lower-dimensional spaces while preserving the underlying structure. It is beneficial for visualizing word vectors, typically represented in high-dimensional spaces.

There are three topics named with the essential terms per topic.

0 - corona_impfpflicht_maskenpflicht - corona_vaccination_mask_obligation

1 - spö_medien_mainstream - spö_media_mainstream

2 - flüchtlinge_türkis_türkei - refugees_turquoise_turkey

The entire weekly comments of the four years can be traced back to this very narrow range of topics. In addition to the COVID-19 issue and the measures, the refugee issue and the mainstream media are dominant.

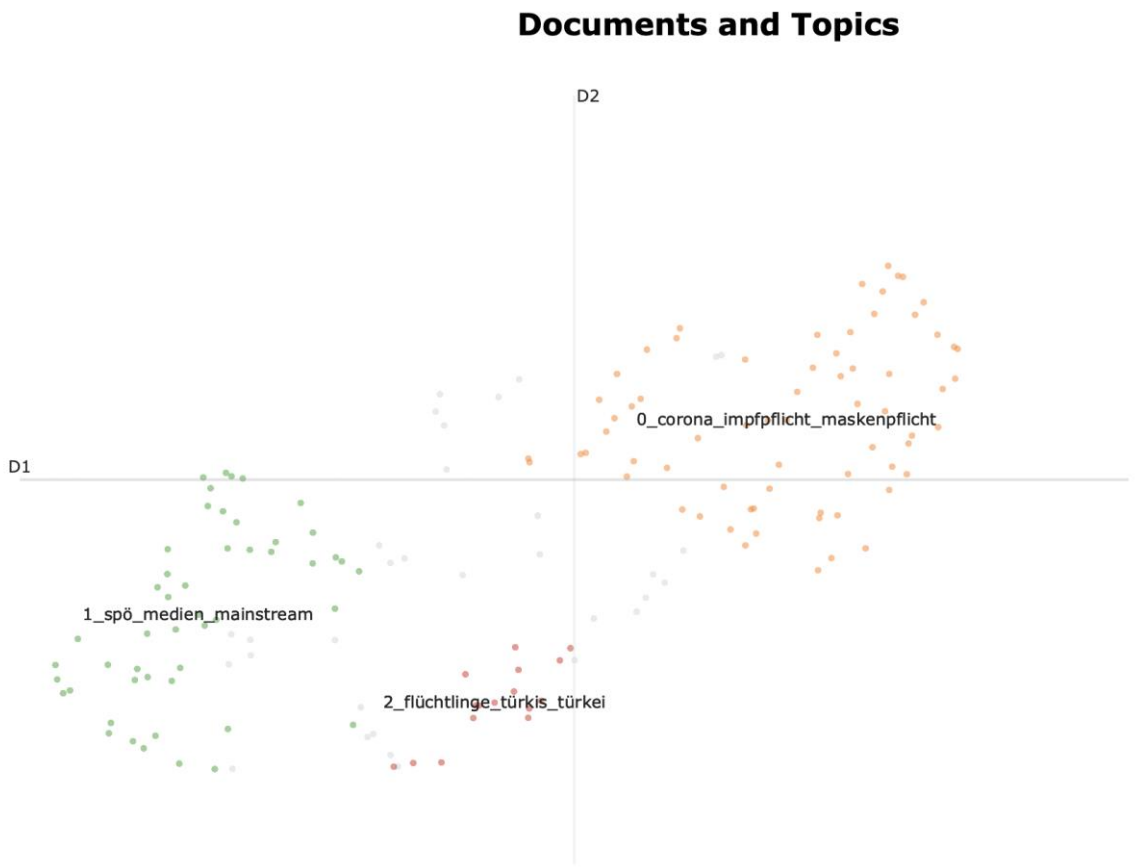

Figure 1: 2D display of transcripts and topic designation.

Word clouds, also known as tag clouds or text clouds, are visual representations of text data where words are displayed in different sizes, with more prominent or more extensive words indicating their higher frequency or importance within the text. They provide a quick and intuitive way to analyze and visualize textual information. In **Fig. 2**, we represent each topic as a word cloud, where the topic's importance weights the words' size. This makes it easier to identify the contents of the individual topics.

Topic 0 deals with compulsory vaccination, lockdown, and mandatory masking, which Wegscheider describes as coercive measures. From Wegscheider's point of view, these are driven by the pharmaceutical industry and mainstream media and represent violations of fundamental rights.

Topic 1 is devoted to the so-called mainstream media, such as the Austrian Broadcasting Corporation (ORF), which is always harshly criticized by Wegscheider and is accused of being close to political parties (SPÖ, FPÖ) and individuals (Kurz, Rendi-Wagner).

Figure 2: Word cloud for the three topics of “Der Wegscheider”.

Topic 2 deals with the refugee problem and the topic of asylum together with the political parties ÖVP, associated with the color turquoise, and the Green Party. Wegscheider is a representative of a restrictive policy here.

Next, we examine the temporal course of the topics. How often are the topics represented in each week's comments?

Fig. 3 shows the development over time. It is clear how the COVID-19 topic emerges and pushes the other two topics back.

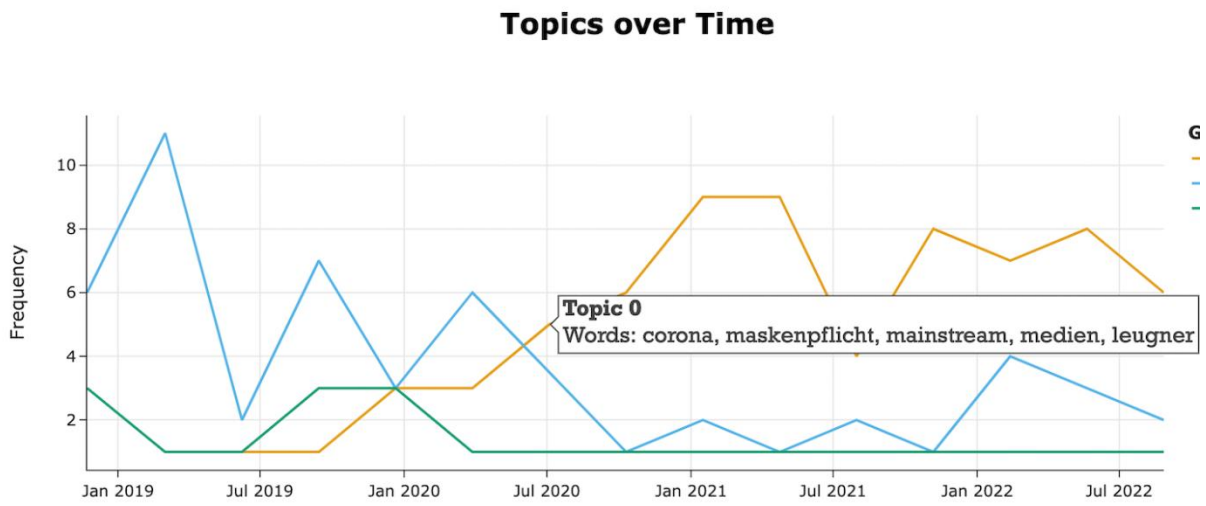

Figure 3: Line plot showing the development of the topics appearing in the broadcasts over time.

6 Summary and Outlook

In our example of the weekly commentary "Der Weigscheider," we have shown how a series of audio or video contributions can be subjected to automated topical analysis. In doing so, we used several powerful tools, including the transcription model "OpenAI Whisper," information summarization using "GPT3", and dynamic topic analysis using "BERTopic." The results were also processed using information visualization to enable better interpretation.

In the context of our concrete use case, we could determine the narrow thematic orientation of the "Der Wegscheider" and which topics predominate with him. In particular, it was interesting to observe how the topic of COVID-19 enormously gained importance over time and ultimately dominated.

This methodology can be applied to audio and video sources, especially podcasts and YouTube channels. This makes it possible to track the development of topics in an automated way. The approach is essential for social and political content but can also be used in different areas, such as technological developments. Automated monitoring of topic development makes it possible to identify trends early and perform a well-founded analysis.

References

Albalawi, R., Yeap, T.H., & Benyoucef, M. (2020). 'Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis.' Frontiers in Artificial Intelligence, 3.

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', Journal of machine Learning research, 3(Jan), pp. 993–1022.

Boyd-Graber, J.L., Hu, Y., & Mimno, D. (2017). 'Applications of Topic Models.' Found. Trends Inf. Retr., 11, 143-296.

Brown, T. et al. (2020) 'Language models are few-shot learners', Advances in neural information processing systems, 33, pp. 1877–1901.

Cambria, E. et al. (2010). SenticNet: A Publicly Available Semantic Resource for Opinion Mining. AAAI Fall Symposium: Commonsense Knowledge.

Chowdhery, A. et al. (2022) 'Palm: Scaling language modeling with pathways', arXiv:2204. 02311.

Devlin, J. et al. (2018) 'Bert: Pre-training of deep bidirectional transformers for language understanding', arXiv preprint arXiv:1810. 04805.

Grootendorst, M. (2022) 'BERTopic: Neural topic modeling with a class-based TF-IDF procedure', arXiv preprint arXiv:2203. 05794.

Hall, D., Jurafsky, D. and Manning, C. D. (2008) 'Studying the history of ideas using topic models', in Proceedings of the 2008 conference on empirical methods in natural language processing, pp. 363–371.

Krauss, O., Aschauer, A. and Stöckl, A. (2022) 'Modelling shifting trends over time via topic analysis of text documents', Proceedings of the 21st International Conference on Modelling and Applied Simulation MAS 2022.

Le, Q. and Mikolov, T. (2014) ‘Distributed representations of sentences and documents’, in International conference on machine learning. PMLR, pp. 1188–1196.

Ma, B. et al. (2016) ‘Public Opinion Analysis based on Probabilistic Topic Modeling and Deep Learning’. Pacific Asia Conference on Information Systems.

Masood, A., & Hashmi, A. (2019). ‘Text Analytics: The Dark Data Frontier.’ Cognitive Computing Recipes.

McInnes, L., Healy, J. and Melville, J. (2018) ‘Umap: Uniform manifold approximation and projection for dimension reduction’, arXiv preprint arXiv:1802. 03426.

Mikolov, T. et al. (2013) ‘Efficient estimation of word representations in vector space’, arXiv preprint arXiv:1301. 3781.

Ouyang, L. et al. (2022) ‘Training language models to follow instructions with human feedback’, arXiv preprint arXiv:2203. 02155.

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), pp. 1532–1543.

Raaijmakers, S., den Hartog, J. and Baan, J. (2002) ‘Multimodal topic segmentation and classification of news video’, Proceedings. IEEE International Conference on Multimedia and Expo, 2, pp. 33–36 vol.2.

Radford, A. et al. (2022) ‘Robust Speech Recognition via Large-Scale Weak Supervision’. Available at: https://cdn.openai.com/papers/whisper.pdf.

Salvador, J., Ruiz, Z., & Rodríguez, J.G. (2017). ‘A Review of Infrastructures to Process Big Multimedia Data.’ Int. J. Comput. Vis. Image Process., 7, 54-64.

Settembre, M. (2012). ‘Towards a hyper-connected world.’ 2012 15th International Telecommunications Network Strategy and Planning Symposium (NETWORKS), 1-5.

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, pp. 63–70.

Stappen, L., Baird, A., Cambria, E., et al. (2021) ‘Sentiment Analysis and Topic Recognition in Video Transcriptions’, IEEE Intelligent Systems, 36, pp. 88–95.

Stappen, L., Baird, A., Schumann, L., et al. (2021) ‘The Multimodal Sentiment Analysis in Car Reviews (MuSe-CaR) Dataset: Collection, Insights and Improvements’, arXiv [cs.MM]. Available at: http://arxiv.org/abs/2101.06053.

Thoppilan, R. et al. (2022) ‘Lamda: Language models for dialog applications’, arXiv:2201. 08239.

Vaswani, A. et al. (2017) ‘Attention is all you need’, Advances in neural information processing systems, 30.

Vayansky, I. and Kumar, S. A. P. (2020) ‘A review of topic modeling methods’, Information Systems. Elsevier, 94, p. 101582.

Wiedemann, G., Yimam, S.M., & Biemann, C. (2018). ‘A Multilingual Information Extraction Pipeline for Investigative Journalism.’ ArXiv, abs/1809.00221.

Zhu, Q., Shyu, M.-L. and Wang, H. (2013) ‘VideoTopic: Content-Based Video Recommendation Using a Topic Model’, 2013 IEEE International Symposium on Multimedia, pp. 219–222.