Universidad de San Andrés
Mathematics and Sciences Department
MSc. in Data Science

A FEW SECTIONS OF
Immigration in Films
A Machine Learning Analysis from Subtitles
(FULL VERSION PUBLISHED IN SPANISH)

Wendy BRAU

Advisor: Marcela Svarc

Ciudad Autónoma de Buenos Aires
2024

Abstract

This thesis analyses the immigration content of films based on text data from a corpus of more than 27.000 subtitles. First, I use a combination of Fast K-Medoids, Random Forest and hierarchical clustering algorithms to obtain stable and meaningful topics that are systematically associated to immigration content in films. Secondly, I train supervised models to construct a continuous index that measures the importance of immigration content in each film. Finally, I explore the association between consumption of films with immigration content in theaters and real immigration dynamics.

Introduction

Films both reflect and influence social dynamics of their context, such as violence, gender roles, and the representation of ethnic minorities [12, 18, 19, 46]. While some studies have analyzed the treatment of immigration topics in other corpora [3, 30, 48, 53], and others the content of films based on subtitles [10, 13, 18, 26, 28, 41, 42, 48, 63], this analysis is situated at the intersection between them, studying the relationship between films and immigration. For so doing, it measures and characterizes the content of immigration in films from a corpus of more than 27,000 subtitles, and explores its relationship with actual immigration dynamics.

This thesis has three different goals. The main one is to assess the topics through which films address immigration themes. The second one is to measure the amount of immigration content in each film. The third one, to explore the relationship between consumption of films with more immigration content and actual immigration dynamics.

As a starting point, a film will be considered to be an 'immigration film' if IMDb labels it using the keywords 'immigration', 'immigrant', 'migration', 'migrant'. Then, to tackle the first goal, I analyse whether there are topics that are systematically associated to this type of films. By 'topic' I will refer to a subset of vocabulary with a shared semantic. The core steps to identify immigration topics are: (i) clustering the unique lemmas present in the corpus of subtitles using Fast K-Medoids, based on the cosine distance between the embeddings of the lemmas from GLoVe; (ii) using the term frequency - inverse document frequency ($TFIDF$) matrix to assign a value to each film in each cluster of lemmas; (iii) getting the clusters that best predict that a film is an immigration film according to a Random Forest model. I repeat these steps using different random seeds to obtain stable and robust immigration topics. Then, I regroup the clusters obtained in different rounds via hierarchical clustering. After the topics are defined, I assign a value to each film in each topic.

Secondly, to measure the importance of immigration content in films, I train supervised models to predict the binary immigration label from IMDb. The predicted probabilities that a film is an immigration film constitute an Index of Immigration Content. This index is a continuous measure, which has more information than a binary label, since immigration can be a main or a secondary topic in a film. I explore two alternative methods to construct it. On the one hand, training supervised models (Naive Bayes, Discriminant Analysis, Logistic Regression, Random Forest Regression, K-Nearest Neighbors) using the internal product of the $TFIDF$ matrix and the embedding matrix as features. On the other hand, fine-tuning RoBERTa, given that transformers usually have the best performance in specific tasks of interest like classification.

Finally, I define a measure of consumption of immigration content in theaters by combining the values of each film in each immigration topic and the Index of Immigration Content with box office data in different countries and years. I explore its association with actual immigration dynamics in different geographies and periods.

Next, Chapter 2 describes the contribution of this thesis to previous studies. Chapter 3 describes the data and methodology. Chapter 4 presents results for each analysis. Chapter 5 discusses the limitations, alternative methodologies and improvements, as well as research ideas for the future. Conclusions at the end.

Literature

This thesis is part of the contributions of Machine Learning and Natural Language Processing (NLP) to the study of social and cultural phenomena such as immigration. A wide literature has analysed public attitudes towards immigration, but most studies did so based on survey data [7, 24, 55]. The availability of lots of text corpora from cultural products, such as film subtitles, allow to study the representation of immigrants from these new data sources by using NLP.

Specifically, this thesis is inspired by studies that have analysed the interaction between consumption of films –or other audiovisual products– and other social phenomena. Studying their relationship is relevant because cultural products not only reflect but also influence the acquisition of social values and accepted behavior. Therefore, their content, biases and stereotypes can favor certain values over others [12, 18, 19, 46]. These studies can be grouped in two:

1. Those that measure the effects of consumption of certain audiovisual content on social behavior. For example, there are studies that found that violent crime decreases in days with more audience in violent movies, which can be explained by a substitution effect, if violent people self-select to watch violent movies; that families that live in areas close to an Al Gore documentary exhibition buy more carbon offsets in the subsequent months; or that people that watch the educational TV show MTV Shuga improve their attitudes towards domestic violence, HIV and risky sexual behavior [5, 6, 15, 31]. All these studies use an ex-ante definition of the set of films or shows referring to the phenomenon of interest.

2. Those that analyse how films capture social phenomena, that is, they quantify biases and stereotypes in films. Some findings are that women and ethnic minorities have less time on screen and speech; women are less associated to intelligence-related vocabulary than men; there are patterns of association between characters' demographics, their relationships and their dialogues (e.g., films with Latin characters have more sexual lines, older characters speak more about achievements or religion); and it is possible to detect hate-speech in films [18, 20, 40, 46, 63]. These studies attempt to describe and measure different types of content in films, many times using NLP based on the script or the subtitles.

This study belongs to the second group. Based on a higher number of film subtitles, it analyses a topic less explored so far using this corpus: immigration. The topic has been analysed in other corpora. For instance, in tweets from US congressmen, to classify their immigration stance; in administrative databases, to identify immigration-oriented ONGs; in British press articles, to describe how they talk about immigration [3, 30, 48, 53]. There are also studies that constructed corpus of immigration-related texts, like the *Corpus*

Multilíngue sobre Migração e Refúgio (COMMIRE), that contains documents and other linguistic materials that migrants and the people that work with migrants use [17]. On the other side, text data related to films' content (synopsis, script or subtitles) have been widely used with other analysis purposes, mainly to predict the film genre and for developing recommendation algorithms [42]. Some of these use unsupervised techniques and topic modelling, while others use supervised methods, like recurrent neural networks or K-Nearest Neighbors (KNN) [10, 13, 26, 28, 41, 42].

I combine supervised and unsupervised methods with the goal of identifying topic associated to immigration content (not general topics). Among supervised techniques, I leverage Fast K-Medoids and hierarchical clustering –similar to [26], but based on 40 times more subtitles and on embeddings from pre-trained models (GloVe) instead of bag-of-words (BOW) embeddings. Among supervised methods, I explore KNN like [28], and finetuned transformers like [63], but with the goal of constructing an Index of Immigration Content.

The methodology for obtaining the immigration topics, unlike many recent studies that use pre-defined labels or pairwise annotation (i.e., indicating whether two documents belong to the same topic), does not require annotation [21, 64]. However, traditional unsupervised methods are usually based on probabilistic models such as Latent Dirichlet Allocation (LDA) that are based in BOW embeddings. This is a limitation, as those representation misses the information about the semantic relationship among the words in the same context. Therefore, I instead follow the most recent literature that uses word embedding models –such as GloVe– and then performs clustering methods based on those embeddings [1, 11, 23, 52, 59].

The evaluation of the quality of the proposed methodology for constructing the topics is based on (i) the topic stability and (i) the topic alignment with meaningful themes or semantic categories. That is, topics should have a clear interpretation. Traditional techniques usually have a better performance in these two aspects than newer ones, like neural topic models. To ensure stability, the methodology involves an ensemble and regrouping method from various rounds of estimations, which also usually renders better results [2].

Hypotheses

The films that should have a higher value in the Index of Immigration Content are those where the main plot is related to the immigration, and those where having an immigration background is a relevant feature of one of the main character. In turn, films with a secondary plot or character related to immigration should have positive though lower values in the index.

Some films that serve as archetypes of immigration films are Gangs of New York (2002), Brooklyn (2015), Tori and Lokita (2022) or The Swimmers (2022). Based on them, I expect to find different immigration topics referring to the origin and destination countries of the characters and the main conflict in the plot.

Following [43], throughout the years, the focus of immigration cinema has moved from urban gangs of Chinese, Irish or Italian origin that searched for wealth and power; to the optimism and hard-work ethics of immigrants to progress and integrate themselves; to the problems of poverty, prejudice, and discrimination that immigrant face; to nostalgia; to the illegal conditions; and finally to the intergenerational gaps between immigrants and their descendants. I expect to identify some of these topics from subtitles.

Regarding the relationship between real immigration dynamics and the consumption of immigration content in films, on the one hand, societies with more immigration may have more interest on watching their own stories or the stories of their new neighbors, resulting in a positive association. Although some studies argue that many new immigration films were released when immigration increased in Europe [4], this could show an increasing interest in immigration from film makers rather than the interest from the general public. On the other hand, the often negative depiction of immigrants in the media –that associates immigration to illegality or crime– might generate reluctance to watch films on the subject. Finally, given that public attitudes towards immigration depend on the sociodemographic profile of immigrants, the sign of the association might also vary among films that talk about immigration from different points of view, topics, or perspectives [7].

Future work (summary)

Some interesting questions to explore in future research are:

- Comparing the performance of simple versus complex models when annotation is incomplete or bad, such as in the case of the IMDb tags of immigration films.
- Recent studies explore the use of large language models for few-shot or zero-shot –that is, with few or no annotations– learning of topics or for evaluating topic models [22, 37, 49, 54]. In future work, it will be interesting to compare with their results for finding immigration topics in films.
- Clustering vocabulary based on the GloVe embeddings has noticeable biases, due to the data used to train the embeddings. To begin with, words like ‘trafficker’ or ‘undocumented’ fall in the same cluster than Latin names. In fact, previous studies have shown that embeddings have historical dynamics and biases that correspond to social changes [19]. Therefore, it would be interesting to train embedding models using subtitles from different historical moments.
- Comparing the immigration content in different film genres. In particular, children films. Comparing immigration content in films of different origins and languages.
- Analysing the immigration content evolution in different parts of the plot. For example, finetuning a RoBERTa model at a group-of-lines levels to visualise the predicted probability of immigration content at different times of the films, on average. Comparing the plot of immigration versus non-immigration films using sentiment analysis as the film progresses, and visualize average narrative arcs in each type of films, like [47] do.
- Analysing the relationship between immigration content consumption in cinemas and emigration, instead of immigration, and by type of migrants (refugees versus non-refugees, immigrants of different origins).

Conclusions

Measuring the immigration content of films helps to understand public attitudes towards immigration over time –which both influence cultural products, and are influenced by them. This thesis explores whether there are topics systematically associated with immigration in films and which and how important those topics are, and it measures general immigration content in more than 27 thousand films, based on their subtitles, NLP, and both supervised and unsupervised ML methods.

Although subtitles cannot capture all the film content –such as the expressed in the images or body language–, they do have relevant information for distinguishing between immigration and non-immigration films, and for finding the topics through which films address immigration.

First of all, combining hierarchical clustering, Random Forest, and Fast K-Medoids based on the cosine distance matrix between the embeddings of more than 27 unique lemmas in the corpus of subtitles is a suitable methodology for finding meaningful, stable topics, systematically associated to immigration. Each topic has a clear meaning: there are geographical references (British & Irish, Europe, New York & US topics), historical references (Nazism, Middle East conflicts), or other immigration-related aspects, such as Language, Immigration Law, Economy & Employment, Religion, Ideology and Cosmovision. When choosing another random seed, or changing the number of initial clusters, the topics found are similar. Over time, there was a peak of films with Europe and Nazism content related in the 40s, there has been an increase of Latin and Technology content during the last years, and a decrease in the New York & US and Economy & Employment topics.

Secondly, both classic ML models and the finetuning of RoBERTa largely improve the ability to discern between immigration and non-immigration films with respect to *ad-hoc* decision rules, such as those based on the presence of certain predefined vocabulary. Moreover, an interpretable supervised ML method like Logistic Regression if useful for measuring the immigration content of films and create an index of immigration content of each film. Although there are many 'false positives' –it predicts many more immigration films than those identified as such by the IMDb tagging–, some of these films actually do contain high immigration content when doing a manual checking. This means that a simple ML model helps to summarize common information referred to immigration in subtitles that may improve an initial imperfect labeling.

Finally, the initial exploration of the association between the consumption of immigration content in cinemas and the actual dynamics of immigrant reception in different times and geographies yields mixed results, but a (provisory) negative association in most cases.

References

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

[2] Miserlis Hoyle, A. M., Goel, P., Sarkar, R. & Resnik, P. (2022). Are Neural Topic Models Broken?. *Findings of the Association for Computational Linguistics: EMNLP 2022*, 5321–5344, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[3] Allen, W. & Blinder, S. (2013). Migration in the News: Portrayals of Immigrants, Migrants, Asylum Seekers and Refugees in National British Newspapers, 2010-2012. *Migration Observatory.*

[4] Ballesteros, I. (2015). Immigration cinema in the New Europe. *Intellect.*

[5] Banerjee, A., La Ferrara, E., & Orozco, V. (2019). Entertainment, education, and attitudes toward domestic violence. *AEA Papers and Proceedings.* Vol. 109, pp. 133-137.

[6] Banerjee, A., La Ferrara, E., & Orozco-Olvera, V. H. (2019). The entertaining way to behavioral change: Fighting HIV with MTV. *National Bureau of Economic Research.* No. w26096.

[7] Bansak, K., Hainmueller, J., & Hangartner, D. (2023). Europeans' support for refugees of varying background is stable over time. *Nature*, 620(7975), 849-854.

[8] Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150.*

[9] Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606.*

[10] Bougiatiotis, K., & Giannakopoulos, T. (2016). Content representation and similarity of movies based on topic extraction from subtitles. *Proceedings of the 9th Hellenic Conference on Artificial Intelligence* (pp. 1-7).

[11] Byrne, C., Horak, D., Moilanen, K. & Mabona, A. (2022). Topic Modeling With Topological Data Analysis. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11514–11533, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[12] Cape, G. S. (2003). Addiction, stigma and movies. *Acta Psychiatrica Scandinavica*, 107(3), 163-169.

[13] Chao, B., & Sirmorya, A. (2016). Automated movie genre classification with LDA-based topic modeling. *International Journal of Computer Applications*, 145(13), 1–5.

[14] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357.

[15] Dahl, G., & DellaVigna, S. (2009). Does movie violence increase violent crime?. *The Quarterly Journal of Economics*, 124(2), 677-734.

[16] Erjavec, Tomaž; et al., 2021, Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1, *Slovenian language resource repository CLARIN.SI*, ISSN 2820-4042.

[17] Furtado, A. B. D., & Teixeira, E. D. (2022). Corpus Multilíngue sobre Migração e Refúgio (COMMIRE): planejamento, compilação e conteúdo, em linhas gerais. Texto Livre, 15, e36965.

[18] Gálvez, R. H., Tiffenberg, V., & Altszyler, E. (2019). Half a century of stereotyping associations between gender and intellectual ability in films. *Sex Roles*, 81(9-10), 643-654.

[19] Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16), E3635-E3644.

[20] Geena Davis Institute on Gender in Media. (2015). The reel truth: Women aren't seen or heard. An automated analysis of gender representation in popular films. Recovered from *seejane.org/research-informs-empowers/data/.*

[21] Goschenhofer, J., Ragupathy, P., Heumann, C., Bischl, B. & Aßenmacher, M. (2022). CC-Top: Constrained Clustering for Dynamic Topic Discovery. *Proceedings of the The First Workshop on Ever Evolving NLP (EvoNLP)*, 26–34, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[22] Gretz, S., Halfon, A., Shnayderman, I., Toledo-Ronen, O., Spector, A., Dankin, L., Katsis, Y., Arviv, O., Katz, Y. Slonim, N. & Dor, L. E. (2023). Zero-shot Topical Text Classification with LLMs-an Experimental Study. *Findings of the Association for Computational Linguistics: EMNLP 2023* (9647-9676).

[23] Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794.*

[24] Hainmueller, J., t Hopkins, D. J. (2014). Public attitudes toward immigration. *Annual Review of Political Science*, 17, 225-249.

[25] Haluszka E, Niclis C, Pareja Lora A, Aballar LR (en prensa). Application of Natural Language Processing for the recognition of obesity-related topics in the discourses of Argentine Twitter users. *Lodz Papers in Pragmatics.*

[26] Hasan, M. M., Dip, S. T., Kamruzzaman, T. M., Akter, S., & Salehin, I. (2021). Movie Subtitle Document Classification Using Unsupervised Machine Learning Approach. *2021 IEEE 6th International Conference on Computing, Communication and Automation (ICCCA)* 219-224. IEEE.

[27] Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* .

[28] Hesham, M., Hani, B., Fouad, N., & Amer, E. (2018). Smart trailer: Automatic generation of movie trailer using only subtitles. *2018 First International Workshop on Deep and Representation Learning (IWDRL)* 26-30. IEEE.

[29] Courtesy information by IMDb (https://www.imdb.com), used with permission.

[30] Islentyeva, A. (2020). *Corpus-based analysis of ideological bias: Migration in the British press.* Routledge.

[31] Jacobsen, G. D. (2011). The Al Gore effect: an inconvenient truth and voluntary carbon offsets. *Journal of Environmental Economics and Management*, 61(1), 67-78.

[32] Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013). The TenTen corpus family. *7th International Corpus Linguistics Conference CL* (pp. 125-127).

[33] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* New York: Springer.

[34] Jurafsky, D., & Martin, J. H. (2024). *Speech and language processing (draft of February 3, 2024).* Chapter 10: Transformers and Large Language Models.

[35] Jurafsky, D., & Martin, J. H. (2024). *Speech and language processing (draft of February 3, 2024).* Chapter 11: Fine-Tuning and Masked Language Models.

[36] Kaufman, L. & Rousseeuw, P. J. (1990). Partitioning Around Medoids (Program PAM). *Wiley Series in Probability and Statistics.* Hoboken, NJ, USA: John Wiley & Sons, Inc., 68–125.

[37] Kim, K., & Lee, Y. (2023). DRAFT: Dense Retrieval Augmented Few-shot Topic classifier Framework. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2278-2294.

[38] Lison, P. & Tiedemann, J. (2016) OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. *Proceedings of the 10th International Conference on Language Resources and Evaluation* (LREC-2016), 2016.

[39] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019). RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint* arXiv:1907.11692

[40] Malik, M., Hopp, F. R., & Weber, R. (2022). Representations of Racial Minorities in Popular Movies: A Content-Analytic Synergy of Computer Vision and Network Science. *Computational Communication Research*, 4(1).

[41] Mangolin, R. B., Pereira, R. M., Britto Jr, A. S., Silla Jr, C. N., Feltrim, V. D., Bertolini, D., & Costa, Y. M. (2022). A multimodal approach for multi-label movie genre classification. *Multimedia Tools and Applications*, 81(14), 19071-19096.

[42] Matthews, P., & Glitre, K. (2021). Genre analysis of movies using a topic model of plot summaries. *Journal of the Association for Information Science and Technology*, 72(12), 1511-1527.

[43] Mintz, S. Historical Context: Movies and Migration. *The Gilder Lehrman Institute of American History.* https://www.gilderlehrman.org/history-resources/teaching-resource/historical-context-movies-and-migration.

[44] Pappagari, R., Zelasko, P., Villalba, J., Carmiel, Y., & Dehak, N. (2019). Hierarchical transformers for long document classification. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 838-844). IEEE.

[45] Pennington, J., Socher, R. & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Empirical Methods in Natural Language Processing (EMNLP)* 1532-1543.

[46] Ramakrishna, A., Martínez, V. R., Malandrakis, N., Singla, K., & Narayanan, S. (2017). Linguistic analysis of differences in portrayal of movie characters. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics.* Vol. 1, pp. 1669-1678.

[47] Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., & Dodds, P. S. (2016). The emotional arcs of stories are dominated by six basic shapes. E*PJ Data Science*, 5(1), 1-12.

[48] Ren, C., & Bloemraad, I. (2022). New Methods and the Study of Vulnerable Groups: Using Machine Learning to Identify Immigrant-Oriented Nonprofit Organizations. *Socius*, 8.

[49] Sarkar, S., Feng, D.,y Santu, S. K. K. (2023). Zero-Shot Multi-Label Topic Inference with Sentence Encoders and LLMs. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing,* 16218-16233.

[50] Schubert, E., & Rousseeuw P.J. (2019). Faster k-Medoids clústering: Improving the PAM, CLARA, and CLARANS Algorithms. *12th International Conference on Similarity Search and Applications (SISAP 2019)*, 171-187.

[51] Schubert, E., & Rousseeuw P.J. (2021). Fast and Eager k-Medoids clustering: O(k) Runtime Improvement of the PAM, CLARA, and CLARANS Algorithms. *Information Systems* (101) 101804.

[52] Sia, S., Dalmia, A., & Mielke, S. J. (2020). Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! *arXiv preprint arXiv:2004.14914.*

[53] Siegel, A. A., Laitin, D., Lawrence, D., Weinstein, J., & Hainmueller, J. (2022). Tracking Legislators' Expressed Policy Agendas in Real Time.

[54] Stammbach, D., Zouhar, V., Hoyle, A., Sachan, M.,y Ash, E. (2023). Re-visiting Automated Topic Model Evaluation with Large Language Models. *arXiv preprint arXiv:2305.12152.*

[55] Steele, L. G., & Abdelaaty, L. (2019). Ethnic diversity and attitudes towards refugees. *Journal of ethnic and migration studies*, 45(11), 1833-1856.

[56] Suchomel, V. Better Web Corpora For Corpus Linguistics And NLP. (2020). Doctoral thesis. Masaryk University, Faculty of Informatics, Brno. Supervised by Pavel Rychly.

[57] Tiedemann, J. (2012), Parallel Data, Tools and Interfaces in OPUS. *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC 2012).

[58] *The British National Corpus*, version 2 (BNC World). 2001. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/.

[59] Thompson, L., & Mimno, D. (2020). Topic modeling with contextualized word representation clusters. *arXiv preprint arXiv:2010.12626.*

[60] United Nations Department of Economic and Social Affairs, Population Division (2020). *International Migrant Stock 2020.*

[61] United Nations High Commissioner for Refugees (2023). *UNHCR Refugee Population Statistics Database.*

[62] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[63] von Boguszewski, N., Moin, S., Bhowmick, A., Yimam, S. M., & Biemann, C. (2021). How hateful are movies? a study and prediction on movie subtitles. *arXiv preprint arXiv:2108.10724.*

[64] Wang, F., Beladev, M., Kleinfeld, O., Frayerman, E., Shachar, T., Fainman, E., Lastmann Assaraf, K., Mizrachi, S. & Wang, B. (2023). Text2Topic: Multi-Label Text Classification System for Efficient Topic Detection in User Generated Content with Zero-Shot Capabilities. *arXiv preprint arXiv:2310.14817.*

[65] Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, p., Ravula, A., Wang, Q., Yang, L. & Ahmed, A. (2020). Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33, 17283-17297.