That's What She Said!
The Office Script Analysis

Luuk Boekestein luuk.boekestein@gmail.com
Shantanu Motiani shantanumotiani@gmail.com
Eline Westerbeek Eline.Westerbeek@gmail.com

Abstract

This report presents a comprehensive analysis of the character dialogue in the popular US TV show "The Office." Through the application of sentiment analysis and topic modelling on the lines in the show, and the development of a dialogue generator using Markov chains, we aim to uncover insights into the differences in emotional undertones, character interactions and patterns of the show's dialogue. This project offers a valuable contribution to the analysis of "The Office," showcasing the potential of computational techniques in analysing character dialogue in TV shows.

1 Introduction

The Office US is a popular television show that has captivated worldwide audiences with its iconic characters. In this paper, we employ natural language processing (NLP) techniques to analyze the sentiments, topics, interactions and patterns in the dialogue of the characters.

To accomplish this, we first conduct a sentiment analysis using BERT uncased. This allows us to gain an understanding of how the sentiments may vary across characters, to uncover more about the personalities of the characters.

Additionally, topic modelling, using both non-negative matrix factorisation (NMF) and latent Dirichlet allocation (LDA), is employed to uncover the recurring themes and topics in the dialogue. This is so we can better comprehend the thematic elements and motifs resonating throughout the show.

Finally, using Markov chains to construct a dialogue generator will give us an interactive way to gain a better understanding of the speech patterns of the characters. By training the model on the lines of dialogue, we can simulate the nuances in the dialogue of each character. Through a combination of these NLP techniques and various visualisations, we aim to answer the following research questions:

1. Sentiment analysis – What sentiments are expressed in the dialogues, and how do they vary per character and season?
2. Topic modelling – What are the predominant topics of discussion and how do these evolve over the seasons?
3. Interactions – What are the interactions between characters throughout the seasons?
4. Dialogue generator – Can we create a tool that generates dialogue for a given character?

2 Related work

2.1 Sentiment analysis

The present study benefited from previous scholarly endeavours in the field of sentiment analysis. For instance, Jahuari (2020) conducted an initial investigation of the dataset used in this paper, providing valuable insights. Moreover, Rábay (2020) and Sifre (2020) conducted preliminary sentiment analysis on the same dataset. Additionally, Allen (2018) employed visualizations to depict the interplay of characters within the show.

2.2 Topic modelling

While topic modelling, especially using LDA, is a widespread method to obtain information across unlabelled textual documents, there has been no previous research which applied either LDA or NMF to television scripts. Nonetheless, there is a wide range of research applying it to other domains, ranging from Twitter posts about the Internet of Things (Bian et al., 2016), online reviews (Titov & McDonald, 2008) and for analysis of scientific trends from academic journals (Liu et al.,

2020).

2.3 Markov chains

Markov chains have been widely used for text generation due to their simplicity and ability to capture local dependencies in the data. Several studies have explored the application of Markov chains in this context. Lowe (1993) investigated the emergence of language by training a Markov chain model on Shakespearean language. The study demonstrated how nonsensical text gradually evolves into coherent language by selecting words based on observed probabilities. Gallagher et al. (2008) introduced a Markov chain Monte Carlo (MCMC) approach for text generation. Recent advancements in natural language processing, such as transformer models (Vaswani et al., 2017), have also surpassed the capabilities of Markov chain models.

3 Dataset

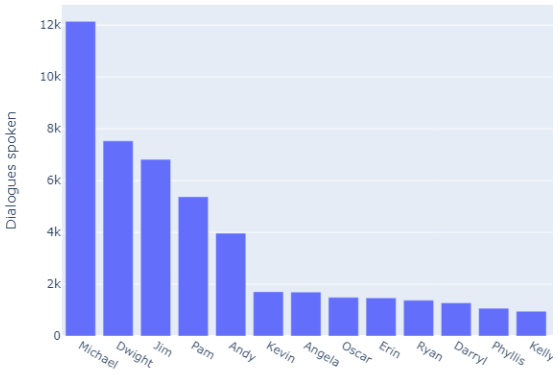

Figure 1: Lines spoken per character

For the project, we used a dataset found on Kaggle (Cominetti, 2022). The dataset contains all character lines that are spoken in The Office, accumulating to 59909 lines. All lines are labelled with season, episode and scene, and a boolean value indicating whether or not a scene was deleted.

Due to the nature of the show, the character dialogue is not evenly distributed, and the main characters have a significantly higher number of lines, as seen in Figure 1.

4 Methods

The code used in this study is available on our GitHub repository `https://github.com/shantanu-555/Text-Mining-Project`, providing a comprehensive resource for reproducibility and further exploration of our methodology.

4.1 Sentiment analysis

4.1.1 Preprocessing

The data underwent a preprocessing pipeline in the context of sentiment analysis, utilizing the inherent preprocessing functionalities incorporated within the employed models. A notable challenge encountered during the data preprocessing was handling character action descriptions such as "[looks in camera]," which were present in the script lines. For the sentiment analysis we completely removed all the descriptions in the text.

4.1.2 Models chosen

We performed a 3-class sentiment analysis classification on the whole dataset, classifying each line as either negative, neutral or positive. We deployed four different sentiment analysis models: BERT uncased/cased (Devlin et al., 2018), DistilBERT (Sanh et al., 2019) and Roberta (Liu et al., 2019), on a small part of the dataset, and compared their performance. For the performance evaluation, we manually annotated 300 lines of the show into each of the three classes. The primary evaluation metric we used for the sentiment analysis was accuracy, along with computation time. We also computed precision, recall, and F1-score to gain further insights into the performance of each model. After testing each of the models on the annotated dataset, we found that BERT uncased resulted in the best performance, with a 63% accuracy score, an MSE of 0.46, and low computational time (see Figure). Roberta had a slightly higher accuracy of 69% but had a significantly higher MSE of 1.29 and was significantly slower. After the initial accuracy testing, we performed sentiment analysis on the entire dataset with BERT uncased.

Model	% Accuracy	MSE
Roberta	67	1.29
BERT Uncased	63	0.46
BERT Cased	54	0.65
DistilBert	41	0.89

Table 1: accuracy and MSE of different models

4.2 Topic modelling

NMF and LDA were employed to generate the topic models. This was done on dialogue grouped by scene, as individual lines are too short and lack context.

For the NMF model, the coherence and reconstruction errors were computed across various topic numbers. Coherence is larger the better the topic model, evaluates the interpretability and quality of topics produced by the model, considering the semantic similarity between highly ranked words within each topic (Stevens et al., 2012). On the other hand, we want to minimise reconstruction error, which entails approximating the original matrix of documents and words by the product of the two lower-dimensional matrices (Chawla, 2017), shown in Figure 2.

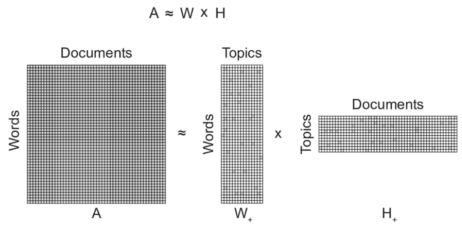

Figure 2: Non-negative matrix factorisation (Goyal, 2021)

To determine the optimal number of topics for LDA, the coherence, log-likelihood, and perplexity were calculated over a range of topic numbers. The log-likelihood assesses how well the latent topics align with the observed data (Loureiro et al., 2021), while perplexity measures the model's ability to predict the remaining words in a document based on a specific topic, after observing only a small portion of it (Guerreiro et al., 2015). A higher log-likelihood and lower perplexity indicate a better fit for the topic model (Loureiro et al., 2021). However, coherence is still the primary criterion for both NMF and LDA.

The clusters of topics were named by interpreting the highly ranked words for each topic and by inspecting scenes that best represented those topics, to gain insight into the topic cluster.

4.3 Dialogue Generator

The model uses Markov Chains, a mathematical system that experiences transitions from one state to another according to certain probabilistic rules. The probability of transitioning to any particular state is dependent solely on the current state and time elapsed. ("Markov Chains — Brilliant Math & Science Wiki").

With dialogues, each word in a character's corpus is 'connected' to every other word with varying probabilities. So given our initial word (or state), the model assigns a probability to every other word in our corpus with its likelihood to follow our initial word (Pernicano).

We used a Python library, Markovify, which can build Markov models of large textual corpora to generate random sentences. Markovify offers options like generating sentences with a maximum character limit, a specified state size and maximum overlap ratio with original sentences.

We created a separate corpus for each of the top 20 characters in the show, which can be found in the 'character lines' sub-directory. Then, we defined the main function which takes a corpus, does the necessary transformations, trains the model on that data, and returns a randomly generated dialogue. We tried different hyper-parameter combinations but settled on the following: a max overlap ratio of 60% with the original sentences, a maximum generated length of 110 characters, and a state size of 2. The model tries a maximum of ten times to generate sentences that align with the above parameters.

5 Results and findings

5.1 Sentiment analysis

The sentiment analysis presented us with interesting findings. The sentiment amongst the lines was distributed fairly equally. The majority of the lines were classified as neutral. Negative lines accounted for 21.6% of the dataset, while positive lines constituted 20.0%.

By analyzing the sentiment at the character level, we found that the most positive characters in the show were Pam and Jim, while Stanley and Angela stood out as the most negatively expressing characters. Additionally, we identified some character development arcs throughout the series. A striking example of this is the sentiment progression of Andy; in season 4, when the character ends his relationship with Angela, a negative trend in sentiment can be observed. Then in season 8, when Andy is promoted to regional manager, his sentiment spikes upwards. This illustrates that the

performed sentiment analysis is able to accurately reflect key events in the progression of the show. An overview of all the observed sentiment trends can be found in Appendix J-N.

5.2 Topic modelling

For the LDA topic model, the coherence, log-likelihood and perplexity for a range of topic numbers are shown in Appendix A. The results indicate that a choice of 5 or 6 topics is ideal, with a good coherence of just over 0.650.

A choice of 6 topics was made as these topics were deemed more coherent. These topics, and their top 10 words are displayed in Appendix B. It was difficult to name these topics, especially topic 3, although other topics, namely topic 2 and topic 4 seemed much more definitive. When looking at the distribution of topics, shown in Appendix C, we see they are relatively balanced.

For the NMF topic models, the coherence and reconstruction error is shown for a range of topics in Appendix D. The ideal number of topics should be either 11 or 13. However, after assessing the topics generated by those, there appeared to be some over-fitting. The next best number of topics, according to the coherence score, is 9, which we deemed much more coherent and easy to name, with a coherence score of 0.655, slightly better than the coherence score for the LDA model.

The topics generated with NMF are shown in Appendix E. The top 10 words are descriptive and portray a clear cluster of topics. Topic 0 however seemed very general, and was named 'general dialogue'. There is an imbalanced distribution of topics, as displayed in Appendix F, as the majority of the scenes were classified as 'general dialogue'. However, this may just represent a feature of the dataset. Apart from the 'general dialogue', the other topics are quite evenly distributed.

We can assess how these topics evolve over the seasons to uncover trends and assess whether these reflect events in the show. The time evolution of the proportion of NMF topics is shown over the seasons in Appendix G, with topic 0 excluded. We can see for instance that topic 7, 'roles/positions in the company', peaks in seasons 1, 3 and 6. In season 1, this may be because the show is establishing the characters and their roles in the company, while in seasons 3 and 6 this reflects the merging of Dunder Mifflin branches and the merger with Sabre respectively, which led to promotions and demotions.

Finally, word clouds for each NMF topic, displayed in Appendix H, were created to give a sense of how heavily weighted the different top words are in each topic. For instance, in topic 1, 'phone calls', we see very clearly that words relating to making phone calls like 'phone', 'hang', 'answer', 'voice', and 'calling', are larger than less related words like 'meeting', 'still' and 'everyone'.

5.3 Interactions

To see how frequently various characters interact, the interactions between the 10 main characters are plotted in Appendix I. We can see that Dwight, Pam, Michael and Jim appear to be the most connected, while Oscar and Ryan have the least interactions. We can see that characters working in the same departments have more interactions as well. For instance, Oscar has more interactions with Kevin and Angela than with Erin, Ryan, Pam and Jim.

5.4 Markov chains

The outcome was a mixed bag. The dialogues generated are plausible but not necessarily coherent. It is evident that the words are put after one another based on pure probabilities and the current state, disregarding the context. This property of Markov Chains especially hurts those dialogues generated that have periods and commas in them because they don't logically follow each other.

Overall, we believe, given the size of the data set and the characteristics of the dialogue written for a sitcom, the dialogue generator does a good job of capturing a character's personality and producing realistic results. The outcomes would change with hyper-parameter tuning for different use cases and the objectives one wants to achieve with it.

6 Conclusion

In closing, a comprehensive analysis of character dialogue in the popular US TV show "The Office" was presented. The results of the sentiment analysis indicated a relatively balanced sentiment distribution, with a significant portion of lines classified as neutral. The sentiment analysis also captured character development arcs and reflected key events in the show, thereby demonstrating the effectiveness of the approach.The study

utilized topic modeling techniques, namely Non-negative Matrix Factorization (NMF) and Latent Dirichlet Allocation (LDA), to gain insights into the recurring themes and topics in the dialogue. The analysis illustrated how these topics evolved over the seasons, reflecting the influence of events and character dynamics. Finally, the paper introduced a dialogue generator based on Markov chains, which was able to capture the unique nuances of each character, enabling users to explore and study their speech patterns in a novel and engaging manner.

7 Author contributions

LB performed preprocessing, sentiment analysis and part of the web app, and wrote the abstract, part of the related work, dataset description, part of the methods, and part of the results. SM did the initial analysis, visualizations of the data, the dialogue generator, and the web app, and wrote part of the results and the conclusion. EW did preprocessing, topic modelling, character interactions, and part of the web app, as well as wrote the introduction, part of the related work, part of the methods and part of the results.

8 References

References

Allen, J. (2018, July 13). Text Mining: Every Line from The Office. Jenna Allen. Retrieved from https://www.jennadallen.com/post/text-analytics-every-line-from-the-office/

Bian, J., Yoshigoe, K., Hicks, A., Yuan, J., He, Z., Xie, M., Guo, Y., Prosperi, M., Salloum, R., & Modave, F. (2016). Mining Twitter to Assess the Public Perception of the "Internet of Things." *PLOS ONE*, 11(7), e0158450. https://doi.org/10.1371/journal.pone.0158450

Chawla, R. (2017, July 30). Topic Modeling with LDA and NMF on the ABC News Headlines dataset. *Medium; ML 2 Vec*. Retrieved from https://medium.com/ml2vec/topic-modeling-is-an-unsupervised-learning-approach-to-clustering-documents-to-discover-topics-fdfbf30e27df

Cominetti, F. (2022). The Office Lines. *www.kaggle.com*. Retrieved from https://www.kaggle.com/datasets/fabriziocominetti/the-office-lines

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018, October 11). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv.org*. Retrieved from https://arxiv.org/abs/1810.04805

Goyal, C. (2021, June 26). Part 15: Step by Step Guide to Master NLP - Topic Modelling using NMF. *Analytics Vidhya*. Retrieved from https://www.analyticsvidhya.com/blog/2021/06/part-15-step-by-step-guide-to-master-nlp-topic-modelling-using-nmf/

Guerreiro, J., Rita, P., & Trigueiros, D. (2015). A Text Mining-Based Review of Cause-Related Marketing Literature. *Journal of Business Ethics*, 139(1), 111–128. https://doi.org/10.1007/s10551-015-2622-4

Jauhari, N. (2020, August 5). The Office Sentiment Analysis. *Kaggle.com*. Retrieved from https://www.kaggle.com/code/nilimajauhari/the-office-sentiment-analysis

Liu, S., Zhang, R.-Y., & Kishimoto, T. (2020). Analysis and prospect of clinical psychology based on topic models: hot research topics and scientific trends in the latest decades. *Psychology, Health & Medicine*, 26(4), 395–407. https://doi.org/10.1080/13548506.2020.1738019

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.org*. Retrieved from https://arxiv.org/abs/1907.11692

Loureiro, S. M. C., Guerreiro, J., & Han, H. (2021). Past, present, and future of pro-environmental behavior in tourism and hospitality: a text-mining approach. *Journal of Sustainable Tourism*, 1–21. https://doi.org/10.1080/09669582.2021.1875477

Rábay, K. (2020, June 2). NLP on The Office series. *Medium*. Retrieved from https://towardsdatascience.com/nlp-on-the-office-series-cf0ed44430d1

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv.org*. Retrieved from https://arxiv.org/abs/1910.01108

Sifre, R. (2020, March 12). Tidy Text Mining with The Office. *Robin Sifre*. Retrieved from https://robin-sifre.netlify.app/post/tidy/tidytues-theoffice/

Stevens, K., Kegelmeyer, P. W., Andrzejewski, D., & Buttler, D. (2012). Exploring Topic Coherence over Many Models and Many Topics. *ACL Anthology*, 952–961.

Titov, I., & McDonald, R. (2008). Modeling online reviews with multi-grain topic models. *Proceeding of the 17th International Conference on World Wide Web - WWW '08*. https://doi.org/10.1145/1367497.1367513

9 Appendix

A LDA scores

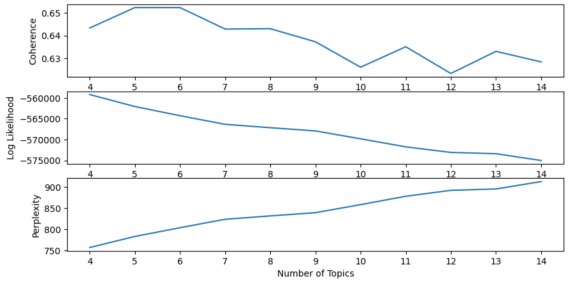

Figure 3: Coherence, Log-likelihood and perplexity for a range of topic numbers for LDA

B LDA topics

No.	Topic name	Top 10 words
0	the company?	phone sorry paper dunder mifflin woman company scranton listen black
1	breaks/meetings?	friend meeting break sorry singing welcome weird lunch family voice
2	parties/holidays?	party money christmas everybody today check throw movie tonight dream
3	??	night year question laugh manager watch got ready happened exactly
4	the office?	getting office month picture excuse sorry talking computer person table
5	humour?	office laugh place better crazy sound remember pretty number funny

Table 2: Topics generated by LDA.

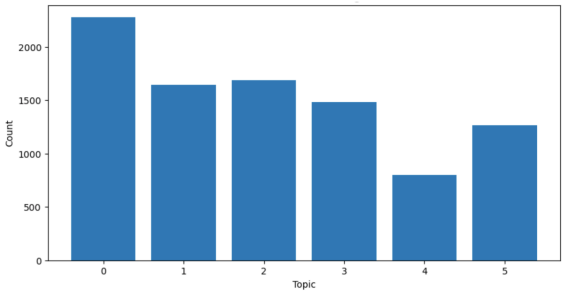

Figure 4: Distribution of LDA topics

C LDA distribution

D NMF scores

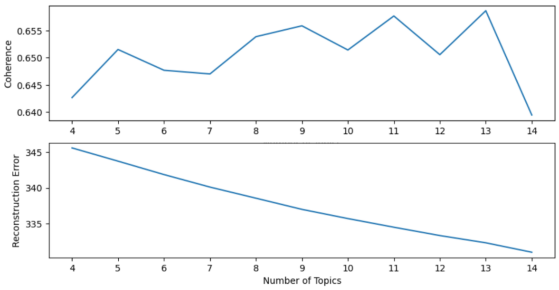

Figure 5: Coherence and reconstruction error for a range of topic numbers for NMF

E NMF topics

F NMF distribution

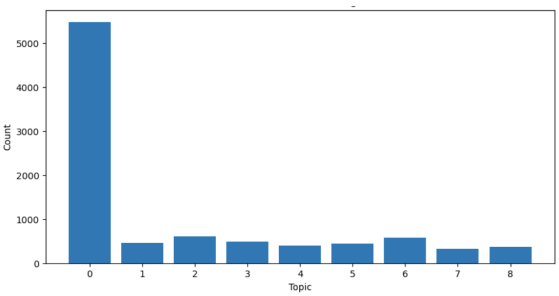

Figure 6: Distribution of NMF topics

No.	Topic name	Top 10 words
0	general dialogue	everybody today talking start better woman meeting night minute place
1	phone calls	phone hang ring answer client calling listen imitating transfer sound
2	the office	office camera chair place knock parking conference walking lunch building
3	the company and business	paper dunder mifflin company business client scranton sale salesman question
4	parties and party planning	party christmas committee planning start throw break everybody starting birthday
5	apologising	sorry voice stupid excuse hand apology client doing probably pretty
6	humour and jokes	laugh funny laughing pretty tonight year place camera welcome drink
7	roles/positions in the company	manager regional assistant branch scranton wallace person position better sale
8	friendships and relationships	friend smart happy totally thinking different understand girlfriend sitting voice

Table 3: Topics generated by NMF.

G NMF evolution

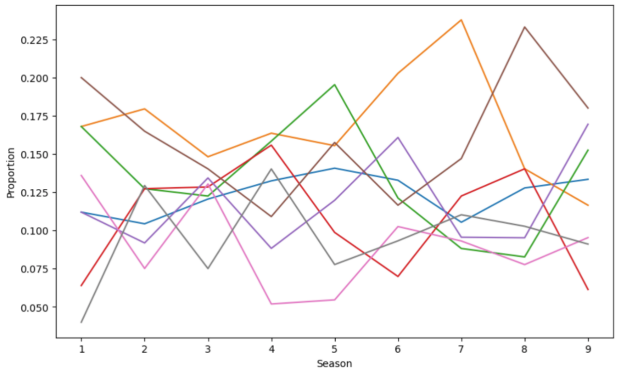

Figure 7: Proportion of NMF topics over the seasons, excluding topic 0

H NMF word clouds

Figure 8: Word clouds for each NMF topic

I Character interactions

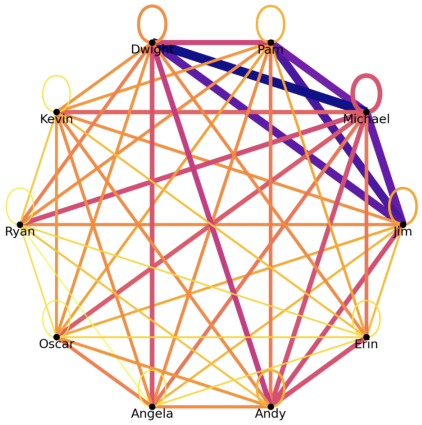

Figure 9: Interactions between 10 main characters

J Character sentiment ranking

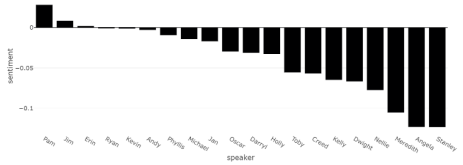

Figure 10: Average sentiment of all characters

K Sentiment of all characters throughout the season

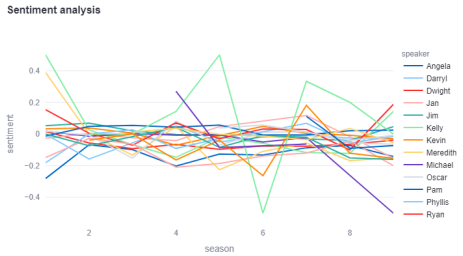

Figure 11: Sentiment of characters throughout the seasons

L Sentiment progression of Andy

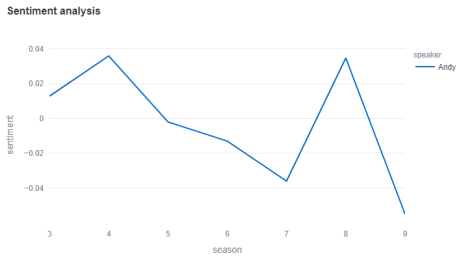

Figure 12: Sentiment progression of Andy

M Sentiment progression of Jim

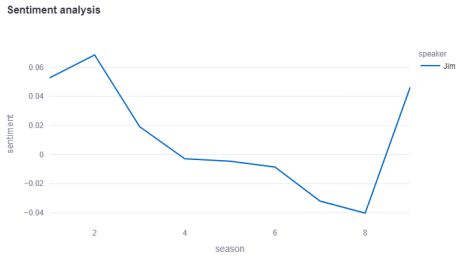

Figure 13: Sentiment progression of Jim

N Sentiment progression of Pam

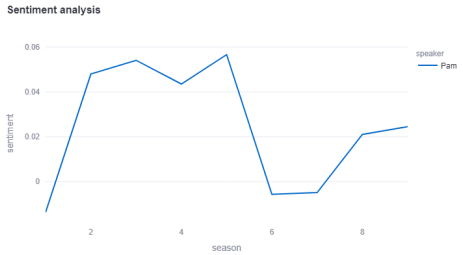

Figure 14: Sentiment progression of Pam

O Sentiment progression of Michael

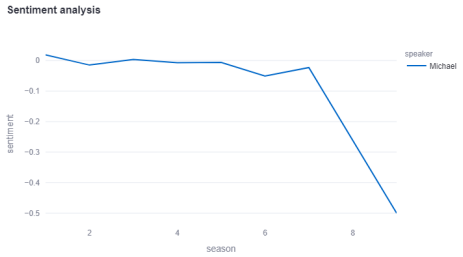

Figure 15: Sentiment progression of Michael