Using Text Embeddings to Predict Plot Points in Novels: Agatha Christie

Ryan Peruski
Department of EECS
University of Tennessee, Knoxville
yhg461@vols.utk.edu

Nolan Coffey
Department of EECS
University of Tennessee, Knoxville
ncoffey3@vols.utk.edu

Triton Eden
Department of EECS
University of Tennessee, Knoxville
teden@vols.utk.edu

William Duff
Department of EECS
University of Tennessee, Knoxville
wduff@vols.utk.edu

***Abstract*—This paper investigates the application of Natural Language Processing (NLP) techniques, specifically text embeddings, to analyze and predict key plot points in Agatha Christie's novels. Our research group used word embedding models, including Continuous Bag of Words (CBOW) and Skip-gram, to explore relationships among various characters and objects within the narratives. The dataset comprises several Agatha Christie novels sourced from Project Gutenberg. Our analysis compares the performance of these embedding models and examines the semantic relationships between significant plot elements, such as the protagonist, antagonist, victim, and murder weapon. Although the study aimed to uncover structural patterns in Christie's mystery writing (to, perhaps, use the embeddings as a predictor for other key plot points), we did not find conclusive results, suggesting that additional methods may be needed to fully capture such patterns. This work illustrates the potential and limitations of statistical NLP methods in literary analysis.**

***Index Terms*—NLP, Natural Language Processing, Agatha Christie, Mystery Novels, Word Embeddings**

I. Introduction

Literary analysis is a long-standing practice used to derive meaning from a work and understand the style, themes, and techniques employed by the author [6]. Traditionally, this process has involved close reading to examine elements such as plot, characterization, and language. However, with the advent of modern computational techniques, new methods have emerged to explore literature in innovative ways. One such method involves using word embeddings, a tool from Natural Language Processing (NLP), to analyze plot structure and uncover deeper patterns in writing. In this report, we apply these "Computational Linguistics" techniques to the works of Agatha Christie, a master of the mystery genre renowned for her intricate plots and unexpected twists. Christie's novels, including classics like *Murder on the Orient Express* and *The Murder of Roger Ackroyd*, have captivated readers with their clever storytelling and suspenseful narratives. By leveraging statistical methods from Natural Language Processing, we aim to explore Christie's distinctive writing style and examine whether patterns in her plots can be identified and even used to predict the outcomes of her novels.

II. Dataset

A. Overview

For this study, we selected our dataset from Project Gutenberg [5], a well-established digital library offering free access to public domain literary works. Project Gutenberg provides reliable, standardized text formats suitable for Natural Language Processing analysis, making it an ideal source for literary datasets. By using this open-access resource, we ensured consistent formatting across texts, facilitating a more straightforward pre-processing pipeline and maintaining a high level of text quality.

Our dataset includes three notable Agatha Christie novels: The Mysterious Affair at Styles, The Murder of Roger Ackroyd, and The Murder on the Links. These works were chosen for their significance in Christie's body of work and their consistent use of Hercule Poirot as the central detective, allowing us to examine a recurring character across multiple narratives. By focusing on novels with the same protagonist, we can more effectively analyze character relationships and thematic patterns unique to Poirot's investigative style, making these selections well-suited for studying structural patterns in Christie's writing. Together, these novels provide a focused yet representative sample of Christie's writing style and plot construction, forming a strong foundation for our NLP-based analysis.

B. Data Cleaning

To begin the data cleaning process, we applied standard pre-processing techniques such as tokenization, lemmatization, and normalization. However, one key step we introduced differs from typical pre-processing: the use of multi-word embeddings to ensure consistency in character references. In many cases, characters are referred to by various names, such as both their first and last names or a shortened version. To address this, we created a single token for each character across the novel. For instance, 'Alfred Inglethorp' and 'Mr. Inglethorp' were both tokenized to the same value. This adjustment ensures more accurate word embeddings and helps maintain consistency in

our analysis. Importantly, this step requires compiling a list of characters for each novel to correctly map the names. Skipping this step would lead to less accurate results, as the model would treat different forms of the same character's name as separate tokens.

III. Methods

A. Algorithms Used

For our word embeddings, we implemented two widely-used algorithms: Continuous Bag of Words (CBOW) and Skip-gram. Both methods belong to the Word2Vec model family and are effective in capturing the relationships between words based on their contextual co-occurrences in text. The CBOW algorithm predicts a target word by considering the surrounding context words, making it useful for capturing common word patterns and frequent terms. In contrast, the Skip-gram algorithm predicts surrounding words given a specific target word, making it particularly effective in identifying less frequent words and nuanced relationships. Utilizing both CBOW and Skip-gram allows us to explore potential connections between characters, objects, and other plot elements in Christie's novels, aiming to uncover patterns that may reveal insights into her narrative structure.

B. Code Libraries

For this study, we used the Word2Vec model from the `gensim.models` library, a robust tool for generating word embeddings that supports both CBOW and Skip-gram algorithms. Additionally, we used the `nltk` library for essential pre-processing steps, including tokenization and lemmatization, which prepared the text data for analysis by standardizing word forms and breaking down the text into manageable tokens. For word embedding visualization, we used *TensorBoard* and the *Projector* options. *Tensorboard* worked extremely well, as they automatically chose "most important dimensions", i.e. the dimensions that displayed the most variance between the embeddings, to display. Together, these libraries provided a comprehensive framework for embedding, analyzing, and visualizing the text data efficiently.

IV. Results

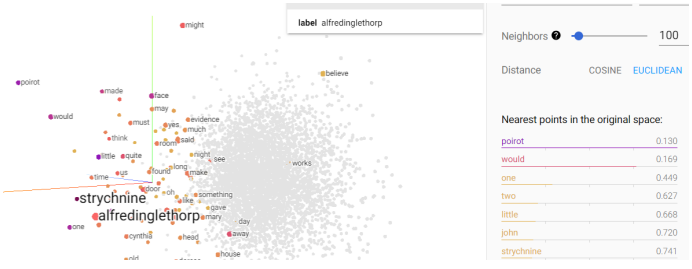

Fig. 1. The Mysterious Affair at Styles: CBOW, unlemmatized,antagonist [4]

Trial	Prot to Antag	Prot to Murder Weapon	Antag to Murder Weapon	Victim to Murder Weapon	Victim to Antag
Styles CBOW	0.130	0.836	0.741	2.750	3.486
Styles Skip-gram	0.295	9.370	0.157	0.130	0.152
Styles CBOW lem	0.244	0.110	0.235	3.110	3.327
Styles Skip-gram lem	0.239	0.333	0.240	0.186	0.166
Ackroyd CBOW	3.094	2.494	0.607	1.342	1.940
Ackroyd Skip-gram	0.163	0.277	0.257	0.225	0.269
Ackroyd CBOW lem	2.353	1.907	0.454	1.380	1.825
Ackroyd Skip-gram lem	0.160	0.246	0.251	0.233	0.228
Links CBOW	2.666	3.127	0.473	2.229	2.692
Links Skip-gram	0.262	0.159	0.227	0.945	1.021
Links CBOW lem	1.945	2.505	0.575	2.346	2.909
Links Skip-gram lem	0.302	0.176	0.251	0.752	0.850

TABLE I
EUCLIDEAN DISTANCES OF GIVEN TRIALS (USE CTRL + TO ZOOM FOR RESULTS IF NEEDED.)

A. Metric Selection

In our analysis, we opted to use Euclidean distance rather than cosine similarity for measuring the relationships between words, characters, and plot elements. While cosine similarity emphasizes the semantic meanings and directional similarity of word vectors, it tends to obscure important differences in magnitude. Euclidean distance, on the other hand, accounts for both magnitude and direction, providing a more nuanced measure of proximity within the novels. Given that our focus is on frequency of co-occurrence rather than purely semantic relationships, Euclidean distance aligns better with our research objectives. Additionally, we observed that cosine similarity often recorded a value of 0 for many words, indicating a strong relationship, but that made ranking those words by proximity difficult.

Please note that, the shorter the euclidean distance, the closer in proximity the tokens are to each other, so the tokens "car" and "bus" would have a lower euclidean distance than that of the tokens "car" and "alien".

B. Discussion

1) The Murder of Roger Ackroyd: The analysis of *The Murder of Roger Ackroyd* revealed interesting patterns in the word embeddings generated through the CBOW and Skip-gram algorithms. The results indicate that the difference between embeddings with and without lemmatization is relatively minor. However, a more significant disparity is observed between the distances produced by the CBOW and Skip-gram models, with CBOW yielding notably higher distances.

The lowest distance in the CBOW models correspond to the relationship between the antagonist and the murder weapon (0.607 and 0.454 for lemmatization and no lemmatization, respectively [4]), which is logical, given that the antagonist employed the murder weapon. In contrast, the Skip-gram algorithms identified the protagonist and antagonist as the most similar entities (0.160 and 0.163 for lemmatization and no lemmatization, respectively [4]) Notably, there was a case where the murder weapon was more closely related to the protagonist than to the antagonist, suggesting a nuanced relationship within the narrative structure.

2) The Mysterious Affair at Styles: In examining *The Mysterious Affair at Styles*, the best performance was observed in the CBOW model without lemmatization. When querying for Poirot, the protagonist and detective, the closest neighbor identified was the antagonist, Alfred Inglethorp (represented as "alfredinglethorp" in the word vectors), with a Euclidean distance of 0.130 [4]. When searching for Inglethorp, Poirot

was found to be the closest neighbor, while the murder weapon, strychnine, ranked seventh with a distance of 0.741 [4]. These results highlight the close relationships among key characters and elements in the narrative.

3) The Murder on the Links: The analysis of *The Murder on the Links* reinforced the trends observed in the previous novels. Similar to *The Murder of Roger Ackroyd*, the Skip-gram distances were notably closer than those produced by the CBOW model. Even with the Skip-gram embeddings, none of the relationships identified were within the top 100 nearest neighbors of one another. This indicates a potential limitation in the model's ability to discern connections in this particular narrative.

4) The Bottom Line: From our analysis, it is evident that, although, as expected, the protagonist, antagonist, victim, and murder weapon are all similar to each other in each of the trials, there is no conclusive way of empirically finding one token by its distance from another token. In other words, say we use a model on an Agatha Christie novel that is not of the three we used. If we knew the protagonist is Poirot, we were hoping we could determine the antagonist by looking at the, say, sixth most similar word, or, the word with a distance of, say, approximately 0.106. Since the ranked most similar words and distances varied too much from trial to trial, we could not use the word embeddings as an accurate predictor of who the antagonist is, given the protagonist, for example.

C. Limitations

Despite the valuable insights gained from our analysis, we encountered several limitations during our research. One significant challenge was the necessity of manually updating the book metadata, which required us to identify key elements such as the murder weapon, protagonist, antagonist, and victim beforehand. This manual process not only increased the complexity of our analysis but also introduced the potential for human error in the data entry. Additionally, we aimed to incorporate a temporal measurement of our results to better understand how distances between characters and plot elements evolve before and after the climaxes of the stories. However, this aspect was not fully realized in our current analysis, highlighting an area for future research to enhance the depth of our findings.

V. Conclusion

In conclusion, this study demonstrates how Natural Language Processing techniques, particularly word embeddings, can contribute to literary analysis. By employing multi-word embeddings to ensure consistent representation of characters, we approached Agatha Christie's novels from a computational perspective. While our analysis did not reveal conclusive patterns in character relationships or plot structures, it underscored the importance of careful pre-processing, particularly in handling character names and references. These findings suggest that while the current techniques may not yield definitive predictions of narrative outcomes, they offer a foundation for further exploration in this field.

Looking ahead, we see several avenues for future research. Implementing co-reference resolution could significantly enhance the performance of our word embeddings by better capturing the relationships among characters throughout the narratives. Additionally, analyzing how embeddings change over the course of each book, chapter by chapter, may reveal interesting dynamics, such as the antagonist getting closer to the murder weapon over time. We could also enrich our analysis by adding data for each character and employing data-backed logic to assess which character is most likely to be the murderer. Lastly, establishing a systematic method for acquiring metadata manually would further streamline our analysis and improve the reliability of our results. Future studies that explore these ideas could deepen our understanding of literary analysis as a whole and enhance the synergy between computational techniques and traditional literary criticism.

VI. Acknowledgment

We would like to thank the University of Tennessee, Knoxville, and the EECS Department for their support. Special thanks to Dr. Edmon Begoli and his assistants for their guidance and mentorship throughout this Natural Language Processing class.

References

[1] Christie, A. (1926). *The murder of Roger Ackroyd*. Project Gutenberg. Retrieved from https://www.gutenberg.org/ebooks/61262
[2] Christie, A. (1920). *The mysterious affair at Styles*. Project Gutenberg. Retrieved from https://www.gutenberg.org/ebooks/863
[3] Christie, A. (1923). *The murder on the links*. Project Gutenberg. Retrieved from https://www.gutenberg.org/ebooks/61075
[4] Peruski, R. Coffey, N. Eden, T. Duff, W. (2024). *nlpproject* [GitHub repository]. GitHub. https://github.com/Silverasdf/nlpproject
[5] Project Gutenberg. (n.d.). Project Gutenberg. https://www.gutenberg.org/
[6] Scribbr. (n.d.). *How to write a literary analysis essay: A step-by-step guide.* Scribbr. https://www.scribbr.com/academic-essay/literary-analysis/