Computing Science

Comparative Analysis of Document-Level Embedding Methods for Similarity Scoring on Shakespeare Sonnets and Taylor Swift Lyrics

Klara Krämer

No portion of the work contained in this document has been submitted in support of an application for a degree or qualification of this or any other university or other institution of learning. All verbatim extracts have been distinguished by quotation marks, and all sources of information have been specifically acknowledged.

Signed:

Date: November 8, 2024

CS4040 Report

Department of Computing Science
University of Aberdeen
King's College
Aberdeen AB24 3UE

November 2024

Comparative Analysis of Document-Level Embedding Methods for Similarity Scoring on Shakespeare Sonnets and Taylor Swift Lyrics

Klara Krämer

1 Introduction

Document similarity assessment plays an important role in various natural language processing (NLP) applications, such as information retrieval, plagiarism detection, recommendation systems, and sentiment analysis [10, 15]. For instance, in recommendation systems, document similarity helps personalise suggestions by finding content that closely matches user preference. These tasks rely on accurate measurements of how similar documents are in terms of their structure, content, and meaning, which depends on the chosen document embedding method. Various methodologies can be employed to obtain document-level embeddings, and the choice of methods directly impacts the accuracy and usefulness of the similarity scores calculated [12, 15].

This paper explores and evaluates three methods of document-level embeddings for document similarity scoring: Averaged Word2Vec [14], TF-IDF weighting [13], and BERT embeddings [5]. Each method offers distinct advantages and limitations, and this comparative study applies these techniques to a diverse set of texts - specifically, the sonnets of William Shakespeare and the lyrics of Taylor Swift. Generating document-level embeddings for documents in these two contrasting genres and analysing their similarity scores will help evaluate how effective, reliable, robust, and adaptable different embedding techniques are for scoring document similarity.

The methodology employed in this project involves creating document-level cosine similarity matrices for each dataset and method, followed by both qualitative and quantitative analyses. This includes examining the most and least similar document pairs, calculating and comparing average similarity scores, and counting document pairs with zero similarity.

2 Background and Related Work

There is a large body of research on document similarity measures, with different studies focusing on various aspects of document similarity and different applications.

Traditionally, methods such as TF-IDF (Term Frequency-Inverse Document Frequency) as proposed by Sparck Jones have been employed to compute the similarity between documents [13]. While TF-IDF captures relative term relevance within documents, the method cannot encode word meaning and contextual information, limiting its effectiveness for nuanced language like poetry or lyrics [11].

The more recent advancement Word2Vec, developed by Mikolov et al., maps words to continuous vector spaces based on their co-occurring context words [14]. This approach captures semantic relationships between words, but how to use these word-level embeddings to create whole-document embeddings remains an open research question. Proposed solutions include the Paragraph Vector model Doc2Vec [8], and simple but effective methods such as averaging word embeddings to obtain sentence- and document-level embeddings from their constituent words [16]. Addressing these limitations, Devlin et al. introduced BERT (Bidirectional Encoder Representation from Transformers), a pre-trained model that generates context-sensitive embeddings [5].

Most studies on document similarity scoring focus on specific application domains, for instance, patent-to-patent comparisons. Younge et al. showed that a TF-IDF-based vector space model outperforms traditional patent classifiers in terms of accuracy, specificity, and generality [17]. However, they did not compare this approach with other vectorisation models.

In a study building on the above, Shahmirzadi et al. analysed the performance of different vector space models in measuring semantic text similarity on patents [12]. Comparing TF-IDF weighting, topic modelling, and Doc2Vec embeddings, they found that pre-trained and highly tuned neural embeddings like Doc2Vec provided the best results. However, they also observed a trade-off between computational efficiency and accuracy, noting that a simple TF-IDF approach was more practical for their application. The authors of this paper noted a lack of studies evaluating the performance of different vectorisation methods on more challenging datasets, which this paper aims to address.

Within the domain of Shakespeare sonnets and Pop lyrics, there is only limited NLP research. Some existing papers utilised clustering approaches to attempt to answer the Shakespeare authorship question [1] [1, 2]. Koppel and Seidman applied first- and second-order document similarity measures to Shakespeare plays in a similar authorship verification task [7]. Their evaluation used 40 plays attributed to Shakespeare, leaving room for the future work this paper focuses on, as a larger corpus of shorter documents such as sonnets may yield different results for document similarity measures.

In a study on music similarity, Knees and Schedl noted that lyrics-based features can sometimes outperform advanced measures such as audio-based features in assessing song semantics and similarity [6]. A more domain-specific paper examined linguistic patterns in Taylor Swift's writing, focusing on word and 3-gram co-occurrences. They conducted a quantitative study on these lyrics based on the assumption that song lyrics often originate as lyric poems [4], an assumption that this paper will help evaluate by calculating similarity scores between Shakespearean poetry and Swift lyrics.

[1] The Shakespeare authorship question asks whether all works commonly attributed to William Shakespeare have truly been written by him, or whether there have been ghostwriters or false attributions.

3 Research Question

As the above body of related work indicates, document similarity scoring is an interesting area of research which so far is quite unexplored within specialised domains such as poetry and lyrics. Therefore, this paper investigates the effectiveness of different techniques in generating document similarity scores for these domains.

The aim is to identify how the three embedding methods under evaluation - namely averaged Word2Vec, TF-IDF weighting, and BERT embeddings - affect the document similarity scores generated. To establish this, the methods are to be applied to (1) a dataset of Taylor Swift lyrics, (2) a dataset of Shakespeare sonnets, and (3) a combination of the above two in one dataset.

Specifically, the following hypotheses are to be tested:

1. TF-IDF weighting will show higher average similarity scores within the Taylor Swift dataset than the other two datasets, reflecting the straightforward narrative style of the lyrics.

2. Averaging the Word2Vec embeddings of the words within each document will result in higher similarity scores for Taylor Swift songs than for Shakespeare sonnets. This is likely because of the more contemporary language of the Swift lyrics, which aligns better with the inherent Word2Vec training data.

3. Averaged Word2Vec embeddings will assign higher scores to Swift-Shakespeare pairs in the joint dataset than the TF-IDF method since Word2Vec can recognise similarities in meaning even when the vocabulary differs significantly.

4. BERT embeddings will perform better than the other methods at identifying similarities in the Shakespeare sonnets due to BERT's ability to capture the contextual nuances in poetic language.

These hypotheses will be evaluated by calculating document-level cosine similarity matrices using each method on each of the 3 datasets. Each matrix will then be analysed qualitatively by examining the 3 most and 3 least similar documents according to each method. Additionally, the average similarity score for each matrix will be calculated, and the number of document pairs with a similarity score of 0 will be counted to test the hypotheses quantitatively. To test hypothesis 3, the average cosine similarity of Swift-Shakespeare pairs in the joint dataset will be compared across the different methods.

The datasets on which the above analyses will be performed are taken from GitHub, utilising the repositories by Marquez [9] and Finch [3] for the Shakespeare sonnets and Swift lyrics respectively. Since the Swift dataset contains 232 songs, but there are only 154 sonnets by William Shakespeare, 154 of the Swift songs are selected at random to ensure equal numbers across the two datasets. See Table 1 in the Appendix for a list of the Taylor Swift songs that were excluded.

References

[1] Refat Aljumily. Hierarchical and non-hierarchical linear and non-linear clustering methods to "shakespeare authorship question". *Social Sciences*, 4(3):758–799, 2015.

[2] Ahmed Shamsul Arefin, Renato Vimieiro, Carlos Riveros, Hugh Craig, and Pablo Moscato. An information theoretic clustering approach for unveiling authorship affinities in shakespearean era plays and poems. *PloS one*, 9(10):e111445, 2014.

[3] Derek Finch. Corpus-of-taylor-swift. `https://github.com/sagesolar/Corpus-of-Taylor-Swift/blob/main/`, 2024.

[4] Faruq Ahmad Kendong, Afifah Sakinah Daud, and Aeisha Joharry. A corpus-driven analysis of taylor swift's song lyrics. *International Journal of Modern Languages and Applied Linguistics*, 7(2):59–82, 2023.

[5] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota, 2019.

[6] Peter Knees and Markus Schedl. A survey of music similarity and recommendation from music context data. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 10(1):1–21, 2013.

[7] Moshe Koppel and Shachar Seidman. Detecting pseudepigraphic texts using novel similarity measures. *Digital Scholarship in the Humanities*, 33(1):72–81, 2018.

[8] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.

[9] Aaron Marquez. Generate-shakespeare-sonnets. `https://github.com/enerrio/Generate-Shakespeare-Sonnets/blob/master/`, 2017.

[10] James H Martin and Daniel Jurafsky. *Speech and Language Processing*. Pearson Prentice Hall, 2024.

[11] Gerard Salton and Chris Buckley. Approaches to text retrieval for structured documents. Technical report, Cornell University, 1990.

[12] Omid Shahmirzadi, Adam Lugowski, and Kenneth Younge. Text similarity in vector space models: a comparative study. In *2019 18th IEEE international conference on machine learning and applications (ICMLA)*, pages 659–666. IEEE, 2019.

[13] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

[14] Greg Corrado Tomas Mikolov, Kai Chen and Jeffrey Dean. Efficient estimation of word representations in vector space. *ICLR Workshop*, 2013.

[15] Jiapeng Wang and Yihong Dong. Measurement of text similarity: a survey. *Information*, 11(9):421, 2020.

[16] Zhibo Wang, Long Ma, and Yanqing Zhang. A novel method for document summarization using word2vec. In *2016 IEEE 15th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, pages 523–529. IEEE, 2016.

[17] Kenneth A Younge and Jeffrey M Kuhn. Patent-to-patent similarity: A vector space model. *Available at SSRN 2709238*, 2016.

4 Appendix

ID	Album	Title
TSW:01	Taylor Swift	Tim McGraw
TSW:03	Taylor Swift	Teardrops On My Guitar
TSW:04	Taylor Swift	A Place In This World
TSW:05	Taylor Swift	Cold As You
TSW:06	Taylor Swift	The Outside
TSW:10	Taylor Swift	Mary's Song (Oh My My My)
TSW:11	Taylor Swift	Our Song
TSW:13	Taylor Swift	Invisible
TSW:14	Taylor Swift	A Perfectly Good Heart
FER:02	Fearless (Taylor's Version)	Fifteen
FER:08	Fearless (Taylor's Version)	Tell Me Why
FER:14	Fearless (Taylor's Version)	Jump Then Fall
FER:17	Fearless (Taylor's Version)	Come In With The Rain
FER:18	Fearless (Taylor's Version)	Superstar
FER:24	Fearless (Taylor's Version)	That's When
FER:26	Fearless (Taylor's Version)	Bye Bye Baby
SPN:04	Speak Now (Taylor's Version)	Speak Now
SPN:06	Speak Now (Taylor's Version)	Mean
SPN:07	Speak Now (Taylor's Version)	The Story Of Us
SPN:09	Speak Now (Taylor's Version)	Enchanted
SPN:14	Speak Now (Taylor's Version)	Long Live
SPN:22	Speak Now (Taylor's Version)	Timeless
RED:01	Red (Taylor's Version)	State Of Grace
RED:04	Red (Taylor's Version)	I Knew You Were Trouble
RED:07	Red (Taylor's Version)	I Almost Do
RED:13	Red (Taylor's Version)	The Lucky One
RED:14	Red (Taylor's Version)	Everything Has Changed
RED:18	Red (Taylor's Version)	Come Back...Be Here
RED:19	Red (Taylor's Version)	Girl At Home
RED:21	Red (Taylor's Version)	Ronan
RED:23	Red (Taylor's Version)	Nothing New
RED:26	Red (Taylor's Version)	I Bet You Think About Me
RED:27	Red (Taylor's Version)	Forever Winter
NEN:01	1989 (Taylor's Version)	Welcome To New York
NEN:11	1989 (Taylor's Version)	This Love

NEN:15	1989 (Taylor's Version)	You Are In Love
NEN:17	1989 (Taylor's Version)	"Slut!"
NEN:18	1989 (Taylor's Version)	Say Don't Go
NEN:21	1989 (Taylor's Version)	Is It Over Now?
NEN:22	1989 (Taylor's Version)	Sweeter Than Fiction
REP:01	Reputation	...Ready For It?
REP:02	Reputation	End Game
REP:03	Reputation	I Did Something Bad
REP:05	Reputation	Delicate
REP:09	Reputation	Getaway Car
REP:11	Reputation	Dancing With Our Hands Tied
REP:13	Reputation	This Is Why We Can't Have Nice Things
LVR:01	Lover	I Forgot That You Existed
LVR:08	Lover	Paper Rings
LVR:10	Lover	Death By A Thousand Cuts
LVR:15	Lover	Afterglow
LVR:17	Lover	It's Nice To Have A Friend
FOL:01	Folklore	The 1
FOL:02	Folklore	Cardigan
FOL:03	Folklore	The Last Great American Dynasty
FOL:09	Folklore	This Is Me Trying
FOL:10	Folklore	Illicit Affairs
FOL:12	Folklore	Mad Woman
FOL:16	Folklore	Hoax
EVE:08	Evermore	Dorothea
EVE:11	Evermore	Cowboy Like Me
EVE:12	Evermore	Long Story Short
MID:06	Midnights (The Til Dawn Edition)	Midnight Rain
MID:09	Midnights (The Til Dawn Edition)	Bejeweled
MID:10	Midnights (The Til Dawn Edition)	Labyrinth
MID:11	Midnights (The Til Dawn Edition)	Karma
MID:16	Midnights (The Til Dawn Edition)	Paris
MID:19	Midnights (The Til Dawn Edition)	Would've, Could've, Should've
TPD:03	The Tortured Poets Department (The Anthology)	My Boy Only Breaks His Favorite Toys
TPD:07	The Tortured Poets Department (The Anthology)	Fresh Out The Slammer
TPD:08	The Tortured Poets Department (The Anthology)	Florida!!!
TPD:09	The Tortured Poets Department (The Anthology)	Guilty As Sin?
TPD:16	The Tortured Poets Department (The Anthology)	Clara Bow

TPD:19	The Tortured Poets Department (The Anthology)	The Albatross
TPD:20	The Tortured Poets Department (The Anthology)	Chloe Or Sam Or Sophia Or Marcus
TPD:22	The Tortured Poets Department (The Anthology)	So High School
TPD:26	The Tortured Poets Department (The Anthology)	The Prophecy
TPD:27	The Tortured Poets Department (The Anthology)	Cassandra

Table 1: Swift Songs that were excluded from the dataset