A Novel Approach for Semantic Similarity in English Text Document

A THESIS SUBMITTED IN PARTIAL FULLFILLMENT OF THE REQUIREMENT FOR DOCTRAL PROGRAMME IN MANAGEMENT

INDIAN INSTITUTE OF MANAGEMENT, INDORE

BY

SUMIT KUMAR (2017FPM17)

INFORMATION SYSTEMS AREA

UNDER THE GUIDANCE OF

Thesis Advisory Committee:

Prof. Shubhamoy Dey [Chair]

Prof. Rajhans Mishra [Member]

Prof. Bhavin J. Shah [Member]

INDIAN INSTITUTE OF MANAGEMENT INDORE

ABSTRACT

The ever-increasing digital landscape generates an estimated 2.5 quintillion bytes of data daily, with 80% of that comprising unstructured textual content. This unprecedented data deluge poses significant challenges for human comprehension and analysis, requiring years to process even a single day's worth of text data. This necessitates the development of intelligent systems capable of autonomously understanding, interpreting, and analysing textual data, free from human biases and limitations.

Natural language processing is the specialised area of Artificial Intelligence (Autonomous intelligent agents capable of mimicking human like intelligence) that focuses on building advanced Natural Language Understanding (NLU) and Natural Language Generation (NLG) capabilities. The pinnacle of achieving human like intelligence to solve multiple diverse & complex tasks by an AI agent is referred as Artificial General Intelligence (AGI). The pursuit to build such a system that truly exhibits all the characteristics of AGI has led to the development of Generative AI (Gen AI).

Large Language Model (LLM), using powerful AI technology, revolutionizes how to handle text interactions and tasks. The emergence of LLMs marks a monumental leap in natural language processing, enabling machines to converse and comprehend human language for the first time. However, this advancement raises concerns regarding potential misuse, particularly in copyright violation and plagiarism. Unfortunately, current plagiarism detection tools, even the state-of-the-art plagiarism detectors dominating the market like Turnitin, rely on lexical and syntactic comparisons, failing to grasp the semantic nuance of text.

This research aims to bridge this critical gap by developing a methodology capable of detecting semantic similarity in textual content. Traditional methods for determining document similarity, like TF-IDF, often fails to capture the true meaning conveyed by the text. This

limitation stems from their extensive dependency on surface-level feature like word frequency which neglects deeper semantic relationships. Of late, researchers have explored alternative features beyond simple word count to address this.

This study investigates the importance of non-overlapping features for achieving accurate semantic similarity between documents. The hypothesis focuses on unique features, providing richer insights into the texts' meaning. The exploration led us to go beyond information content of mere word frequencies. This dissertation delved into topic frequencies and order, scrutinizing how the distribution and sequence of topics shape the document's meaning. Additionally, the work investigated the power of named entities and their order, recognizing their potential to reveal hidden connections and relationships. By incorporating these novel features, this research work aims to elevate the bar on semantic similarity, achieving a more nuanced and accurate understanding of the relationships between text documents.

The research highlights the limitations of existing term frequency (TF) approaches. While TF captures word occurrences, it needs to be more nuanced regarding the intricacies of natural language. It treats each word as an isolated entity, blind to the more profound meaning woven by the interplay of topics and their sequence.

An exhaustive literature survey helped to discover a hidden gem i.e. relationship between the topic and its order. Unfortunately its potential for revolutionizing semantic similarity calculations has remained largely untapped. This study compares and contrasts two techniques for topic discovery: Latent Semanic Analysis (LSA) and Latent Dirichlet Allocation (LDA) in their topic frequency and order handling. Examining the different approaches of these methods for topic extraction will reflect on topic frequency and order, the research uncovers intriguing insights into their effectiveness in capturing the core meaning of text, showing each of strengths and weaknesses.

While topics provide valuable information about a document's core content, relying solely on them can limit systems ability to glean deeper insights. This is where the power of named entities (NEs) comes into play. These unique entities like people, organizations, or locations possess distinct features that can offer vital clues for differentiating between documents.

By incorporating NE features into any given text analysis, research can move beyond topic similarities and delve into the specific entities that shape the text's meaning. This allows us to identify subtle connections and nuanced differences that might remain hidden. Driven by the potential of Named Entity Recognition (NER) to enrich semantic similarity measures, this research investigated its capabilities more deeply. It is required to look beyond the common named entities and explore the potential of identifying less-frequent entities, hypothesizing that they could hold hidden semantic value. In order to begin this investigation, this work harnessed the power of a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model, leveraging its ability to understand the complex context of language. To further refine its NER ability, the BERT model is fine-tuned on the diverse and challenging texts of the Kaggle dataset. This targeted training allowed the model to improve its results in identifying even the most subtle named entities.

The methodological framework draws upon a diverse toolbox, leveraging the power of Python libraries and carefully tunes the parameters while training these models. This complex approach comprehensively analyses the relationship between independent variables and semantic similarity, paving the way for surprising discoveries and a potential paradigm shift in understanding text semantic relationships.

The results paint an astonishing picture, revealing that a select group of variables emerge as clear champions in their contribution to model inference for semantic similarity. Specifically, topic order, extracted from LSA, along with topic frequency and named entity occurrence, consistently outperforms all other variables across nearly all model configurations. This unexpected finding suggests that these features are crucial to unlocking the nuances of semantic relationships between texts. Furthermore, the overall performance of some models surpasses even established benchmarks of tools like Turnitin. Compared to the gold standard for semantic similarity, the proposed models have shown their effectiveness in accurately determining the true meaning and connections between texts.

This research disrupts pre-established beliefs, challenging the perceived infallibility of human ratings. Findings redefine previous knowledge and unlock practical applications, particularly in semantic plagiarism detection, question answering, and recommendation systems. Despite data limitations, the study marks a transformative step forward in semantic similarity analysis, embracing data-driven approaches to unlock new frontiers in understanding natural language.

Keywords: Semantic similarity, Natural language processing, Term Frequency, Latent Semantic Analysis, Latent Dirichlet Allocation, Named Entity Recognition, Topic frequency, Topic order, Named Entity Frequency, Named Entity order, BERT, Turnitin

TABLE OF CONTENTS

CHAPTER 1. SEMANTICS AND NATURAL LANGUAGE PROCESSING 2

1.1 Introduction 2

1.2 Natural Language Processing 2

1.3 Natural Language Processing: History 7

1.4 Rationale 9

1.5 Semantic Similarity 12

1.5.1 Key pillars of semantic similarities 12

1.5.2 Current approaches: Limitation & Challenge 13

1.6 Research Question 15

1.7 Objective 17

1.8 Theorizing the proposed work 20

1.9 Challenges in understanding Semantics: Unstructured Data 22

1.10 Scope 23

1.11 Meaningful Features 24

1.12 Conclusion 26

CHAPTER 2. PROPOSED APPROACH 30

2.1 Introduction 30

2.2 Limitation of using TF-IDF for similarity 30

2.2.1 Illustration dataset 31

2.2.2 Analysis 33

2.3 Order and Topic Intertwinement 33

2.4 Semantic Similarity: Deriving the Latent variables 35

2.5 Understanding the importance of variable and Architecture 41

2.6 Conclusion....................44

CHAPTER 3. Background work on Latent Features....................47

3.1 Introduction 47

3.2 Term Frequency and Inverse document Frequency (tf-idf) 47

3.3 Latent Semantic Analysis....................51

3.4 Latent Dirichlet Allocation....................56

3.5 Named Entity Recognition 70

B. Machine Learning Approaches:....................75

C. Deep Learning Approaches: 75

3.6 Co-reference Resolution....................76

3.7 Conclusion....................80

CHAPTER 4. VECTORIZING TERM FREQUENCY 82

4.1 Introduction 82

4.2 Term Frequency: Quantifying Word Importance....................82

4.2.1 Methodology 84

4.3 Conclusion....................96

CHAPTER 5. Vectorizing Topics From LATENT SEMANTIC ANALYSIS 98

5.1 Introduction 98

5.2 Methodology 98

5.3 Implementation ... 101

5.4 LSA: Content Similarity ... 102

5.5 Content Similarity: Implementation ... 109

5.6 Topic Order ... 113

5.7 LSA order similarity - Implementation: ... 126

5.8 Conclusion ... 127

CHAPTER 6. Vectorizing Topics From LATENT DIRICHLET ALLOCATION ... 129

6.1 Introduction ... 129

6.2 LDA: Content Similarity ... 129

6.2.1 Topic Word Labelling ... 130

6.2.2 Categorizing Document into Topics ... 131

6.2.3 Sliding Window Technique ... 131

6.2.4 Pseudo Code ... 132

6.2.5 Similarity Calculation ... 133

6.3 Content Similarity: Implementation ... 137

6.4 Topic Order ... 141

6.5 LDA order similarity - Implementation: ... 153

6.6 Conclusion ... 154

CHAPTER 7. Vectorizing NAMED ENTITY RECOGNITION ... 157

7.1 Introduction ... 157

7.2 Method: Coreference ... 157

7.3 Implementation: Coreference Resolution 158

7.4 Methodology: Named Entity Recognition 159

7.5 Implementation 161

7.6 Content Similarity: Named Entity 165

7.6.1 Extraction of NEs from the document 165

7.6.2 Similarity estimation 166

7.7 Order Similarity: Named Entities 170

7.7.1. Extract Named Entity Order 171

7.7.2. Order Similarity 171

7.8 Implementation: Content and order similarity NEs 186

7.8.1 NE: Content Similarity 187

7.8.2 NE: Order Similarity 188

7.9 Conclusion 188

CHAPTER 8. METHODOLOGY 190

8.1 Introduction 190

8.2 Exploring Text Similarity Using Parametric and Non-Parametric Machine Learning Models 190

8.3 Data Collection 192

8.4 Methodology 193

8.4.1 Linear Regression 196

8.4.2 Polynomial Regression 196

8.4.3 Random Forest 196

8.4.4 Implementation 196
8.5 Conclusion 199
CHAPTER 9. RESULTS 201
9.1 Introduction 201
9.2 Experimental Setup 201
9.3 Results 202
9.4 Linearity Assessment of Semantic Similarity Scores 203
9.4.1 Linearity Plot 203
9.4.2 Multi-collinearity 212
9.5 Linear Regression 214
9.5.1 Model Summary 215
9.5.2 Model Specification 215
9.5.3 Plots 220
9.6 Polynomial Regression 229
9.6.1 Errors 229
9.6.2 Model Fit 230
9.6.3 Model Complexity: AIC and BIC 230
9.6.4 Diagnostic Plots 231
9.7 Decision Tree 237
9.7.1 Error 237
9.7.2 Model Fit 238

9.7.3 Diagnostic Plots of the Decision Tree 240

9.8 Random Forest 245

9.8.1 Error 245

9.8.2 Model Fit....247

9.8.3 Diagnostic plot of Random Forest 247

9.9 Neural Network 252

9.9.1 Error 252

9.9.2 Model Fit....253

9.9.3 Learning Plot....254

9.10 Conclusion....258

CHAPTER 10. DISCUSSION....262

10.1 Introduction 262

10.2 Discussion 262

10.2.1 Linearity Assessment 262

10.2.2 Multicollinearity 266

10.3 Linear Regression....266

10.3.1 Feature Coefficients 266

10.3.2 Model Fitting 268

10.3.3 Residual Normality 269

10.4 Polynomial Regression....270

10.4.1 Error 270

10.4.2 Model Fit......270

10.4.3 Residual Distribution......271

10.5 Decision Tree......271

10.5.1 Error......272

10.5.2 Model Fit......272

10.5.3 Model Plot......273

10.5.4 Feature importance......273

10.6 Random Forest......274

10.6.1 Error......274

10.6.2 Model Fit......274

10.6.3 Feature Importance......275

10.7 Neural Network......276

10.7.1 Error......276

10.7.2 Model Fit......276

10.7.3 Feature Importance......277

10.8 Inference: Linear Regression, Polynomial Regression, Decision Tree, Random Forest, Neural Network......278

10.9 Comparison with Semantic Similarity based Techniques......282

10.9.1 Word Net Semantic Similarity......282

10.9.2 Numerical Embedding based Semantic Similarity......283

10.9.3 Statistical Feature Based Techniques......284

10.10 Comparison of Proposed work with Plagiarism Software on Plagiarised documents 285
10.10.1 Data 286
10.10.2 Experiment Setup 286
10.10.3 Methodology 287
10.11 Conclusion 290
CHAPTER 11. IMPLICATIONS AND LIMITATIONS 293
11.1 Introduction 293
11.2 Theoretical Implication 293
11.2.1 Breaking Myth on Human Rating 293
11.2.2 Additional Latent Feature Extraction 294
11.2.3 Topic based features 294
11.2.4 Order features 295
11.2.5 Most significant Features 295
11.2.6 Comparison between LSA and LDA features 296
11.2.7 Linear and Non-linear approach 297
11.3 Practical Implementation 297
11.3.1 Information Retrieval 297
11.3.2 Plagiarism Detection 298
11.3.3 Recommendation System 298
11.3.4 Question Answering Systems 298
11.3.5 Answersheet Evaluation 298

11.4 Legal and Policy Implications 299

11.5 Limitation 301

11.6 Future Work 302

11.7 Conclusion 304

APPENDIX 1 343

APPENDIX 2 362

Appendix 2.1 Pseudo Code Latent Dirichlet Allocation 362

APPENDIX 3 364

APPENDIX 4 367

APPENDIX 5 372

APPENDIX 6 376

APPENDIX 7 Gold Standard 378

Table of Tables

Table 2. 1 TF-IDF similarity between the documents (storyA, storyB, and storyC) 33

Table 2. 2 Topic frequency count in documents (StoryA, storyB, storyC) 43

Table 2. 3 Topic count similarity between documents (storyA, storyB, and storyC).............43

Table 2. 4 Topic order similarity between documents (storyA, storyB, and storyC).............43

Table 2. 5 NE Frequency similarity between documents 44

Table 2. 6 NE order similarity between documents..44

Table 3. 1 Transformation from original attributes to constructed attributes..........................51

Table 3. 2 Word document vector matrix 52

Table 3. 3 Probability of each word to the topics 57

Table 3. 4 Example: named entities...70

Table 4. 1 Words and its frequency and normalised term frequency Document A.................90

Table 4. 2 Words and its frequency and normalised term frequency Document B 91

Table 4. 3 Common words and their respective normalized frequencies from Document A, and B...92

Table 4. 4 Square of the normalized frequency for both the documents................................93

Table 4. 5 Square normalized frequency 94

Table 4. 6 Multiplication of common word's and normalized frequencies 95

Table 4. 7 Measurements for cosine similarity..95

Table 4. 8 Sum multiplication of normalized frequencies...95

Table 4. 9 tf cosine similarity 96

Table 5. 1 Topic and its corresponding topic count..107

Table 5. 2 Square topic count 107

Table 5. 3 Sum the square of topic counts...108

Table 5. 4 Square root of the sum of squares...108

Table 5. 5 Product of the sum 109

Table 5. 6 Topic flow in Document A 115

Table 5. 7 Topic flow in Document B 116
Table 5. 8 Topic index in Document A....... 117
Table 5. 9 Topic index in Document B....... 117
Table 5. 10 Topic index in Document A, and B 119
Table 5. 11 Square Topic index in Document A 120
Table 5. 12 Square Topic index in Document A 121
Table 5. 13 Sum Square Topic index in Document A 122
Table 5. 14 Sum Square Topic index in Document A, ... 122
Table 5. 15 Square root of sum of square Topic index in Document A 123
Table 5. 16 Square root of sum of square Topic index in Document B 124
Table 5. 17 Product Topic index from the Document A, and B 125
Table 5. 18 Sum of Product Topic index from the Document A, and B 125

Table 6. 1 Document's A, and B topic count....... 134
Table 6. 2 Square of frequency....... 135
Table 6. 3 Sum of square for document A, and B....... 135
Table 6. 4 Square root of the sum of squares for documents A and B 136
Table 6. 5 Product of the sum from the count of topic for documents A and B....... 136
Table 6. 6 Topic flow in Document A 143
Table 6. 7 Topic flow in Document B 143
Table 6. 8 Positional index of Document A....... 144
Table 6. 9 Positional index of Document B....... 145
Table 6. 10 Positional index of Document A, and B 147
Table 6. 11 Square positional index of Document A....... 148
Table 6. 12 Square positional index of Document B....... 149
Table 6. 13 Sum of square positional index of Document A....... 149
Table 6. 14 Sum of square positional index of Document A....... 150
Table 6. 15 Square root of sum of square of the index values (Document A)....... 151
Table 6. 16 Square root of sum of square of the index values (Document B)....... 151
Table 6. 17 Product positional index values....... 152
Table 6. 18 Sum of product positional index values....... 153

Table 7. 1 Frequency of named entities from text 1, and 2 168
Table 7. 2 Common Named Entities and their frequency from text 1, and 2 168

Table 7. 3 Square of common NE frequency from text 1, and 2 169
Table 7. 4 Product of common NE frequency from text 1, and 2 169
Table 7. 5 Sum of Square of common NE frequency from text 1, and 2 170
Table 7. 6 Sum of product of common NE frequency from text 1, and 2 170
Table 7. 7 List of NE's from text 1, and 2 173
Table 7. 8 Position of NE's from text 1 177
Table 7. 9 Position of NE's from text 2 181
Table 7. 10 Position of Common NE's from text 1, and 2 181
Table 7. 11 Ascending order Common NE's position from text 1, and 2 182
Table 7. 12 Reorder Common NE's position from text 1, and 2 183
Table 7. 13 Square of NE's position from text 1, and 2 183
Table 7. 14 Union of Common and uncommon NE's and their position from text 1, and 2. 184
Table 7. 15 Alphabetically arrange Union of Common and uncommon NE's and their position from text 1, and 2 185
Table 7. 16 Product of order index from text 1, and 2 185
Table 7. 17 Square root of sum of square of order 186
Table 7. 18 Sum of product of order of NEs 186

Table 9. 1 Required hardware resources 202
Table 9. 2 Required software resources 202

Table 10. 1 Errors measurements in different models 279
Table 10. 2 Model fit for all the models 280
Table 10. 3 Normal Distribution of the Residuals 281
Table 10. 4 Coefficient of each feature in each model 282
Table 10. 5 Correlation between model rating and human rating 285
Table 10. 6 Comparison between different models with Turnitin 288

Table of Figures

Figure 1. 1 Core task of NLP (Gudivada, 2018)..4

Figure 3. 1 Latent Semantic Analysis (SVD decomposition)...53
Figure 3. 2 Adopted model of LDA (Blei *et al.*, 2003) 65
Figure 3. 3 Co-reference demonstration 77

Figure 5. 1 Sliding window demonstration 1..104
Figure 5. 2 Sliding window demonstration 2..105

Figure 6. 1 Sliding window demonstration 1..132
Figure 6. 2 Sliding window demonstration 2..132

Figure 9. 1 Scatter plot lda_content vs. similarity score...204
Figure 9. 2 Scatter plot lda_order vs. similarity score 205
Figure 9. 3 Scatter plot lsa_content vs. similarity score 205
Figure 9. 4 Scatter plot lsa_order vs. similarity score..206
Figure 9. 5 Scatter plot ne_content vs. similarity score...206
Figure 9. 6 Scatter plot ne_order vs. similarity score 207
Figure 9. 7 Scatter plot tf_score vs. similarity score..207
Figure 9. 8 Partial regression plot lda_content vs Similarity Score.....................................209
Figure 9. 9 Partial regression plot lda_order vs Similarity Score 209
Figure 9. 10 Partial regression plot lsa_order vs Similarity Score 210
Figure 9. 11 Partial regression plot lsa_order vs Similarity Score 210
Figure 9. 12 Partial regression plot ne_content vs Similarity Score.....................................211
Figure 9. 13 Partial regression plot ne_order vs Similarity Score 211

Figure 9. 14 Partial regression plot tf_score vs Similarity Score 212
Figure 9. 15 Heat map of correlation matrix between independent variables 214
Figure 9. 16 A screen shot of the Model Summary 215
Figure 9. 17 Scattered plot of residual vs. fitted value 221
Figure 9. 18 Q-Q plot of observed vs theoretical quantiles 222
Figure 9. 19 A frequency distribution of the residuals 223
Figure 9. 20 Scatter plot of Leverage vs normalised residual...225
Figure 9. 21 Cook's distance on outlier present in the observation.......................................226
Figure 9. 22 Scatter plot of square root of absolute residuals vs fitted values 227
Figure 9. 23 Scatter plot of independent variables vs residual 228
Figure 9. 24 Model fit: actual vs predicted value (degree 3) 232
Figure 9. 25 Model fit: actual vs predicted value (degree 2) 233
Figure 9. 26 Scatter plot Predicted value vs Residual (degree 3) 233
Figure 9. 27 Scatter plot Predicted value vs Residual (degree 2) 233
Figure 9. 28 Q-Q plot Theoretical vs observed quantiles (degree 2)....................................234
Figure 9. 29 Q-Q plot Theoretical vs observed quantiles (degree 2)....................................234
Figure 9. 30 Frequency distribution of the residual (degree 2)...235
Figure 9. 31 Frequency distribution of the residual (degree 3)...236
Figure 9. 32 Cook's distance plot 237
Figure 9. 33 Learning curve of training and cross validation MSE.......................................239
Figure 9. 34 Model fit between actual and predicted values 241
Figure 9. 35 Residual vs predicted value plot...242
Figure 9. 36 Q-Q plot between observed and theoretical quantiles.......................................243
Figure 9. 37 Feature importance bar graph...244
Figure 9. 38 Frequency distribution of the residual..245
Figure 9. 39 Learning curve to compare training MSE, and cross-validation MSE..............247
Figure 9. 40 Model fit between the actual and predicted values 248
Figure 9. 41 Plot of residual vs predicted values..249
Figure 9. 42 Frequency distribution of the residual..250
Figure 9. 43 Q-Q plot between observed and theoretical quantiles.......................................251
Figure 9. 44 Bar Plot for important independent features 252
Figure 9. 45 Learning curve comparing between training and validation loss......................254
Figure 9. 46 Model fit between predicted and actual values 255
Figure 9. 47 Scattered plot of residual vs predicted values 256

Figure 9. 48 Q-Q plot of theoretical vs observed values 257
Figure 9. 49 Bar plot of frequency of residual and its density..258
Figure 9. 50 Bar plot for shows importance of features in prediction 259
Figure 9. 51 SHAP plot for feature importance at observation level....................................259

Figure 10. 1 Distance between human rating and proposed work on plagiarised documents 289
Figure 10. 2 Average absolute difference between human rating and other methods 290

Table of Equations

(3. 1)..53
(3. 2)..59
(3. 3)..64
(3. 4)..64
(3. 5)..64
(3. 6)..65
(3. 7)..66

(5. 1)..109

(6. 1)..136

Appendix Eq. 1. 1 359
Appendix Eq. 1. 2 359
Appendix Eq. 1. 3 359
Appendix Eq. 1. 4 359
Appendix Eq. 1. 5 359

Table of Algorithms

Algorithm 3. 1 Algorithm Latent Dirichlet Allocation 363

Algorithm 4. 1 Term frequency 364
Algorithm 4. 2 Cosine similarity with term frequency 366

Algorithm 5. 1 Modified Latent Semantic Analysis 367
Algorithm 5. 2 Labelling words with the topics 367
Algorithm 5. 3 Topic allocation to each sliding window 368
Algorithm 5. 4 Count of each topic in the document 369
Algorithm 5. 5 Cosine similarity based on each topic count 370
Algorithm 5. 6 Indexing all windows of document 370
Algorithm 5. 7 Creating a shared dataframe for topics extracted from document A, and document B 371

Algorithm 6. 1 Labelling topics to each words in the document 372
Algorithm 6. 2 Assigning topic name to each window 373
Algorithm 6. 3 Counting different topics assign to each window 373
Algorithm 6. 4 Cosine similarity between two documents based on topic count frequency (LDA) 374
Algorithm 6. 5 Creating index of each window and assign topic (LDA) 375
Algorithm 6. 6 Merging index file from document A, and document B (LDA) 375

Algorithm 8. 1 Pseudo for the model training for drawing the inference 377

proposed method incorporates diverse parameters for document retrieval, enhancing relevance to the query. Future research could explore extending this model within the realm of information retrieval, evaluating its efficacy in comparison to established methodologies.

f) Conversational AI

This dissertation did not explore the application of ontologies. However, integrating ontologies with the proposed method holds promise for significant improvement. Ontologies provide a structured schema for classifying search results into categories. This allows users to refine their search by selecting relevant categories, particularly beneficial for domain-specific documents like marketing materials or medical records. Additionally, ontologies can enrich user queries by suggesting related terms or concepts. They can then assign relevance scores to search results based on the semantic relationships between the query and the documents.

11.7 Conclusion

Our experiments have shattered some established beliefs in the field of semantic similarity. For example, the "gold standard" of human rating has been shown to have limitations, opening doors to alternative assessment methods. Additionally, while previous research on LDA has received some experimental support, LSA features produced different results, suggesting deeper intricacies in topic extraction.

These findings unlock exciting possibilities for practical applications. The newly-developed variables enable semantic plagiarism detection, allowing for more nuanced identification of content theft. Furthermore, the potential extends to diverse areas like question answering and recommendation systems, where understanding deeper meaning is crucial.

However, data remains a major obstacle. The current sample size needs significant expansion for robust NLP tasks. Natural language processing requires vast amounts of data, and pronoun resolution further complicates matters. While the pre-trained allennlp model performs well, fine-tuning specifically for co-reference resolution will significantly improve accuracy.

Overall, this research paves the way for transformative advancements in semantic similarity analysis. By addressing the limitations of traditional methods and embracing data-driven approaches, we can unlock new frontiers in understanding and utilizing the richness of natural language.

REFERENCE

A quote by Salvador Dalí. (n.d.). https://www.goodreads.com/quotes/18874-have-no-fear-of-perfection---you-ll-never-reach-it

Abdi, H. (2007). Singular value decomposition (SVD) and generalized singular value decomposition. *Encyclopedia of measurement and statistics*, *907*(912), 44.

Ades, A. E., & Steedman, M. J. (1982). On the order of words. *Linguistics and philosophy*, *4*, 517-558.

Admin, & Admin. (2024, May 16). The case of copyright infringement in the use of training AI. IIPRD |. https://www.iiprd.com/the-case-of-copyright-infringement-in-the-use-of-training-artificial-intelligence-vis-a-vis-the-positions-in-india-us-a-critical-analysis/#:~:text=A%20latest%20update%20around%20these,to%20work%20for%20commercial%20motives.

Ahmad, F., & Faisal, M. (2022). A novel hybrid methodology for computing semantic similarity between sentences through various word senses. *International Journal of Cognitive Computing in Engineering*, *3*, 58-77.

Ahmed, H. (2017). *Detecting opinion spam and fake news using n-gram analysis and semantic similarity* (Doctoral dissertation).

Ajinaja, M. O., Adetunmbi, A. O., Ugwu, C. C., & Popoola, O. S. (2023). Semantic similarity measure for topic modeling using latent Dirichlet allocation and collapsed Gibbs sampling. *Iran Journal of Computer Science*, *6*(1), 81-94.

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika*, *60*(2), 255-265.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*, 317-332.

Akbik, A., Bergmann, T., & Vollgraf, R. (2019, June). Pooled contextualized embeddings for named entity recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 724-728).

Albert, A., & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, *71*(1), 1-10.

Amigó, E., Fang, H., Mizzaro, S., & Zhai, C. (2018, June). Are we on the right track? An examination of information retrieval methodologies. In The *41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 997-1000).

Anne, C., Mishra, A., Hoque, M. T., & Tu, S. (2018). Multiclass patent document classification. *Artif. Intell. Res.*, *7*(1), 1-14.

Arabi, H., & Akbari, M. (2022). Improving plagiarism detection in text document using hybrid weighted similarity. *Expert Systems with Applications*, *207*, 118034.

Archive Articles from Web. (n.d.). The Hindu. https://www.thehindu.com/archive/

Armstrong, M. (August 6, 2021). How Many Websites Are There? [Digital image]. Retrieved December 30, 2022, from https://www.statista.com/chart/19058/number-of-websites-online/

Auret, L., & Aldrich, C. (2012). Interpretation of nonlinear relationships between process variables by use of random forests. *Minerals Engineering*, *35*, 27-42.

Awasthi, S. (2019). Plagiarism and academic misconduct: A systematic review. *DESIDOC Journal of Library & Information Technology*, *39*(2).

Bah, M. J., & Wang, H. (2020). A Parametric and Non-Parametric Approach for High-Accurate Outlier Detection. *Journal of Information Science & Engineering*, *36*(2).

Balk, B., & Elder, K. (2000). Combining binary decision tree and geostatistical methods to estimate snow distribution in a mountain watershed. *Water Resources Research*, *36*(1), 13-26.

Barde, B. V., & Bainwad, A. M. (2017, June). An overview of topic modeling methods and tools. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 745-750). IEEE.

Bastani, K., Namavari, H., & Shaffer, J. (2019). Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. *Expert Systems with Applications*, *127*, 256-271.

Beheshti, S. M. R., Benatallah, B., Venugopal, S., Ryu, S. H., Motahari-Nezhad, H. R., & Wang, W. (2017). A systematic review and comparative analysis of cross-document coreference resolution methods and tools. *Computing, 99,* 313-349.

Bellaouar, S., Bellaouar, M. M., & Ghada, I. E. (2021, February). Topic modeling: Comparison of LSA and LDA on scientific publications. In *2021 4th international conference on data storage and data engineering* (pp. 59-64).

Belli, S., Raventós, C. L., & Guarda, T. (2020, January). Plagiarism detection in the classroom: Honesty and trust through the Urkund and Turnitin software. In *International Conference on Information Technology & Systems* (pp. 660-668). Cham: Springer International Publishing.

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Ben Ishak, A. (2016). Variable selection using support vector regression and random forests: A comparative study. *Intelligent Data Analysis*, *20*(1), 83-104.

Bengtson, E., & Roth, D. (2008, October). Understanding the value of features for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 294-303).

Berragan, C., Singleton, A., Calafiore, A., & Morley, J. (2023). Transformer based named entity recognition for place name extraction from unstructured text. *International Journal of Geographical Information Science, 37(4),* 747-766.

Bhattacharjee, S., Haque, R., de Buy Wenniger, G. M., & Way, A. (2020, June). Investigating query expansion and coreference resolution in question answering on BERT. In *International conference on applications of natural language to information systems* (pp. 47-59). Cham: Springer International Publishing.

Bikel, D. M., Schwartz, R., & Weischedel, R. M. (1999). An algorithm that learns what's in a name. *Machine learning, 34,* 211-231.

Birkle, C., Pendlebury, D. A., Schnell, J., & Adams, J. (2020). Web of Science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies*, *1*(1), 363-376.

Blanco, E., & Moldovan, D. (2015). A semantic logic-based approach to determine textual similarity. *Ieee/acm transactions on audio, speech, and language processing*, *23*(4), 683-693.

Blei, D. M., & Lafferty, J. D. (2006, June). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning* (pp. 113-120).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research,* 3(Jan), 993-1022.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3*(Jan), 993-1022.

Bliss, R. L., Katz, J. N., Wright, E. A., & Losina, E. (2012). Estimating proximity to care: are straight line and zipcode centroid distances acceptable proxy measures?. *Medical care*, *50*(1), 99.

Bøhn, C., & Nørvåg, K. (2010, April). Extracting named entities and synonyms from wikipedia. In *2010 24th IEEE International Conference on Advanced Information Networking and Applications* (pp. 1300-1307). IEEE.

Bollegala, D., Matsuo, Y., & Ishizuka, M. (2007). Measuring semantic similarity between words using web search engines. *www*, *7*(2007), 757-766.

Bontcheva, K., Dimitrov, M., Maynard, D., Tablan, V., & Cunningham, H. (2002, June). Shallow methods for named entity coreference resolution. In *Chaınes de références et résolveurs d'anaphores, workshop TALN* (pp. 24-27).

Borah, A., Barman, M. P., & Awekar, A. (2021, August). Are word embedding methods stable and should we care about it?. In *Proceedings of the 32nd ACM Conference on Hypertext and social media* (pp. 45-55).

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology, 3(2)*, 77-101.

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, *30*(1-7), 107-117.

Bruton, S., & Childers, D. (2016). The ethics and politics of policing plagiarism: A qualitative study of faculty views on student plagiarism and Turnitin®. *Assessment & Evaluation in Higher Education*, *41*(2), 316-330.

Budanitsky, A. (1999). *Lexical semantic relatedness and its application in natural language processing*. technical report CSRG-390, Department of Computer Science, University of Toronto.

Burgess, C., & Lund, K. (1995). Hyperspace analog to language (hal): A general model of semantic representation. In *Proceedings of the annual meeting of the Psychonomic Society* (Vol. 12, pp. 177-210).

Cai, D., Mei, Q., Han, J., & Zhai, C. (2008, October). Modeling hidden topics on document manifold. In *Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 911-920).

Cai, Y., Zhang, Q., Lu, W., & Che, X. (2018). A hybrid approach for measuring semantic similarity based on IC-weighted path distance in WordNet. *Journal of intelligent information systems*, *51*, 23-47.

Celikyilmaz, A., Hakkani-Tur, D., & Tür, G. (2010, June). LDA based similarity modeling for question answering. In *Proceedings of the NAACL HLT 2010 Workshop on Semantic Search* (pp. 1-9).

Chandrasekaran, D., & Mago, V. (2021). Evolution of semantic similarity—a survey. *ACM Computing Surveys (CSUR)*, *54*(2), 1-37.

Chandrasekaran, D., & Mago, V. (2021). Evolution of semantic similarity—a survey. *ACM Computing Surveys (CSUR)*, *54*(2), 1-37.

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in cognitive sciences*, *10*(7), 335-344.

Chen, L. C. (2017). An effective LDA-based time topic model to improve blog search performance. *Information Processing & Management*, *53*(6), 1299-1319.

Cheng, H., Shen, Y., Liu, X., He, P., Chen, W., & Gao, J. (2021). UnitedQA: A hybrid approach for open domain question answering. *arXiv preprint arXiv:2101.00178*.

Chomsky, N. (1966). Explanatory models in linguistics. In *Studies in Logic and the Foundations of Mathematics* (Vol. 44, pp. 528-550). Elsevier.

Chomsky, N. (1998). Some Observations on Economy in Generative Grammar Noam Chomsky, Massachusetts Institute of Technology. *Is the Best Good Enough?: Optimality and competition in syntax*, 115.

Chowdhary, K., & Chowdhary, K. R. (2020). Natural language processing. *Fundamentals of artificial intelligence*, 603-649.

Christian, H., Agus, M. P., & Suhartono, D. (2016). Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF). *ComTech: Computer, Mathematics and Engineering Applications, 7(4)*, 285-294.

Clark, K., & Manning, C. D. (2016). Deep reinforcement learning for mention-ranking coreference models. *arXiv preprint arXiv:1609.08667*.

Costa, S. D., Barcellos, M. P., & Falbo, R. D. A. (2021). Ontologies in human–computer interaction: A systematic literature review. *Applied Ontology*, *16*(4), 421-452.

Cox, N. (2008). Copyright in Primary Legal Materials in Common Law Jurisdiction. *Australian Intellectual Property Journal, Forthcoming*.

Croft, W. B., Turtle, H. R., & Lewis, D. D. (1991, September). The use of phrases and structured queries in information retrieval. In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 32-45).

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior research methods*, *51*, 14-27.

Crossno, P. J., Wilson, A. T., Shead, T. M., & Dunlavy, D. M. (2011, November). Topicview: Visually comparing topic models of text collections. In *2011 ieee 23rd international conference on tools with artificial intelligence* (pp. 936-943). IEEE.

Cutrone, L., & Chang, M. (2011, July). Auto-assessor: computerized assessment system for marking student's short-answers automatically. In *2011 IEEE International Conference on Technology for Education* (pp. 81-88). IEEE.

Cvitanic, T., Lee, B., Song, H. I., Fu, K., & Rosen, D. (2016, January). LDA v. LSA: A comparison of two computational text analysis tools for the functional categorization of patents. In *International Conference on Case-Based Reasoning*.

Damessie, T. T., Scholer, F., & Culpepper, J. S. (2016, December). The influence of topic difficulty, relevance level, and document ordering on relevance judging. In *Proceedings of the 21st Australasian Document Computing Symposium* (pp. 41-48).

Damessie, T. T., Scholer, F., Järvelin, K., & Culpepper, J. S. (2016, September). The effect of document order and topic difficulty on assessor agreement. In Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval (pp. 73-76).

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391-407.

Deloria, V. (2020). Research, redskins, and reality. In *American nations* (pp. 458-467). Routledge.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dogra, V., Verma, S., Chatterjee, P., Shafi, J., Choi, J., & Ijaz, M. F. (2022). A complete process of text classification system using state-of-the-art NLP models. Computational Intelligence and Neuroscience, 2022.

dslim/bert-large-NER · Hugging Face. (2001, January 18). https://huggingface.co/dslim/bert-large-NER?text=The+NER+model+uses+the+pre-trained+BERT%2C+available+on+the+hugging+face.+The+BERT+model+needs+fine-tuning+before+using+for+the+NER.+BERT+is+trained+on+the+Kaggle+dataset.

Du, L., Buntine, W., Jin, H., & Chen, C. (2012). Sequential latent Dirichlet allocation. *Knowledge and information systems*, *31*, 475-503.

Dufour, J. M., & Dagenais, M. G. (1985). Durbin-Watson tests for serial correlation in regressions with missing observations. *Journal of Econometrics*, *27*(3), 371-381.

Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology (ARIST), 38,* 189-230.

Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, *297*(3), 1178-1192.

East, J. (2006). The problem of plagiarism in academic culture. *International Journal for Educational Integrity*, *2*(2).

Ebeling, W., & Nicolis, G. (1992). Word frequency and entropy of symbolic sequences: a dynamical perspective. *Chaos, Solitons & Fractals*, *2*(6), 635-650.

Egger, R. (2022). Topic Modelling: Modelling Hidden Semantic Structures in Textual Data. In *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications* (pp. 375-403). Cham: Springer International Publishing.

Eisenstein, J. (2019). *Introduction to natural language processing*. MIT press.

Eret, E., & Gokmenoglu, T. (2010). Plagiarism in higher education: A case study with prospective academicians. *Procedia-Social and Behavioral Sciences*, *2*(2), 3303-3307.

Fabri, R., & Borg, A. (2002). Topic, focus and word order in Maltese. AY *et al.*, ed., Aspects of Dialects of Arabic Today, 354-363.

Farahat, A. K., & Kamel, M. S. (2011). Statistical semantics for enhancing document clustering. *Knowledge and information systems, 28,* 365-393.

Farahian, M., Parhamnia, F., & Avarzamani, F. (2020). Plagiarism in theses: A nationwide concern from the perspective of university instructors. *Cogent Social Sciences*, *6*(1), 1751532.

Ferreira, R., Lins, R. D., Simske, S. J., Freitas, F., & Riss, M. (2016). Assessing sentence similarity through lexical, syntactic and semantic analysis. *Computer Speech & Language*, *39*, 1-28.

Finkel, J. R., Grenager, T., & Manning, C. D. (2005, June). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)* (pp. 363-370).

Fodeh, S., Punch, B., & Tan, P. N. (2011). On ontology-driven document clustering using core semantic features. *Knowledge and information systems, 28,* 395-421.

Foltz, P. W. (2007). Discourse coherence and LSA. *Handbook of latent semantic analysis*, *167*, 184.

Fothergill, R., Cook, P., & Baldwin, T. (2016, May). Evaluating a topic modelling approach to measuring corpus similarity. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 273-279).

Friburger, N., Maurel, D., & Giacometti, A. (2002, August). Textual similarity based on proper names. In *Proc. of the workshop Mathematical/Formal Methods in Information Retrieval* (pp. 155-167).

Garbhapu, V., & Bodapati, P. (2020). A comparative analysis of Latent Semantic analysis and Latent Dirichlet allocation topic modeling methods using Bible data. *Indian Journal of Science and Technology*, *13*(44), 4474-4482.

Ghahramani, Z. (2013). Bayesian non-parametrics and the probabilistic approach to modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 371*(1984), 20110553.

Gokul, A. (2023). Llms and ai: Understanding its reach and impact.

Gorrell, G. (2006, April). Generalized Hebbian algorithm for incremental singular value decomposition in natural language processing. In *11th conference of the European chapter of the association for computational linguistics* (pp. 97-104).

Gorrell, G., Petrak, J., & Bontcheva, K. (2015). Using@ Twitter conventions to improve# LOD-based named entity disambiguation. In *The Semantic Web. Latest Advances and New Domains: 12th European Semantic Web Conference, ESWC 2015, Portoroz, Slovenia, May 31--June 4, 2015. Proceedings 12* (pp. 171-186). Springer International Publishing.

Goyal, A., Gupta, V., & Kumar, M. (2018). Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review, 29*, 21-43.

Grant-Muller, S. M., Gal-Tzur, A., Minkov, E., Nocera, S., Kuflik, T., & Shoor, I. (2015). Enhancing transport data collection through social media sources: methods, challenges and opportunities for textual data. *IET Intelligent Transport Systems*, *9*(4), 407-417.

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. Universals of language, 2, 73-113.

Gregorich, M., Strohmaier, S., Dunkler, D., & Heinze, G. (2021). Regression with highly correlated predictors: variable omission is not the solution. *International journal of environmental research and public health*, *18*(8), 4259.

Grishman, R., & Sundheim, B. M. (1996). Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Gu, Y., Qu, X., Wang, Z., Zheng, Y., Huai, B., & Yuan, N. J. (2022). Delving deep into regularity: a simple but effective method for Chinese named entity recognition. *arXiv preprint arXiv:2204.05544.*

Gudivada, V. N., Rao, D. L., & Gudivada, A. R. (2018). Information retrieval: concepts, models, and systems. In *Handbook of statistics* (Vol. 38, pp. 331-401). Elsevier.

Gupta, A., Kumar, A., Gautam, J., Gupta, A., Kumar, M. A., & Gautam, J. (2017). A survey on semantic similarity measures. *IJIRST-International Journal for Innovative Research in Science & Technology*, *3*(12).

Hadi, M. A., Aruldhas, J., Chow, L. F., & Wattleworth, J. A. (1995). Estimating safety effects of cross-section design for various highway types using negative binomial regression. *Transportation Research Record*, *1500*, 169.

Han, L., Martineau, J., Cheng, D., & Thomas, C. (2015, June). Samsung: Align-and-differentiate approach to semantic textual similarity. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (pp. 172-177).

Han, M., Zhang, X., Yuan, X., Jiang, J., Yun, W., & Gao, C. (2021). A survey on the techniques, applications, and performance of short text semantic similarity. *Concurrency and Computation: Practice and Experience*, *33*(5), e5971.

Hänig, C., Remus, R., & De La Puente, X. (2015, June). Exb themis: Extensive feature extraction from word alignments for semantic textual similarity. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)* (pp. 264-268).

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients.

Herring, S. C. (1990, August). Information structure as a consequence of word order type. In Annual Meeting of the Berkeley Linguistics Society (Vol. 16, No. 1, pp. 163-174).

Ho, C., Azmi Murad, M. A., Doraisamy, S., & Abdul Kadir, R. (2014). Extracting lexical and phrasal paraphrases: a review of the literature. *Artificial Intelligence Review*, *42*, 851-894.

Hobbs, J. R. (1977). Pronoun resolution. *ACM SIGART Bulletin*, (61), 28-28.

Hoblos, J. (2020, December). Experimenting with latent semantic analysis and latent dirichlet allocation on automated essay grading. In *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)* (pp. 1-7). IEEE.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, *42*, 177-196.

Hofstätter, S., Rekabsaz, N., Eickhoff, C., & Hanbury, A. (2019, July). On the effect of low-frequency terms on neural-IR models. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1137-1140).

Hong, Z., Ward, L., Chard, K., Blaiszik, B., & Foster, I. (2021). Challenges and advances in information extraction from scientific literature: a review. *JOM*, *73*(11), 3383-3400.

Hossain, D. M. (2011). Qualitative research process. *Postmodern Openings*, *2*(7), 143-156.

HS, C., & Shenoy, M. K. (2020). Advanced text documents information retrieval system for search services. *Cogent Engineering*, *7*(1), 1856467.

Humerick, M. (2017). Taking AI personally: how the EU must learn to balance the interests of personal data privacy & artificial intelligence. *Santa Clara High Tech. LJ*, *34*, 393.

Ibrahim, O. A. S., & Landa-Silva, D. (2016). Term frequency with average term occurrences for textual information retrieval. *Soft Computing*, *20*, 3045-3061.

Indolia, S., Goswami, A. K., Mishra, S. P., & Asopa, P. (2018). Conceptual understanding of convolutional neural network-a deep learning approach. *Procedia computer science*, *132*, 679-688.

Iosif, E., & Potamianos, A. (2009). Unsupervised semantic similarity computation between terms using web documents. *IEEE Transactions on knowledge and data engineering*, *22*(11), 1637-1647.

Iriberri, A., & Leroy, G. (2007, August). Natural language processing and e-government: Extracting reusable crime report information. In 2007 IEEE International Conference on Information Reuse and Integration (pp. 221-226). IEEE.

Ise, O. A. (2016). Integration and analysis of unstructured data for decision making: Text analytics approach. *International Journal of Open Information Technologies*, *4*(10), 82-88.

Islam, A., & Inkpen, D. (2008). Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *2*(2), 1-25.

Ivanov, A. A., & Holtzer, S. V. (2021). A modified algorithm of the latent semantic analysis for text processing in the Russian language. In *Journal of Physics: Conference Series* (Vol. 1715, No. 1, p. 012009). IOP Publishing.

Jain, S., Seeja, K. R., & Jindal, R. (2020). A new methodology for computing semantic relatedness: modified latent semantic analysis by fuzzy formal concept analysis. *Procedia Computer Science*, *167*, 1102-1109.

Jiang, C., Nian, Z., Guo, K., Chu, S., Zhao, Y., Shen, L., & Tu, K. (2020, November). Learning numeral embedding. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 2586-2599).

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., & Levy, O. (2020). Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics, 8,* 64-77.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., & Levy, O. (2020). Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics*, *8,* 64-77.

Joshi, M., Levy, O., Weld, D. S., & Zettlemoyer, L. (2019). BERT for coreference resolution: Baselines and analysis. *arXiv preprint arXiv:1908.09091.*

Joshi, M., Wang, H., & McClean, S. (2018). Dense semantic graph and its application in single document summarisation. *Emerging ideas on information filtering and retrieval: DART 2013: Revised and invited papers*, 55-67.

Kabbaj, A., Moulin, B., Gancef, J., Nadeau, D., & Rouleau, O. (2001). Uses, improvements, and extensions of Prolog+ CG: Case studies. In *Conceptual Structures: Broadening the Base:*

9th International Conference on Conceptual Structures, ICCS 2001 Stanford, CA, USA, July 30–August 3, 2001 Proceedings 9 (pp. 346-359). Springer Berlin Heidelberg.

Kabbaj, A., Moulin, B., Gancef, J., Nadeau, D., & Rouleau, O. (2001). Uses, improvements, and extensions of Prolog+ CG: Case studies. In *Conceptual Structures: Broadening the Base: 9th International Conference on Conceptual Structures, ICCS 2001 Stanford, CA, USA, July 30–August 3, 2001 Proceedings 9* (pp. 346-359). Springer Berlin Heidelberg.

Kadhim, A. I. (2019, April). Term weighting for feature extraction on Twitter: A comparison between BM25 and TF-IDF. In 2019 international conference on advanced science and engineering (ICOASE) (pp. 124-128). IEEE.

Kakkonen, T., Myller, N., Timonen, J., & Sutinen, E. (2005, June). Automatic essay grading with probabilistic latent semantic analysis. In *Proceedings of the second workshop on Building Educational Applications Using NLP* (pp. 29-36).

Kalepalli, Y., Tasneem, S., Teja, P. D. P., & Manne, S. (2020, May). Effective comparison of LDA with LSA for topic modelling. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1245-1250). IEEE.

Kalia, A., Kumar, N., & Namdev, N. (2022). Classifying Case Facts and Predicting Legal Decisions of the Indian Central Information Commission: a Natural Language Processing Approach. In *Advances in Deep Learning, Artificial Intelligence and Robotics: Proceedings of the 2nd International Conference on Deep Learning, Artificial Intelligence and Robotics, (ICDLAIR) 2020* (pp. 35-45). Springer International Publishing.

Kalmukov, Y. (2022). Comparison of latent semantic analysis and vector space model for automatic identification of competent reviewers to evaluate papers. *International Journal of Advanced Computer Science and Applications*, *13*(2).

Kalogeratos, A., Zagorisios, P., & Likas, A. (2016, May). Improving text stream clustering using term burstiness and co-burstiness. In *Proceedings of the 9th hellenic conference on artificial intelligence* (pp. 1-9).

Kamińska, J. A. (2018). Residuals in the modelling of pollution concentration depending on meteorological conditions and traffic flow, employing decision trees. In *ITM Web of Conferences* (Vol. 23, p. 00016). EDP Sciences.

Kaplan, R. M., & Bresnan, J. (1981). *Lexical-functional grammar: A formal system for grammatical representation*. Massachusetts Institute Of Technology, Center For Cognitive Science.

Kashyap, A., Han, L., Yus, R., Sleeman, J., Satyapanich, T., Gandhi, S., & Finin, T. (2016). Robust semantic text similarity using LSA, machine learning, and linguistic resources. *Language Resources and Evaluation, 50,* 125-161.

Kee, Y. H., Li, C., Kong, L. C., Tang, C. J., & Chuang, K. L. (2019). Scoping review of mindfulness research: A topic modelling approach. *Mindfulness, 10,* 1474-1488.

Keraghel, I., Morbieu, S., & Nadif, M. (2024, April). Beyond words: a comparative analysis of LLM embeddings for effective clustering. In *International Symposium on Intelligent Data Analysis* (pp. 205-216). Cham: Springer Nature Switzerland.

Khadse, V. M., Mahalle, P. N., & Shinde, G. R. (2020). Statistical study of machine learning algorithms using parametric and non-parametric tests: A comparative analysis and recommendations. *International Journal of Ambient Computing and Intelligence (IJACI), 11*(3), 80-105.

Khandare, A., Agarwal, N., Bodhankar, A., Kulkarni, A., & Mane, I. (2023). Study of Python libraries for NLP. *International Journal of Data Analysis Techniques and Strategies, 15*(1-2), 116-128.

Kherwa, P., & Bansal, P. (2017, September). Latent semantic analysis: an approach to understand semantic of text. In *2017 international conference on current trends in computer, electrical, electronics and communication (CTCEEC)* (pp. 870-874). IEEE.

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia tools and applications*, *82*(3), 3713-3744.

Kiatkawsin, K., Sutherland, I., & Kim, J. Y. (2020). A comparative automated text analysis of airbnb reviews in Hong Kong and Singapore using latent dirichlet allocation. *Sustainability*, *12*(16), 6673.

Kilicoglu, H., & Bergler, S. (2008). Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC bioinformatics*, *9*, 1-10.

Koller, D., & Sahami, M. (1996). *Toward optimal feature selection*. Stanford InfoLab.

Kopp, M., Pevný, T., & Holena, M. (2014). Interpreting and clustering outliers with sapling random forests. *Information Technologies—Applications and Theory.*

Koroteev, M. V. (2021). BERT: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943.*

Krishnan, V., & Manning, C. D. (2006, July). An effective two-stage model for exploiting non-local dependencies in named entity recognition. In *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics* (pp. 1121-1128).

Kroese, D. P., Botev, Z., & Taimre, T. (2019). *Data science and machine learning: mathematical and statistical methods*. Chapman and Hall/CRC.

Kubal, D. R., & Nimkar, A. V. (2019). A survey on word embedding techniques and semantic similarity for paraphrase identification. *International Journal of Computational Systems Engineering*, *5(1),* 36-52.

Kulkarni, D., Thakur, A., & Kshirsagar, J. (2017). Automatic Answer Sheet Evaluation Using Efficient & Reliable OCR System.

Kumar, A. (2019). Mastering pandas: A complete guide to pandas, from installation to advanced data analysis techniques. *Packt Publishing Ltd*.

Kumar, A., Sangwan, S. R., & Nayyar, A. (2020). Multimedia social big data: Mining. *Multimedia Big Data Computing for IoT Applications: Concepts, Paradigms and Solutions*, 289-321.

Kumar, R., & Raghuveer, K. (2012). Legal document summarization using latent dirichlet allocation. *International Journal of Computer Science and Telecommunications*, *3*(7), 8-23.

Kumaran, G., & Allan, J. (2004, July). Text classification and named entities for new event detection. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 297-304).

Kundeti, S. R., Vijayananda, J., Mujjiga, S., & Kalyan, M. (2016, December). Clinical named entity recognition: Challenges and opportunities. In *2016 IEEE International Conference on Big Data (Big Data)* (pp. 1937-1945). IEEE.

L'heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. (2017). Machine learning with big data: Challenges and approaches. *Ieee Access*, *5*, 7776-7797.

Lahitani, A. R., Permanasari, A. E., & Setiawan, N. A. (2016, April). Cosine similarity to determine similarity measure: Study case in online essay assessment. In *2016 4th International Conference on Cyber and IT Service Management* (pp. 1-6). IEEE.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes, 25(2-3),* 259-284.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse processes, 25(2-3), 259-284.

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997, August). How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In *Proceedings of the 19th annual meeting of the Cognitive Science Society* (pp. 412-417).

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997, August). How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In *Proceedings of the 19th annual meeting of the Cognitive Science Society* (pp. 412-417).

Lee Rodgers, J., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, *42*(1), 59-66.

Lee, H., Surdeanu, M., & Jurafsky, D. (2017). A scaffolding approach to coreference resolution integrating statistical and rule-based models. *Natural Language Engineering, 23(5),* 733-762.

Lee, J., Sung, M., Kang, J., & Chen, D. (2020). Learning dense representations of phrases at scale. *arXiv preprint arXiv:2012.12624*.

Levy, D. M. (2000, January). Topics in document research. In *Proceedings of the ACM Conference on Document Processing Systems* (pp. 187-193).

Li, B., & Fonseca, F. (2006). TDD: A comprehensive model for qualitative spatial similarity assessment. *Spatial Cognition and Computation*, *6*(1), 31-62.

Li, F., Liao, L., Zhang, L., Zhu, X., Zhang, B., & Wang, Z. (2020). An efficient approach for measuring semantic similarity combining wordnet and wikipedia. *IEEE Access*, *8*, 184318-184338.

Li, J., Sun, A., Han, J., & Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering, 34(1),* 50-70.

Li, L., Dai, S., Cao, Z., Hong, J., Jiang, S., & Yang, K. (2020). Using improved gradient-boosted decision tree algorithm based on Kalman filter (GBDT-KF) in time series prediction. *The Journal of Supercomputing*, *76*, 6887-6900.

Li, Y., Bandar, Z. A., & McLean, D. (2003). An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on knowledge and data engineering*, *15*(4), 871-882.

Li, Y., Bandar, Z. A., & McLean, D. (2003). An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on knowledge and data engineering*, *15*(4), 871-882.

Lim, D. (2021). Saving substantial similarity. *Fla. L. Rev.*, *73*, 591.

Linstead, E., Rigor, P., Bajracharya, S., Lopes, C., & Baldi, P. (2007, November). Mining concepts from code with probabilistic topic models. In *Proceedings of the 22nd IEEE/ACM International Conference on Automated Software Engineering* (pp. 461-464).

Liu, C. Z., Sheng, Y. X., Wei, Z. Q., & Yang, Y. Q. (2018, August). Research of text classification based on improved TF-IDF algorithm. In 2018 *IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE)* (pp. 218-222). IEEE.

Liu, M., Lang, B., & Gu, Z. (2017). Calculating semantic similarity between academic articles using topic event and ontology. *arXiv preprint arXiv:1711.11508.*

Liu, R., Mao, R., Luu, A. T., & Cambria, E. (2023). A brief survey on recent advances in coreference resolution. *Artificial Intelligence Review*, 1-43.

Liu, X. Y., Zhou, Y. M., & Zheng, R. S. (2008, July). Measuring semantic similarity within sentences. In *2008 International Conference on Machine Learning and Cybernetics* (Vol. 5, pp. 2558-2562). IEEE.

Lucchi, N. (2023). ChatGPT: a case study on copyright challenges for generative artificial intelligence systems. *European Journal of Risk Regulation*, 1-23.

Ma, X., & Hovy, E. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Mahdawi, A. (2022). A Brief History of Dreams: A Mini-Review, Analysis of Dream Articles With Natural Language Processing, and Direction of Dream Research. *Analysis of Dream Articles With Natural Language Processing, and Direction of Dream Research (December 22, 2022).*

Majumder, G., Pakray, P., Gelbukh, A., & Pinto, D. (2016). Semantic textual similarity methods, tools, and applications: A survey. *Computación y Sistemas*, *20*(4), 647-665.

Malte, A., & Ratadiya, P. (2019). Evolution of transfer learning in natural language processing. *arXiv preprint arXiv:1910.07370*.

Manabu, O., & Hajime, M. (2000). Query-Biased Summarization Based on Lexical Chaining. *Computational Intelligence*, *16*(4), 578-585.

Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., & Zamparelli, R. (2014, August). Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)* (pp. 1-8).

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., & Gómez-Berbís, J. M. (2013). Named entity recognition: fallacies, challenges and opportunities. *Computer Standards & Interfaces, 35(5),* 482-489.

Masic, I. (2014). Plagiarism in scientific research and publications and how to prevent it. *Materia socio-médica*, *26*(2), 141.

Matić, D. (2003). Topic, focus, and discourse structure: Ancient Greek word order. Studies in Language. International Journal sponsored by the Foundation "Foundations of Language", 27(3), 573-633.

McCallum, A. (2012). Efficiently inducing features of conditional random fields. *arXiv preprint arXiv:1212.2504*.

Medelyan, O., Milne, D., Legg, C., & Witten, I. H. (2009). Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, *67*(9), 716-754.

Mohammed, S. H., & Al-augby, S. (2020). Lsa & lda topic modeling classification: Comparison study on e-books. *Indonesian Journal of Electrical Engineering and Computer Science*, *19*(1), 353-362.

Moody, C. E. (2016). Mixing dirichlet topic models and word embeddings to make lda2vec. *arXiv preprint arXiv:1605.02019*.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Name Entity Recognition (NER) Dataset. (2021, March 25). *Kaggle. https://www.kaggle.com/datasets/debasisdotcom/name-entity-recognition-ner-dataset*

NER Dataset. (2020, October 2). Kaggle. https://www.kaggle.com/datasets/rohitr4307/ner-dataset

NER_NER_dataset.(2020,August 20).Kaggle. https://www.kaggle.com/datasets/namanj27/ner-dataset

Nguyen, H. T., Duong, P. H., & Cambria, E. (2019). Learning short-text semantic similarity with word embeddings and external knowledge sources. *Knowledge-Based Systems*, *182*, 104842.

Nguyen, M. T., Le, D. T., & Le, L. (2021). Transformers-based information extraction with limited data for domain-specific business documents. *Engineering Applications of Artificial Intelligence, 97,* 104100.

Odden, T. O. B., Marin, A., & Rudolph, J. L. (2021). How has Science Education changed over the last 100 years? An analysis using natural language processing. *Science Education*, *105*(4), 653-680.

Okaka, R. A. (2018). *A Hybrid Approach for Personalized Recommender System Using Weighted Term Frequency Inverse Document Frequency* (Doctoral dissertation, JKUAT-COPAS).

Ostertagová, E. (2012). Modelling using polynomial regression. *Procedia engineering*, *48*, 500-506.

Ozili, P. K. (2023). The acceptable R-square in empirical modelling for social science research. In *Social research methodology and publishing results: A guide to non-native english speakers* (pp. 134-143). IGI global.

Pakray, P., Bandyopadhyay, S., & Gelbukh, A. (2011). Textual entailment using lexical and syntactic similarity. *International Journal of Artificial Intelligence and Applications*, *2*(1), 43-58.

Patil, R., Boit, S., Gudivada, V., & Nandigam, J. (2023). A survey of text representation and embedding techniques in nlp. *IEEE Access*, *11*, 36120-36146.

Pawar, A., & Mago, V. (2018). Calculating the similarity between words and sentences using a lexical database and corpus statistics. *arXiv preprint arXiv:1802.05667*.

Peinelt, N., Nguyen, D., & Liakata, M. (2020, July). tBERT: Topic models and BERT joining forces for semantic similarity detection. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7047-7055).

Petrova, A., Kostylev, E. V., Cuenca Grau, B., & Horrocks, I. (2019). Query-based entity comparison in knowledge graphs revisited. In *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part I 18* (pp. 558-575). Springer International Publishing.

Pincombe, B. (2004). *Comparison of human and latent semantic analysis (LSA) judgements of pairwise document similarities for a news corpus*. Edinburgh: DSTO Information Sciences Laboratory.

Pinto, D., Civera, J., Barrón-Cedeno, A., Juan, A., & Rosso, P. (2009). A statistical approach to crosslingual natural language tasks. *Journal of Algorithms*, *64*(1), 51-60.

Ponte, J. M., & Croft, W. B. (1997). Text segmentation by topic. In *Research and Advanced Technology for Digital Libraries: First European Conference, ECDL'97 Pisa, Italy, September 1–3, 1997 Proceedings 1* (pp. 113-125). Springer Berlin Heidelberg.

Prakoso, D. W., Abdi, A., & Amrit, C. (2021). Short text similarity measurement methods: a review. *Soft Computing*, *25*, 4699-4723.

Qaiser, S., & Ali, R. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications*, *181(1)*, 25-29.

Qi, Y. (2012). Random forest for bioinformatics. *Ensemble machine learning: Methods and applications*, 307-323.

Quick, P. (2005). 10 Topic continuity, voice and word order in Pendau. The many faces of Austronesian voice systems: Some new empirical studies, 221.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*, 81-106.

Raschka, S., & Mirjalili, V. (2019). *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. Packt publishing ltd.

Rau, L. F. (1991, January). Extracting company names from text. In *Proceedings the Seventh IEEE Conference on Artificial Intelligence Application* (pp. 29-30). IEEE Computer Society.

Reimers, N., Beyer, P., & Gurevych, I. (2016, December). Task-oriented intrinsic evaluation of semantic textual similarity. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 87-96).

Ren, P., Chen, Z., Ren, Z., Wei, F., Nie, L., Ma, J., & De Rijke, M. (2018). Sentence relations for extractive summarization with deep neural networks. *ACM Transactions on Information Systems (TOIS)*, *36*(4), 1-32.

Ren, P., Chen, Z., Ren, Z., Wei, F., Nie, L., Ma, J., & De Rijke, M. (2018). Sentence relations for extractive summarization with deep neural networks. *ACM Transactions on Information Systems (TOIS), 36(4)*, 1-32.

Rice, D. B., Raffoul, H., Ioannidis, J. P., & Moher, D. (2020). Academic criteria for promotion and tenure in biomedical sciences faculties: cross sectional analysis of international sample of universities. *Bmj, 369*.

R-square in empirical modelling for social science research. In *Social research methodology and publishing results: A guide to non-native english speakers* (pp. 134-143). IGI Global.

Rus, V., Banjade, R., & Lintean, M. C. (2014, May). On Paraphrase Identification Corpora. In *LREC* (pp. 2422-2429).

Sáenz, C. A. C., & Becker, K. (2023). Understanding stance classification of BERT models: an attention-based framework. *Knowledge and Information Systems*, 1-33.

Sag, I. A. (1991, November). Linguistic theory and natural language processing. In *Natural Language and Speech: Symposium Proceedings Brussels, November 26/27, 1991* (pp. 69-83). Berlin, Heidelberg: Springer Berlin Heidelberg.

Salloum, S. A., Khan, R., & Shaalan, K. (2020). A survey of semantic analysis approaches. In *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020)* (pp. 61-70). Springer International Publishing.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), 513-523.

Samuelson, P. (1999). Privacy as intellectual property. *Stan. L. Rev.*, *52*, 1125.

Savage-Rumbaugh, E. S., Murphy, J., Sevcik, R. A., Brakke, K. E., Williams, S. L., Rumbaugh, D. M., & Bates, E. (1993). Language comprehension in ape and child. *Monographs of the society for research in child development*, i-252.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461-464.

Seber, G. A., & Lee, A. J. (2012). *Linear regression analysis*. John Wiley & Sons.

Sha, F., & Pereira, F. (2003). Shallow parsing with conditional random fields. In *Proceedings of the 2003 human language technology conference of the North American Chapter of the Association for Computational Linguistics* (pp. 213-220).

Shaik, A. B., & Srinivasan, S. (2019). A brief survey on random forest ensembles in classification model. In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2018, Volume 2* (pp. 253-260). Springer Singapore.

Shams, M., & Baraani-Dastjerdi, A. (2017). Enriched LDA (ELDA): Combination of latent Dirichlet allocation with word co-occurrence analysis for aspect extraction. *Expert Systems with Applications*, *80*, 136-146.

Shao, M., & Qin, L. (2014, March). Text similarity computing based on LDA topic model and word co-occurrence. In *2014 2nd International Conference on Software Engineering, Knowledge Engineering and Information Engineering (SEKEIE 2014))* (pp. 199-203). Atlantis Press.

Sharnagat, R. (2014). Named entity recognition: A literature survey. *Center For Indian Language Technology*, 1-27.

Si, J., Li, Q., Qian, T., & Deng, X. (2014). Users' interest grouping from online reviews based on topic frequency and order. World Wide Web, 17, 1321-1342.

Silveira, T., Zhang, M., Lin, X., Liu, Y., & Ma, S. (2019). How good your recommender system is? A survey on evaluations in recommendation. *International Journal of Machine Learning and Cybernetics*, *10*, 813-831.

Singh, A. K., & Shashi, M. (2019). Vectorization of text documents for identifying unifiable news articles. *International Journal of Advanced Computer Science and Applications*, *10*(7).

Singh, R., & Singh, S. (2021). Text similarity measures in news articles by vector space model using NLP. Journal of The Institution of Engineers (India): Series B, 102, 329-338.

Singhal, A., Salton, G., Mitra, M., & Buckley, C. (1996). Document length normalization. *Information Processing & Management*, *32*(5), 619-633.

Sintia, S., Defit, S., & Nurcahyo, G. W. (2021). Product Codefication Accuracy With Cosine Similarity And Weighted Term Frequency And Inverse Document Frequency (TF-IDF). *Journal of Applied Engineering and Technological Science (JAETS)*, *2*(2), 62-69.

Sintia, S., Defit, S., & Nurcahyo, G. W. (2021). Product Codefication Accuracy With Cosine Similarity And Weighted Term Frequency And Inverse Document Frequency (TF-IDF). Journal of Applied Engineering and Technological Science (JAETS), 2(2), 62-69.

Sitikhu, P., Pahi, K., Thapa, P., & Shakya, S. (2019, November). A comparison of semantic similarity methods for maximum human interpretability. In *2019 artificial intelligence for transforming business and society (AITB)* (Vol. 1, pp. 1-4). IEEE.

Slimani, T. (2013). Description and evaluation of semantic similarity measures approaches. *arXiv preprint arXiv:1310.8059*.

Soares, L. B., FitzGerald, N., Ling, J., & Kwiatkowski, T. (2019). Matching the blanks: Distributional similarity for relation learning. arXiv preprint arXiv:1906.03158.

Soares, V. H. A., Campello, R. J., Nourashrafeddin, S., Milios, E., & Naldi, M. C. (2019). Combining semantic and term frequency similarities for text clustering. *Knowledge and Information Systems*, *61*, 1485-1516.

Song, C., Ristenpart, T., & Shmatikov, V. (2017, October). Machine learning models that remember too much. In *Proceedings of the 2017 ACM SIGSAC Conference on computer and communications security* (pp. 587-601).

Soon, W. M., Ng, H. T., & Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, *27*(4), 521-544.

Sozou, P. D., Cootes, T. F., Taylor, C. J., Di Mauro, E. C., & Lanitis, A. (1997). Non-linear point distribution modelling using a multi-layer perceptron. *Image and Vision Computing*, *15*(6), 457-463.

Srinivasa-Desikan, B. (2018). Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras. Packt Publishing Ltd.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012, July). Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 952-961).

Stim, R. (2022). *Getting permission: Using & licensing copyright-protected materials online & off*. Nolo.

Stone, B., Dennis, S., & Kwantes, P. J. (2011). Comparing methods for single paragraph similarity analysis. *Topics in Cognitive Science*, *3*(1), 92-122.

Strzalkowski, T. (1995). Natural language information retrieval. *Information Processing & Management, 31(3),* 397-417.

Subramanian, J., & Simon, R. (2013). Overfitting in prediction models–is it a problem only in high dimensions?. *Contemporary clinical trials*, *36*(2), 636-641.

Sukthanker, R., Poria, S., Cambria, E., & Thirunavukarasu, R. (2020). Anaphora and coreference resolution: A review. *Information Fusion, 59*, 139-162.

Suleman, R. M., & Korkontzelos, I. (2021). Extending latent semantic analysis to manage its syntactic blindness. *Expert Systems with Applications*, *165*, 114130.

Sultan, M. A., Bethard, S., & Sumner, T. (2014, August). Dls@ cu: Sentence similarity from word alignment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)* (pp. 241-246).

Syed, S., & Spruit, M. (2017, October). Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE International conference on data science and advanced analytics (DSAA)* (pp. 165-174). IEEE.

Takale, S. A., & Nandgaonkar, S. S. (2010). Measuring semantic similarity between words using web documents. *International Journal of Advanced Computer Science and Applications*, *1*(4).

Terra, E. L., & Clarke, C. L. (2003). Frequency estimates for statistical word similarity measures. In Proceedings of the 2003 human language technology conference of the North American Chapter of the Association for Computational Linguistics (pp. 244-251).

The New York Times - Search. (n.d.). https://www.nytimes.com/search/?srchst=nyt

Tomlin, R. S. (2014). Basic Word Order (RLE Linguistics B: Grammar): Functional Principles. Routledge.

Travis, L. D. (1984). Parameters and effects of word order variation (Doctoral dissertation, Massachusetts Institute of Technology).

Tsekouras, L., Varlamis, I., & Giannakopoulos, G. (2017, September). A Graph-based Text Similarity Measure That Employs Named Entity Information. In RANLP (pp. 765-771).

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, *37*, 141-188.

Tursunbayeva, A., Pagliari, C., Di Lauro, S., & Antonelli, G. (2022). The ethics of people analytics: risks, opportunities and recommendations. *Personnel Review*, *51*(3), 900-921.

Vajjala, S., Majumder, B., Gupta, A., & Surana, H. (2020). *Practical natural language processing: A comprehensive guide to building real-world NLP systems*. O'Reilly Media.

Van Aken, B., Winter, B., Löser, A., & Gers, F. A. (2019, November). How does bert answer questions? a layer-wise analysis of transformer representations. In *Proceedings of the 28th ACM international conference on information and knowledge management* (pp. 1823-1832).

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

Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7-16.

Vijaymeena, M. K., & Kavitha, K. (2016). A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, *3*(2), 19-28.

Wali, W., Gargouri, B., & Hamadou, A. B. (2015). Supervised learning to measure the semantic similarity between arabic sentences. In Computational Collective Intelligence: *7th International Conference, ICCCI 2015, Madrid, Spain, September 21-23, 2015, Proceedings, Part I* (pp. 158-167). Springer International Publishing.

Wang, C., & Blei, D. M. (2011, August). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 448-456).

Wang, J., & Dong, Y. (2020). Measurement of text similarity: a survey. *Information*, *11*(9), 421.

Wang, J., & Dong, Y. (2020). Measurement of text similarity: a survey. *Information*, *11*(9), 421.

Wang, M. C., & Bushman, B. J. (1998). Using the normal quantile plot to explore meta-analytic data sets. *Psychological Methods*, *3*(1), 46.

Wang, W., Feng, Y., & Dai, W. (2018). Topic analysis of online reviews for two competitive products using latent Dirichlet allocation. *Electronic Commerce Research and Applications*, *29*, 142-156.

Wang, X., Jiang, Y., Bach, N., Wang, T., Huang, Z., Huang, F., & Tu, K. (2021). Improving named entity recognition by external context retrieving and cooperative learning. *arXiv preprint arXiv:2105.03654*.

Warner, B., & Misra, M. (1996). Understanding neural networks as statistical tools. *The american statistician*, *50*(4), 284-293.

Wasti, S. H., Hussain, M. J., Huang, G., Akram, A., Jiang, Y., & Tang, Y. (2020). Assessing semantic similarity between concepts: A weighted-feature-based approach. *Concurrency and Computation: Practice and Experience*, *32*(7), e5594.

Weiss, S. M., Indurkhya, N., Zhang, T., Damerau, F. J., Weiss, S. M., Indurkhya, N., ... & Damerau, F. J. (2005). From textual information to numerical vectors. *Text Mining: Predictive Methods for Analyzing Unstructured Information*, 15-46.

Wikipedia contributors. (2024, June 22). *Adolf Hitler* - Wikipedia. https://en.wikipedia.org/wiki/Adolf_Hitler

Wikipedia contributors. (2024a, June 14). *Causes of World War II.* Wikipedia. https://en.wikipedia.org/wiki/Causes_of_World_War_II

Wikipedia contributors. (2024c, June 30). *World War II* - Wikipedia. https://en.wikipedia.org/wiki/World_War_II

Wu, J., Yang, S., Zhan, R., Yuan, Y., Wong, D. F., & Chao, L. S. (2023). A survey on llm-gernerated text detection: Necessity, methods, and future directions. *arXiv preprint arXiv:2310.14724*.

Xian, J., Teofili, T., Pradeep, R., & Lin, J. (2024, March). Vector search with OpenAI embeddings: Lucene is all you need. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining* (pp. 1090-1093).

Yang, M., Cui, T., & Tu, W. (2015, February). Ordering-sensitive and semantic-aware topic modeling. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 29, No. 1).

Yang, Y., Zhang, J., Meng, Z. L., Qin, L., Liu, Y. F., & Bi, H. Y. (2018). Neural correlates of orthographic access in Mandarin Chinese writing: An fMRI study of the word-frequency effect. *Frontiers in Behavioral Neuroscience, 12,* 288.

Ye, X., Shen, H., Ma, X., Bunescu, R., & Liu, C. (2016, May). From word embeddings to document similarities for improved information retrieval in software engineering. In *Proceedings of the 38th international conference on software engineering* (pp. 404-415).

Yin, P., & Fan, X. (2001). Estimating R 2 shrinkage in multiple regression: A comparison of different analytical methods. *The Journal of Experimental Education, 69*(2), 203-224.

Zhang, Q., & Zhang, R. (2021). An Evaluation on Entity Extraction and Semantic Similarity Metrics to Facilitate Medical Text Analysis Based on WordNet. In *Design, Operation and Evaluation of Mobile Communications: Second International Conference, MOBILE 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings 23* (pp. 138-151). Springer International Publishing.

Zhang, W., Li, Z., Wang, Q., & Li, J. (2019). FineLocator: A novel approach to method-level fine-grained bug localization by query expansion. *Information and Software Technology*, *110*, 121-135.

Zhang, Y., Chen, M., Huang, D., Wu, D., & Li, Y. (2017). iDoctor: Personalized and professionalized medical recommendations based on hybrid matrix factorization. *Future Generation Computer Systems*, *66*, 30-35.

Zhang, Z., & Wu, Z. (2021, July). Improved TF-IDF algorithm combined with multiple factors. In *2021 3rd International Conference on Applied Machine Learning (ICAML)* (pp. 492-495). IEEE.

Zhao, F., Ren, X., Yang, S., Han, Q., Zhao, P., & Yang, X. (2020). Latent dirichlet allocation model training with differential privacy. *IEEE Transactions on Information Forensics and Security*, *16*, 1290-1305.

Zhou, Y., Booth, S., Ribeiro, M. T., & Shah, J. (2022, June). Do feature attribution methods correctly attribute features?. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 9, pp. 9623-9633).

Zhu, G., & Iglesias, C. A. (2018). Exploiting semantic similarity for named entity disambiguation in knowledge graphs. *Expert Systems with Applications*, *101*, 8-24.

Zhu, Z., Liang, J., Li, D., Yu, H., & Liu, G. (2019). Hot topic detection based on a refined TF-IDF algorithm. *IEEE access*, *7*, 26996-27007.

Zulkifeli, W. R. W., & Rusila, W. (2013). *Term frequency and inverse document frequency with position score and mean value for mining web content outliers* (Doctoral dissertation, Universiti Putra Malaysia).