THE UNIVERSITY OF CHICAGO

Linking Scientific Research: Extracting Sentences Containing Dataset Information

By

Yilun Xu

December 2021

A paper submitted in partial fulfillment of the requirements for the Master of Arts degree in the Master of Arts in Computational Social Science

Faculty Advisor: Professor James Allen Evans

Preceptor: Shilin Jia

Linking Scientific Research: Extracting Sentences Containing Dataset Information

December 2021

Contents

Acknowledgements

Officially and academically.

I would like to thank Professor James Allen Evans, my preceptor, Shilin Jia, Paulino Diaz Mejia, Yutao Chen, and Donghyun Kang for their significant assistance with my thesis. Professor Evans offered us a vast amount of knowledge on how to proceed with the research, especially when we hit unexpected roadblocks. Professor Evans and Shilin have been extremely helpful throughout my master's program in computational social science, in addition to my thesis. Paulino contributed a great deal in the initial stages of coding, and he provides valuable support throughout the course of this project. Yutao and Donghyun's support on technique issues is highly appreciated as well.

Unofficially and sentimentally.

I saw James, smiling, take off his white half-rim glasses after he finally gave a "begrudging" nod, at the mere thought of the shadows of the blinds in my bedroom. Thousands of times my body, stretching or scrunching, mottled on the blinds as with a floating ghost or a fleeing bird, above the Lake, staring at the prohibitive far afield. The window mirrored days and nights completely over the two shattered years, serving as a roll of films burning after shining in the ashes of time. This pair of glasses, a bit postmodern, resembles the survivors bantering life and death, or solely a span celebration by trembling wings of a slow sparrow.

Semi-officially and privately.

In her typical voice like mine, my mom was insistent that my grandmother and she be highlighted above my grandfather and my father. However, the disguised threat flies through the internet. As a promise, I will spare no effort to acknowledge them in my Ph.D. dissertation in the future. In particular, I would like to thank Mr. Xiang Zhang who has offered me tremendous selfless support since my dream began at UChicago, and consistently demonstrated his ability in drama (sarcasm, to be honest), and perhaps

in economics as well. Wish him all the best at Princeton and support me as he did, to secure his position in my next acknowledgements. Much thanks to Doctor to-be Dimitrios Tanoglidis. As he said he has been watching me grow up from a *baby* step by step, and his encouragement equipped me with the wings to weather ups and downs in the last two years at UChicago. I feel honored to feel loved by my superiors and contemporaries, in Chicago or any other corner of the globe, who soothe me whenever my fried feathers flap. Each of my dreams depicts my ideal for our reunions.

To the arrival and departure in my life. To my upcoming 25th birthday.

Abstract

Despite the popularity of data-driven research in scientific fields, we are intrigued by the combined value of datasets in a given area. Our research seeks to establish strategies for retrieving words containing dataset information from academic publications using a specific example of COVID-19 epidemiological papers, which was encouraged by previous studies concerning research originality and how combinatorial work improves science. We deployed LDA and word embedding algorithms to filter epidemiological papers versus clinical ones. We also annotated sentences based on whether each sentence in the abstract and title parts mentions dataset information. "Pre-trained" word representations enabled classification models to discriminate between data and non-data sentences. The unexpected finding is that, while more diverse terms in a publication's abstract and title help advertise it in terms of citation, they make this document less likely to be one of the top-cited papers. In conclusion, while we have not reached accurate conclusions for identifying data sentences in papers, we have uncovered techniques for filtering *possible* data sentences. We suggest inspecting a larger corpus in the next stage to evaluate the impact of alternative datasets and gather more information for the paper's word representations and citation.

1 Introduction

Humans have made it feasible to gather, store, and use large data sets thanks to the advancement of computer science. As a result, the use of large datasets has permeated every aspect of scientific study. The explosive appearance of massive datasets in recent years has piqued scholars' interest in investigating its role in advancing scientific research novelty [1, 2]. For example, Agarwal and Dhar [1] have examined how large-scale data shape studies on information systems, showing that it has empowered the study of information systems, which is consistent with the aim of *ISR*, the journal Information Systems Research, without reaching quantitative conclusions.

It has been demonstrated that big data has been incorporated into science research, and in most studies, multiple datasets are examined in order to test hypotheses. The datasets from various sources can be combined to form "the same units" [3], or they can serve multiple purposes within a research. The integration and exchange of datasets involved in the same research has influenced the course that science takes [4]. This occurrence prompted us to investigate the utility of dataset combination. The ultimate mission is to demonstrate how datasets can be integrated to examine a research issue in the advancement of science in a specific area of study.

This research used a situational analysis to extract words providing dataset information from scholarly publications on a particular academic topic. In this study, we utilize data sentences to refer to sentences holding dataset information, and non-data sentences to indicate the reverse. To be more specific, we attempted to collect data sentences regarding COVID-19 from epidemiological publications. Since the outbreak of COVID-19, many data and hundreds of relevant papers have been released concerning this global crisis, and the use of large datasets is becoming more fashionable, especially during the pandemic when global researchers engage together to combat COVID-19 [5]. Diverse disciplines align their efforts to combat this global challenge. There are two distinct groups of studies within epidemiological and clinical research as evident from the available literature. These facts highlight the significance of our study in assessing human attempts to counteract

COVID-19 and ensure the validity of our findings owing to the large data volumes in our corpus. We can collect the relevant textual components from the epidemiological and clinical publications and develop a methodology for choosing studies reflecting specific themes that can be applied to other subject areas since we possess a large number of COVID-19-related papers. We are committed to locating data sentences within the works selected by the auto-selection procedure for epidemiology research.

We conducted an empirical study in which we attempted to separate data sentences from epidemiological papers from clinical papers that discussed COVID-19 and focused on specific academic topics. In other words, COVID-19 epidemiological papers are an example of scholarly publications specializing in a particular academic area, from which our strategies can be applied to any scholarly work. Epidemiological papers are not chosen because clinical papers do not utilize datasets. In clinical papers, the data used is more frequently derived from laboratory experiments designed for specific studies than those used in epidemiological papers. In general, these experiments are less likely to be combined with data from other studies. Since epidemiology publications are believed to have a more substantial number of common datasets, we chose to refer to them as an example of scholarly publications. As a result, the following portions of our work will be discussed in this case study:

1. How to distinguish epidemiological papers against clinical papers among COVID-19 related work, as an example of how to select papers focusing on any specific topic?

2. How to uncover sentences carrying dataset information in each COVID-19 epidemiological paper, as an example of how to discover data sentences information from papers with homogeneous topics?

Detecting data sentences in scholarly papers pertaining to a specific academic topic is not our ultimate goal. By discovering data sentences, we are setting the groundwork for the next step in the process. This step is to discover the relationship between datasets and hypotheses, whether in science in general or in a specific area of science.

2 Literature Review

Existing studies in Science of Science

For decades, academic publications have been used to study the progress of science in diverse fields, and citation indices have survived and thrived in these works (see, e.g., [6–10]) because citation can not only reflect how researchers gain wisdom from existing studies, but it also works as an effective measurement to estimate the impact of an academic publication [6]. Additionally, scholars have proved that the combination of precedent work enlightens the followers [11–13], while no direct evidence has provided a hint on the value of dataset combination. We benefited from studies where novelty was explored. Specifically, measurements relied on citation has become increasingly popular to evaluate papers for the paper since the last 50s [14, 15]. Citation rates, distributions, and averages are considered to be research evaluation metrics [16, 17]. With the advancement of computational technologies, more studies focusing on citation data and issues in the science of science [12, 15]. For example, citation data from the Microsoft Academic Graph (MAG) is leveraged in a publication by Lin et al. [15] to allude to how people learnt from and incorporated prior work in their research to investigate the disruptive implications of works.

The preceding papers provide directions for implementing citation analysis in our investigation. Although citation indices show its dominant influence in studying scientific collaboration and impact, other criteria including co-authorship [18] and specific professional domains [19] also contributes in previous studies depicting the scientific collaboration networks. However, prior studies centering heavily on the advancement of science, on the other hand, focused on the research entities, including individuals and organizations, rather than the materials employed during the studies. Examining how entities interact, directly or indirectly, to support scientific discovery has gotten more sophisticated, yet collaboration across datasets and ideas lacks emphasis in contrast. This phenomenon is where our study's worth shines, since it delivers pipelines to extract information for papers concealed in the texts.

Computational Technologies at the Cutting Edge

Word representations have long been a key matter in natural language processing (NLP), acting as a bridge between textual data and computation approaches for numeric data [20]. Those technologies have been applied to classify text data like sentiment analysis [21, 22] and sarcasm detection [23], and also intersect with topic modeling to extract information from corpus [24–26] theoretically and empirically. For example, Ren et al. gained "topic-enhanced" word representations with a recursive autoencoder model, and conducted sentiment classification with LDA using Twitter corpus [27], where large datasets bolstered research findings. Naseem et al. [20] evaluated the effectiveness of several word embedding approaches on classification and regression models in a range of NLP-oriented tasks. The achievement of deep learning classifiers on semantic analysis and text classification tasks suggests that natural language processing (NLP) would become more popular in extracting information buried in textual data, including multilingual data, in the future. These studies motivate us on how to filter data sentences given a specific academic topic, based on which classification and regression algorithms are applied.

Previous research provided us with the idea of categorizing textual data. Admittedly, the concepts of word embedding centroid and cosine distance between each pair of word representation vectors have been explored by previous studies and the distances among word representation vectors have been introduced to discuss the diversity of information contained in each document, both theoretically and practically (see, e.g., [28–30]). These ideas have been proved to be effective in empirical studies including recognizing the emotions and attitudes of different entities [28]. In other words, such approaches are effective in classifying texts on a variety of themes. In this work, we advanced text classification by integrating current word representation approaches and creating new characteristics for each text that served our study aim in addition to word embedding centroid and cosine distance.

3 Data

The datasets used in this study consist of three parts: The COVID-19 Open Research Dataset [31], Web of Science Database [32,33] and Microsoft Academic Graph Database [7,34,35].

The COVID-19 Open Research Dataset

As part of the ongoing fight against this global challenge and future scientific research, papers focusing on COVID-19, also called CORD-19, were compiled during the pandemic. Researchers at the Allen Institute for AI developed natural language processing tools to automatically release updated datasets weekly or biweekly [31]. We can utilize the data based on their online releases [1], where the title, abstract, and paper ids are provided. Generally, later releases always contain the papers in the previous releases. When some publications considered irrelevant to COVID-19 are subsequently recognized as qualified ones, there may still be some differences in their algorithms. We have largely relied on the version of August 5th, 2020 for our data. Essentially, this corpus is used for extracting data from epidemiological studies and comparing them with clinical publications, as well as creating annotations.

Web of Science Database

Known as the WoS dataset, it contains publications from a variety of fields, including COVID-19 related papers, together with corresponding information, including citations and abstracts. It is primarily used for building a word2vec model in order to analyze annotations. In other words, the word2vec model created using corpus from this database served as a *pre-trained* embedding model in our study. Due to limitations in computing power, we collected only 1 million samples to build the word2vec model and obtain the *pre-trained* embeddings for each paper in our CORD-19 dataset.

[1] `https://ai2-semanticscholar-cord-19.s3-us-west-2.amazonaws.com/historical_releases.html`

Microsoft Academic Graph Database

This database is developed by researchers at Microsoft, where the academic publications are linked to graphs by scholarly relationships including citation [7, 34, 35], where we can see the collaboration based on citation of each paper with the graph database. For each graph, academic entities are the nodes with the edges displaying their connections [7]. This database has been used for studying research developments across different scientific issues including COVID-19 (see, e.g., [36–38]), single or combined with other datasets [38]. Based on data from this database, we have extracted citation times for papers in the CORD-19 corpus and analyzed the relationship between citation times and paper content diversity.

4 Methodologies

The following methodologies were applied in this research: TF-IDF, word embeddings, topic modeling, regression and classification, and dimension reduction.

TF-IDF

TF-IDF, referring to Term Frequency-Inverse Document Frequency, is a data mining technique for extracting essential words from a corpus based on the counts that a specific word appears in a single document and the whole corpus [39]. Namely, this technique is a measurement for the significance of a word to the document [40]. This technique is commonly applied in information retrieval systems [41]. This technique is defined as follows:

$$\begin{aligned} T_{w,d} &= \frac{c_{w,d}}{\mid d \mid} \\ I_{w,C} &= \log(\frac{\mid C \mid}{\mid \{d_i : p_{w,d_i} \neq 0, d_i \in C\} \mid}) \\ \text{TF-IDF}_{w,d,C} &= T_{w,d} \cdot I_{w,C} \end{aligned} \tag{1}$$

where we have

- $c_{w,d}$: the count of word w in a single document d.
- $\mid d \mid$: the length (or the total number of all words) of the document d.
- $\mid C \mid$: the length (or the total number of documents) of the corpus C.
- $\mid \{d_i : p_{w,d_i} \neq 0, d_i \in C\} \mid$: the number of documents in the corpus C where the word w appears at least one time [41]. [42]

As the above algorithm suggests, the more a word appears in a document, the higher frequency it has in terms of the TF-IDF coefficient, which indicates its higher relevance to the document [40]. TF-IDF is mainly used to extract words related to dataset information and help us filter data sentences in this study. Specifically, we examine which words are more important in each text, especially data sentence clusters. This methodology is adopted to verify our assumption that there might be some obvious word signals in a text indicating that it is a data sentence.

Topic Modeling

Topic modeling algorithms [43,44], as the term implies, aspire to create models to represent the topics covered by the target corpus, with words or vectors [45]. In this study, we intend to distinguish papers or sentences with a particular focus. For example, we plan to inspect which sentences will be highly possible to contain dataset information and which papers are more likely to study epidemiological issues against clinical ones. These technologies allow us to select works specializing in specific and random themes that we aim to investigate in a particular academic context. We could find it easier to check scholarly works regarding COVID-19 simply by entering it as a keyword in a search engine to filter academic publications. However, identifying epidemiological works among the COVID-19-related publications is complicated since "epidemiology" and its cognates are not required for those papers to be written. Two topic models are tested, including latent Dirichlet allocation (LDA) and Discourse atoms.

LDA "is a generative probabilistic model" that constructs corpus topics relied on word count and distribution [46]. This model is used to distinguish epidemiological papers against clinical work. We adopted a *static* LDA model in this study where we do not consider the topic changes over time, against dynamic topic models [47,48] like Hidden Markov Model Latent Dirichlet Allocation (HMM-LDA) [49], because all papers utilized in this part are about COVID-19 which are published no later than 2019, which makes the general topics consistent to some extent.

The discourse atom model is another model we will employ to discover what issues are discussed in the corpus based on word embedding [43,44,50]. Specifically, this model recognizes a set of vectors to represent the word embedding space and then maps each of the vectors into a group of words. This model is used to help us distinguish data sentences against non-data ones as well.

Word Representation

Word representations are computational linguistics tools that aim to convert text data into mathematical representations [51,52]. Consequently, we can apply various algorithms to analyze text data, including dimension reduction techniques to exclude less critical information from our study. In this study, two embedding techniques are mainly used, including word2vec embedding and doc2vec embedding, both based on neural networks. The computation is conducted via the Gensim package in Python [53]. In addition to the embeddings trained by data of smaller size, we also created a "pre-trained" model using the data in our study to represent the textual data for future classification and regression tasks.

Word2vec Models

Word2vec model aims to create a vector for a single word in the corpus where the skip-gram model is employed to predict the word representations from nearby word vectors with hierarchical softmax, or negative sampling techniques drew on maximum likelihood [30, 50, 54]. This approach has a significant impact on retrieving word-level information in this study. With word embedding vectors derived from a trained word2vec model, we obtain the centroid of a sentence/paragraph, represented by the mean of the loading of each word inside. The word2vec embedding centroids are utilized to show the content of each text, relied on which we divide the sentences/paragraphs into different topic clusters. Additionally, we also calculated the standard deviation of each text using the word2vec embedding of each word. The standard deviation values serve as a measurement showing how diverse the information carried by the text is. Specifically, the word2vec standard deviation is defined as follows:

$$\begin{gathered} CD(\vec{v_i}, \vec{v_j}) = 1 - \frac{\vec{v_i} \cdot \vec{v_j}}{\|\vec{v_i}\| \cdot \|\vec{v_j}\|} \\ SD_{text} = \sqrt{\frac{\sum_{i=1}^{n} (CD(\vec{v_i}, \vec{\overline{v}}))^2}{n}} \end{gathered} \tag{2}$$

where we have

- $CD(\vec{v_i}, \vec{v_j})$: the cosine distance between two vectors, $\vec{v_i}$ and $\vec{v_j}$ [55].
- $\vec{\overline{v}}$: the word2vec embedding centroid of the given text.
- SD_{text}: the standard deviation of a given text.
- n: the number of words in the text.

Doc2vec Models

Doc2vec model, also known as *Paragraph Vector* and developed by researchers from Google, creates vectors for each document in the whole corpus. Doc2vec model is inspired by word2vec embedding and borrowed its idea of valuing word semantics in creating the vectors. In addition, previous texts have an impact on the following texts' vectors [56]. Doc2vec model also sheds light on later pre-trained models mainly neural ones [30, 50, 56, 57]. We use word2vec embeddings to study word-level information for various corpus. In contrast and for comparison, this model shapes our understanding of the general information contained in each document and plays an essential role in this study in regression and classification tasks.

The Pre-trained Model

As the name suggests, pre-trained embeddings of a word or a sentence are obtained from a large corpus and will be used to represent new data [58], and there has been some substantial work including BERT [59], GloVe [60] and some other models in Keras in Python [61] as well. Pre-trained word embeddings have been widely accepted in real-word applications and practices (see, e.g., [62–64]). In this study, we do not use any of those existing embeddings but train one word2vec model on around 1 million academic publications using WoS data samples randomly chosen regardless of their academic topics, when the intention to train the word2vec model failed limited by computing power. The reason is that there have not been any pre-trained word embedding models that mainly focus on scholarly works because of the uniqueness of academic writing style and other types of writing and that many terms possibly carry distinct meanings compared with the ones in daily life settings.

Regression and Classification

These tools are used to analyze different features among words, sentences, and paragraphs with word representations. For regression tasks, we mainly use linear regression [65] in this study to explore the standard deviation of a sentence and the text including the title and the abstract of the paper the sentence belonging to and excluding this sentence.

In the meanwhile, the classification methods include Naive Bayes [66, 67], C-support vector classification (linear and polynomial) [68], k-nearest neighbors [69], logistic regression [70], decision trees [71], random forests [72], neural networks [73], and gradient boosting [74]. The computation was completed with the packages scikit-learn and statsmodel in Python [75–77]. We train these models with cross validation [75, 78] with folds to be 5. The classification models are measured and compared by precision, recall and f1-score [75].

These methods aid our understanding of how distinct document characteristics may be used to predict if a text includes dataset information. The models will be trained with labeled samples, and the supervised learning processes shed light on the following unsupervised prediction. We pay more attention to the model performances on data sentence clusters since we seek to understand how we can efficiently decide on whether a sentence is a data one or not.

Dimension Reduction

Dimension reduction techniques aim at transforming higher-dimensional data points into lower-dimensional ones [79], which contribute to reducing computation cost and selecting more critical information carried by the data. We take the dimension reduction technique into account in this study because we assume not all words inside a document are sufficiently valuable to be considered in classifying the text's topic or its property of being a data sentence or not. It is anticipated that dimension techniques work to improve classification model performance in our study, since by intuition not all words in a sentence will contribution to or influence our judgement of whether this sentence contains dataset information or not.

In the computation, we select Uniform Manifold Approximation and Projection (UMAP) [80] to reduce the dimension of word embedding loadings. This non-linear dimension reduction technique is developed relied on manifold theory and topology and is also considered as an "alternative" to t-SNE [80,81]. The computation is completed with the package umap-learn in Python [82].

5 Experiments and Results

Two goals are served by our work: filtering epidemiological articles against clinical paper publications and establishing a pipeline for distinguishing the data sentences contained within epidemiological articles.

5.1 Epidemiological Studies Separation

In this section, we aim to filter paper clusters with a higher proportion of epidemiological papers. Integrated with our efforts in filtering data sentences, results from this section also contributed to our exploration of partitioning data sentences in the next step. To be detailed, the word2vec model, the doc2vec model, the LDA model, and TF-IDF were employed to distinguish epidemiological papers against clinical papers among the CORD-19 corpus of August 5th, 2020, using the title and abstract of each paper. Hierarchical clustering only succeeded in containing 23% to 28% of papers in different levels of clusters using the results of the doc2vec model. As figure 1 shows, the coherence score inspired us to extract seven topics from the sample corpus whose size is around 10% of the full corpus. Then we extracted 7 topics from the full corpus and the topics are shown in table 1.

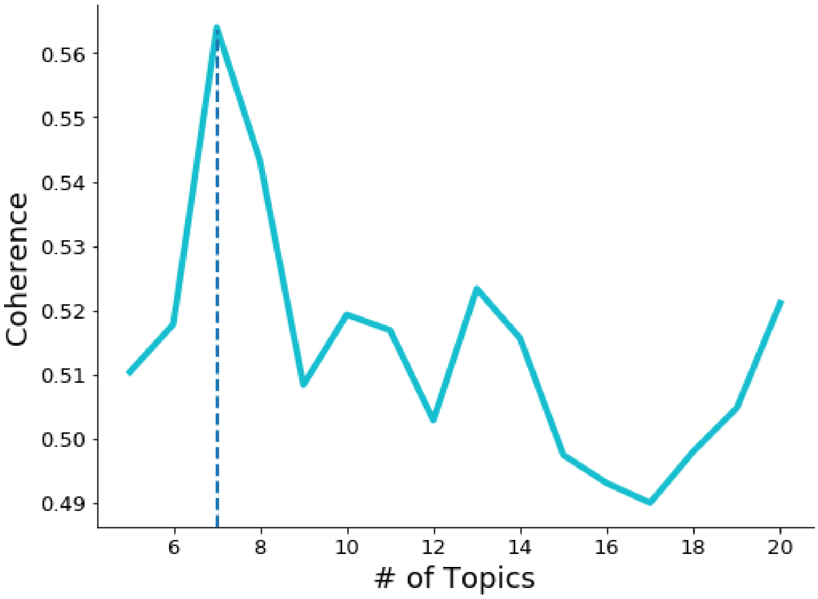

Figure 1: Coherence Scores for Different Numbers of Topics

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
virus	patients	health	treatment	model	respiratory	patients
cells	surgery	pandemic	disease	data	sars	study
protein	care	disease	patients	based	infection	group
viral	laparoscopic	public	clinical	results	coronavirus	results
cell	patient	diseases	studies	time	virus	compared
infection	surgical	research	lung	analysis	influenza	methods
rna	hospital	care	review	study	infections	treatment
viruses	use	infectious	acute	number	severe	age
proteins	procedures	risk	therapy	epidemic	viral	days
human	technique	global	associated	cases	cov	ci

Table 1: LDA Topics of CORD-19 Dataset (2020-08-05)

As a result of a close inspection of the key vocabulary of each extracted topic, Topic 4 appears to be more closely related to the datasets used in each article than the other topics. Hence, papers where topic 4 is the most dominant topic are considered to be epidemiological works. An impressive number of the predicted epidemiological articles are actually epidemiological articles, and the titles of the top five papers with the greatest topic 4 (in table 1) shares are included below.

1. Mathematical model of COVID-19 spread in Turkey and South Africa

2. A Simulation of a COVID-19 Epidemic Based on a Deterministic SEIR Model

3. Understanding Spatio-Temporal Variability in the Reproduction Ratio of the Bluetongue (BTV-1) Epidemic in Southern Spain (Andalusia) in 2007 Using Epidemic Trees

4. Recalibrating disease parameters for increasing realism in modeling epidemics in closed settings

5. Models of epidemics: when contact repetition and clustering should be included

Inspired by the work mentioned above, we kept the exploration in the LDA model and doc2vec model. To begin with, we tried to determine the threshold for labeling a paper to be “topic 4”, and the keywords of topic 4 are shown in table 1. The decision was made by manually evaluating which papers belong to the epidemiological group after all papers were sorted according to their topic 4 shares. Starting from papers ranking 4050th to 4075th, there appeared more clinical papers. The topic 4 probability of the 4050th paper is 86.2081%, and the one of the 4051st paper is 86.2078%. Thus 86.208% is selected to be the threshold. The papers with topic 4 probability exceeding the threshold are considered epidemiological papers, and they are part of our /textbfnew corpus for further study, which contains **78,340** papers.

5.2 Data Sentences Selection

Our analysis can be divided into four modules that serve the second goal: extracting significant words from manually labeled data sentences with manually annotated samples and an existing lexicon database using TF-IDF; classifying data and non-data sentences using word representations and topic model representations; conducting linear regressions on the standard deviation of different sub-sections of the same papers; regressing citation times on the different measurements of text diversity. These four modules do not necessarily conflict with one another, and we may choose to incorporate multiple techniques in order to verify our assumptions.

5.2.1 TF-IDF Frequencies

We tried four strategies to compare the words with higher frequencies using some manually annotated samples in terms of TF-IDF. For each paper, the input for the TF-IDF application becomes one of the following contents:

1. Full Paper minus Data/methods section
2. Title plus Abstract minus Data sentences in the abstract
3. The methods section is compared against all non-methods sections in the paper based on the section names.
4. The data sentences are compared against all non-data sentences in the abstract.

TF-IDF's power to extract words from limited annotated samples revealed the diversity of vocabulary used in papers on various topics.

We repeated the hierarchical clustering method with doc2Vec and word2vec embeddings and the LDA model on the revised corpus with **78,340** papers. We applied TF-IDF to newly created clusters, but we found that the words with the higher tf-idf weights of different clusters were quite similar to ones in other clusters even when the number of clusters increased. When we increased the number of topics for the LDA model, we labeled each sentence as data or non-data. We found similar results when comparing the labels predicted by the highest LDA loading and the full LDA loadings.

With the cluster computed from the doc2vec model, we noticed that in one specific cluster (cluster 11), there are 86 data sentences included among 311 data sentences totally in our annotated samples. Therefore, we decided to build a word dictionary that may be more related to data information using this cluster corpus. At first, we applied TF-IDF and extracted the top 50 words in terms of the frequencies. Princeton WordNet [83] [84] is an English linguistic database that displays lexical relationships among words, developed by researchers from Princeton University. With the online database [85], we included the synonyms of the selected vocabulary using TF-IDF in our dictionary as well. Furthermore, we have incorporated the words related to the selected vocabulary using TF-IDF based on the word relationship shown in the previously trained word2vec model and doc2vec model. In the newly compiled dictionary, we included information on the TF-IDF frequencies of each word that we included.

5.2.2 Classifications on Representations

In the meanwhile, we employed various classification methods on the annotated samples and applied the trained model to the whole corpus where it contained predicted epidemiological papers using the threshold determined in section 5.1. We noticed that Random Forest classifiers performed the most effectively among the classifiers regardless of using word2vec embedding and doc2vec embedding. We applied the trained models to the whole epidemiological corpus, but the frequencies of words in our vocabulary dictionary in the predicted data sentences turned out to be pretty low, which is less than 0.1. Furthermore, we compared the effects of the dictionary only containing words with high TF-IDF frequencies and the extended version, and we found that the extended version failed to improve the results significantly.

The next step was to classify the annotated samples by using discourse atoms, but several of the atoms referred to data-related information, making it difficult to differentiate between data and non-data sentences. In order to retrain the classification models, we made a larger annotation corpus where we had 416 sentences and derived the word probability for each sentence as follows:

$$\text{Word probability} = \frac{\text{Word count in the sentence appearing in the TF-IDF dictionary}}{\text{Sentence length}} \tag{3}$$

where the TF-IDF vocabulary is built from annotated data sentences and comprises only the top 50 words. To filter sentences where TF-IDF vocabulary appears more frequently than in other cases with different thresholds, we selected 0.06 as the benchmark. However, we could not find a large fraction of sentences in which dataset information is directly mentioned.

With the annotated samples consisting of 1311 sentences, we constructed different sets of discourse atoms in which the partition numbers included 25, 35, 50, 65, 75, and 99. To perform a logistic regression, we manually checked the keywords in the discourse atom and selected four atoms per partition number. We computed the minimum cosine distance of a discourse atom and the word2vec loading of each word. Vectors for the discourse atom are derived from the centroid of word2vec loadings of the words contained in this discourse atom. The dependent variable in the logistic regressions is a dummy variable showing whether the sentence is a data sentence. The independent variables vary among the combinations of the top 1, 2, 3, or 4 atoms with the atom(s) added together. As we take the top 4 atoms into account in one of our logistic regression models, McFadden's pseudo-R-squared can reach 0.558.

Furthermore, we knew that the technique of dimension reduction might be able to help us capture more important features in the loadings, as well as improve classification and regression accuracy. We employed UMAP to the word2vec loading centroid of each sentence and re-trained the classification models with five cross-validation folds. Nevertheless, for the data sentences, the accuracy, recall, and F-score of classification models with and without dimension reduction did not vary significantly. The model performances of loadings whose dimensions were reduced with UMAP are shown in table 2.

precision	recall	f1-score	model
0.454545	0.259366	0.330275	Random_forest (cv = 5)
0.447205	0.207493	0.283465	Ensemble (cv = 5)
0.400000	0.288184	0.335008	KNN (cv = 5)
0.392157	0.403458	0.397727	Decision_tree (cv = 5)
0.324324	0.034582	0.062500	Neural_network (cv = 5)
0.314549	0.884726	0.464097	Bayes (cv = 5)
0.285714	0.034582	0.061697	Logistic_regression (cv = 5)
0.000000	0.000000	0.000000	SVC_linear (cv = 5)
0.000000	0.000000	0.000000	SVC_poly (cv = 5)

Table 2: Classification Model Performances of Annotated Samples (UMAP)

In the following step, we decided to utilize word2vec loadings derived from a larger corpus. The Web of Science database contains approximately one million papers published between 1990 and 2019. We trained the word2vec model using the title and abstract of each paper. After updating the word2vec loadings and ones whose dimensions were reduced with UMAP, we repeated the process of logistic regression using newly created discourse atoms. The size change in a corpus, to some extent, improved some classification model performances that are shown in table 3, when we had 2390 annotated sentences with 625 data sentences.

5.2.3 Linear Regressions on Standard Deviations

We removed some annotated samples due to their language and duplicate papers. The linear regression analysis is carried out using 1,868 annotated samples in order to explore

precision	recall	f1-score	model
0.518519	0.0672	0.118980	Logistic_regression (cv = 5)
0.473282	0.1984	0.279594	Ensemble (cv = 5)
0.468468	0.2496	0.325678	Random_forest (cv = 5)
0.463964	0.3296	0.385407	KNN (cv = 5)
0.395238	0.3984	0.396813	Decision_tree (cv = 5)
0.293527	0.9360	0.446906	Bayes (cv = 5)
0.000000	0.0000	0.000000	SVC_poly (cv = 5)
0.000000	0.0000	0.000000	Neural_network (cv = 5)
0.000000	0.0000	0.000000	SVC_linear (cv = 5)

Table 3: Classification Model Performances of Annotated Samples (UMAP & WoS data & word2vec model)

the relationship between each sentence and the abstract and title of the paper. For convenience, *the content* refers to the combination of a paper's abstract and the title, while *the rest content* against a sentence indicates the content excluding this sentence. We regressed the text variety of the rest content against the text variety of the sentence. A text's text variety is determined by its word2vec loading standard deviation. The data sentence corpus, the non-data sentence corpus, and the whole corpus were all explored. Figure 2 visualized all the data points, and table 4 displays the regression results for the data and non-data corpus separately.

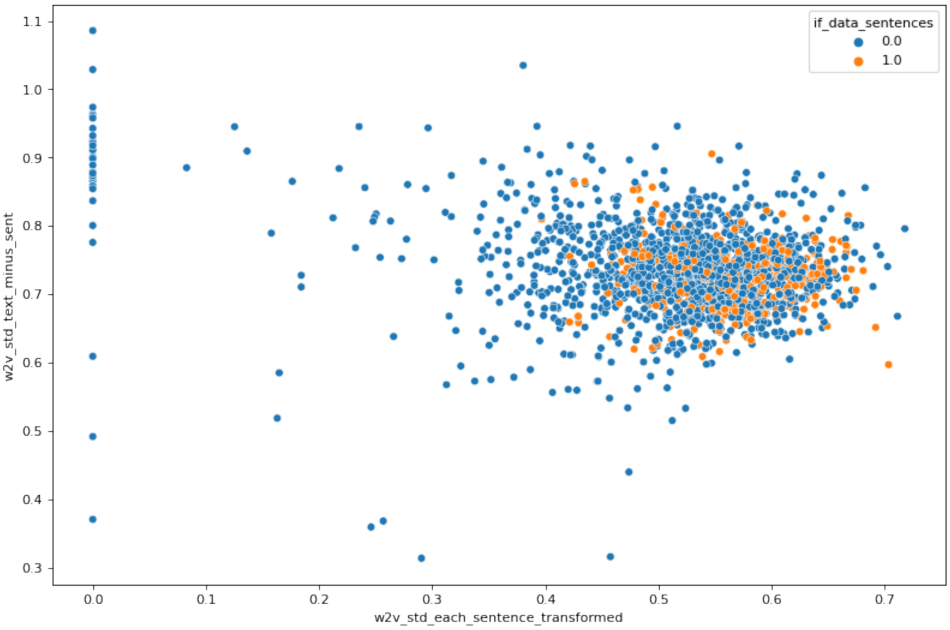

Figure 2: Word2vec Loading Standard Deviation of 1,868 Annotated Samples

Variable	**Data Corpus Model**	**Non-data Corpus Model**
Intercept	0.7589***	0.8117***
	(0.000)	(0.000)
Sentence Text Variety	-0.0588	-0.1495***
	(0.167)	(0.000)
N	407	1461
R^2	0.005	0.058
F-statistic	1.921	90.21
Log-Likelihood	677.26	1868.2
AIC	-1351.	-3732.

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 4: Regression Results for 1,868 Annotated Sentences

Despite the fact that $R * 2$ is not very high in either of the two regressions, we noticed that the coefficients are significant. Additionally, there is a negative correlation between the word2vec loading standard deviation of the sentence and its content. Nevertheless, the coefficients of the independent variable in the two regressions do not vary very much. Both coefficients are close to 0.

Additionally, we re-trained the classification models with five cross-validation folds, with the standard deviation as the independent variable while the dummy variable indicating whether the sentence was a data sentence or not acted as the dependent variable. The results of the data sentence part with 407 samples are shown in table 5. We saw a substantial performance improvement compared with the statistics in tables 2 and 3. In terms of precision, the trained random forest model outperforms other models, while the trained Naive Bayes model shows the greatest recall. We used both models to predict whether a sentence contains data information or not. For each group of predicted data, non-data, and all of the sentences, a linear regression model was trained using the standard deviation of each sentence as the independent variable.

Figures 3 and 4 show the predicted results by the two re-trained classification models. The details of regression results are shown in table 6 and 7. One interesting phenomenon is that we have 775 data sentences by prediction using the Naive Bayes model. However, there are only 64 data sentences by prediction given the random forest model. It is

Precision	Recall	F1-score	Model
0.759259	0.100737	0.177874	Random_forest
0.696429	0.191646	0.300578	SVC_poly
0.631579	0.235872	0.343470	Ensemble
0.498498	0.407862	0.448649	Neural_network
0.473684	0.221130	0.301508	Logistic_regression
0.449275	0.076167	0.130252	SVC_linear
0.408824	0.341523	0.372155	KNN
0.360000	0.685504	0.472081	Bayes
0.330120	0.336609	0.333333	Decision_tree

Table 5: Classification Model Performances of 407 Annotated Data Sentences (WoS data & word2vec model)

understandable since the Naive Bayes model aims to assign more sentences to be data ones based on the features of the actual data sentences, while the Random Forest model seeks to ensure that more predicted data sentences are data sentences in fact.

The coefficients in both models regressing against sentence text variety in the data and non-data corpus are significant, among which only the one in the data corpus model using the Naive Bayes model is positive. The positive coefficient indicates that the more diverse the sentence, the more diverse the rest of the text, while the negative coefficient supports the converse. The sharp difference in the number of predicted data sentences using the two models is believed to contribute to this phenomenon. It is also comprehensible that there will be a shift from a negative relationship between the sentence variety and the rest of the content variety. This is because an increasing number of sentences in the content are designated as data ones. This transition illuminated us that it is possible that data sentences have coherent and common writing styles. Specifically, when datasets are mentioned or introduced in the paper, it is highly possible that the expressions are similar among various authors and papers. Our finding emphasizes the importance of combining topics and text variety together in order to separate data sentences from non-data sentences. This led to our endeavor in Section 5.2.4.

Our findings suggest that data sentences and non-data sentences are not equal in length based on the p-value, whether with the actual or predicted labels by the fitted random

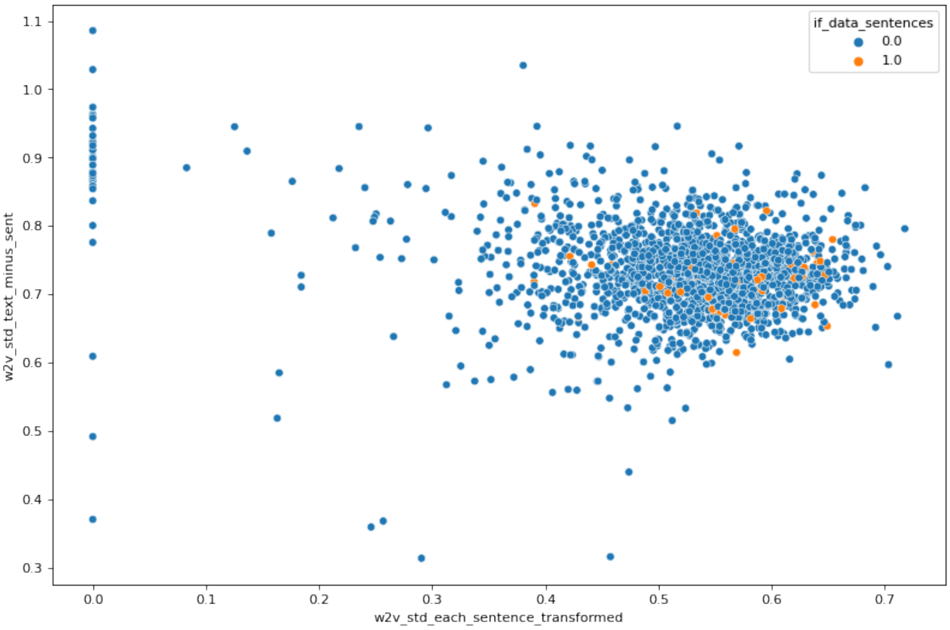

Figure 3: Word2vec Loading Standard Deviation of Annotated Sentences (Classified by Random Forest)

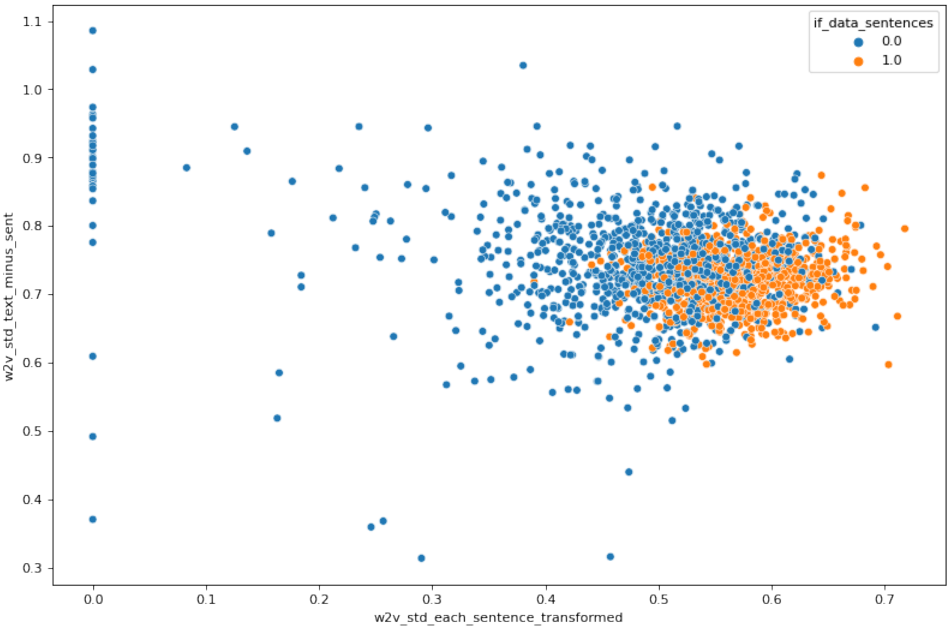

Figure 4: Word2vec Loading Standard Deviation of Annotated Sentences (Classified by Naive Bayes)

forest model showing whether a sentence is data or not. An alternative to using all words from a sentence or a complete paper was to extract word pairs from the sentences and the content of a paper. This analysis was undertaken in order to determine if there was a relationship between the word2vec loading standard deviations and the corpus, where sentences were assumed to be data sentences for the trained random forest model. The

Variable	Data Corpus Model	Non-data Corpus Model
Intercept	0.8379***	0.8091***
	(0.000)	(0.000)
Sentence Text Variety	-0.1997**	-0.1451***
	(0.018)	(0.000)
N	64	1804
R^2	0.088	0.054
F-statistic	5.952	103.5
Log-Likelihood	114.45	2399.1
AIC	-224.9	-4794.

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 6: Regression Results for 1,868 Annotated Sentences (Using Predicted Labels by Random Forest)

Variable	Data Corpus Model	Non-data Corpus Model
Intercept	0.6594***	0.8129***
	(0.000)	(0.000)
Sentence Text Variety	0.1041***	-0.1432***
	(0.001)	(0.000)
N	775	1093
R^2	0.015	0.048
F-statistic	11.42	55.15
Log-Likelihood	1375.9	1291.7
AIC	-2748.	-2579.

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 7: Regression Results for 1,868 Annotated Sentences (Using Predicted Labels by Naive Bayes)

positive correlation between the standard deviation of the sentence and the rest of the corpus was confirmed once again. We tried various thresholds for training the random forest classifier ranging from 0.3, 0.4, 0.45, to 0.5. However, this time we observed that the lengths of predicted data sentences and non-data sentences were equally long, while its relationship with the rest of the corpus was predicted to be positive. We also requested eight-word samples from the text of a sentence and 16 samples from the rest of the corpus, assuming that the rest of the paragraph is longer than the text of the specific sentence. The standard deviations of the word2ve loadings were calculated for each group of word samples, and these standard deviations were regressed against the standard deviation of the eight-word samples as indicated in figure 5. In the annotated samples used to conduct this regression, there are 1435 sentences, 51 data sentences, and 1384 non-data sentences. Rest of the sentences are filtered since they are not long enough to construct word samples. The regression results are shown in table 8. The coefficient of sentence diversity is positive and significant in non-data corpus, which suggests that the more diverse a non-data sentence is, the more diverse the rest paragraph will be. Despite the fact that the coefficient of sentence diversity is positive in the data corpus, it is not significant according to the p-value. We are therefore cautious in our assessment of the relationship between the text diversity of data sentences and the rest of the content. This limits the likelihood of detecting data sentences solely based on text diversity. Given that we only used less than 2,000 sentences in the article, we do not dispute this correlation. With an expanding sample size, we may be able to re-run linear regressions and find that the coefficient becomes significant.

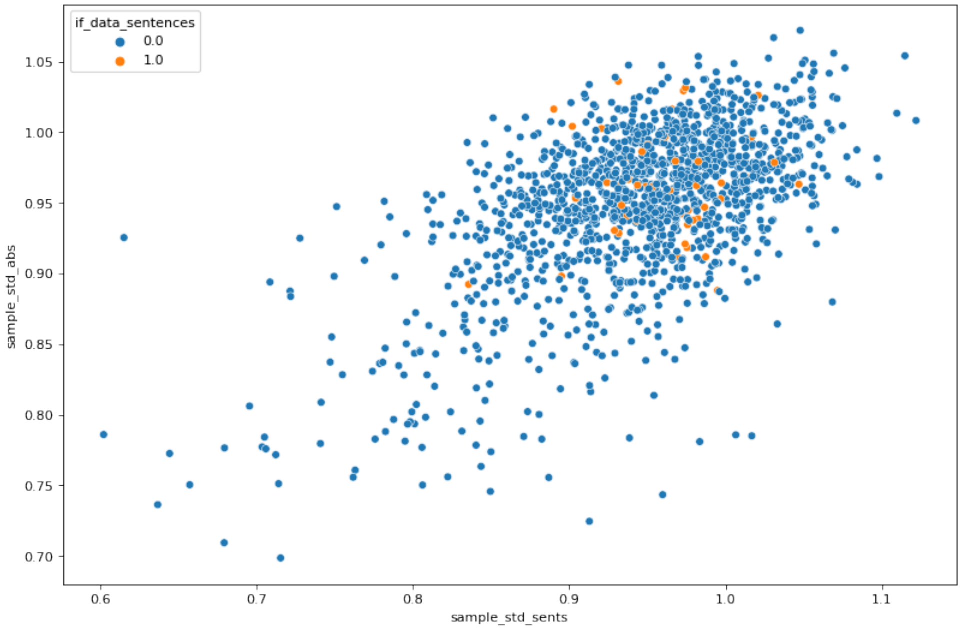

Figure 5: Word2vec Loading Standard Deviation of Sample words in Annotated Sentences (Classified by Random Forest)

5.2.4 Linear Regressions of Citation on Text Diversity

In light of the regression results, we argue that sentence diversity may contribute to filtering sentences with data information. Consequently, this part of the analysis is also

Variable	Data Corpus Model	Non-Data Corpus Model
Intercept	0.8650***	0.5075***
	(0.000)	(0.000)
Sampled Words Diversity	0.1010	0.4682***
	(0.449)	(0.000)
N	51	1384
R^2	0.012	0.343
F-statistic	0.5823	721.7
Log-Likelihood	98.395	2327.4
AIC	-192.8	-4651.

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 8: Regression Results for Sample Words in 1,435 Annotated Sentences

expected to provide insight into how text diversity can impact the impact of a paper. In combining these two relationships, we can make inferences regarding the relationship between dataset information and the impact of a paper, with text diversity acting as a bridge relating the two. Thus, once we have the citation to a paper, we are able to grasp more aspects of data sentences. To be detailed, we designed three diversity measurements in this section, which are inspired by the text variety utilized in Section 5.2.3 and as follows:

1. For the corpus (abstract plus title) of each paper, we **sampled 3 or 4 sentence loadings** which are defined by the word2vec loading centroid and calculate its max inner product. (The number of samples depends on the length of our sentences. In this case study, we tried to sample 3 sentence loadings for each observation.)

2. For the corpus (abstract plus title) of each paper, we **do not sample but use all the sentence loadings**, and calculate its max inner product.

3. For each paper, we calculated the abstract cosine distance between each sentence loading and the corpus centroid and pick the biggest value.

In terms of the calculation of the max inner product, suppose we have three vectors,

$[a_1, a_2, a_3]^T$, $[b_1, b_2, b_3]^T$ and $[c_1, c_2, c_3]^T$.

$$\text{Max inner product} = max(\mid a_1 \cdot b_1 \cdot c_1 \mid, \mid a_2 \cdot b_2 \cdot c_2 \mid, \mid a_3 \cdot b_3 \cdot c_3 \mid) \tag{4}$$

We regressed the times that a paper is cited against each of the three measurements, and figures 6, 7 and 8 display the data points. Despite the positive coefficients of the three independent variables, there is no significant coefficient for the exhaustive inner product values. Particularly, the greater the inner product of the sampled sentences loadings or the greater the abstract cosine distance between the sentence loading and the paragraph centroid, the greater the likelyhood of citation of the paper. In Section 5.2.3, although we were unable to conclude that there is a relationship between the text diversity of a data sentence and the rest of the content, we note that text diversity affects the paper impact.

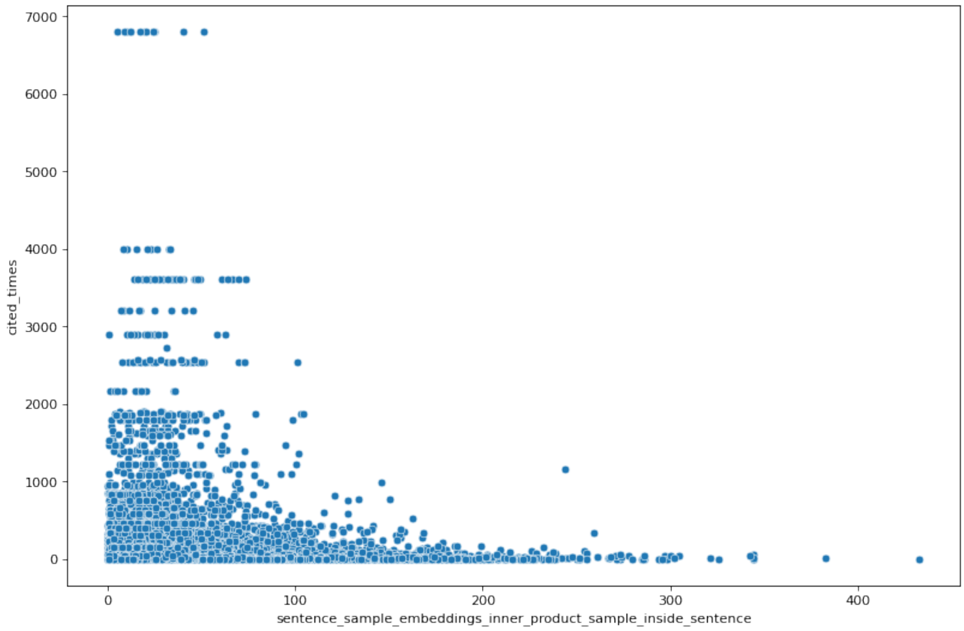

Figure 6: Citation Times and Sampled Max Inner Product of 135,404 Sentences

In addition, we divided the corpus into two parts: sentences with a paper citation in the top 10% and the rest. We applied three logistic regression models on the three diversity measurements and the dummy variable separately. It surprised us that the coefficients of the three independent variables are all negative and significant. This indicates that

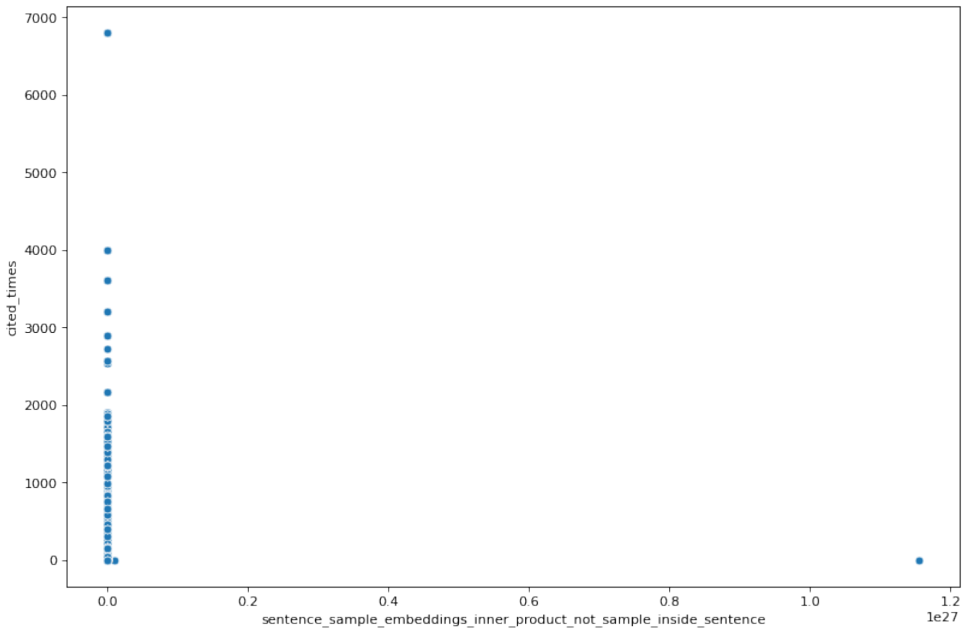

Figure 7: Citation Times and Exhaustive Max Inner Product of 135,404 Sentences

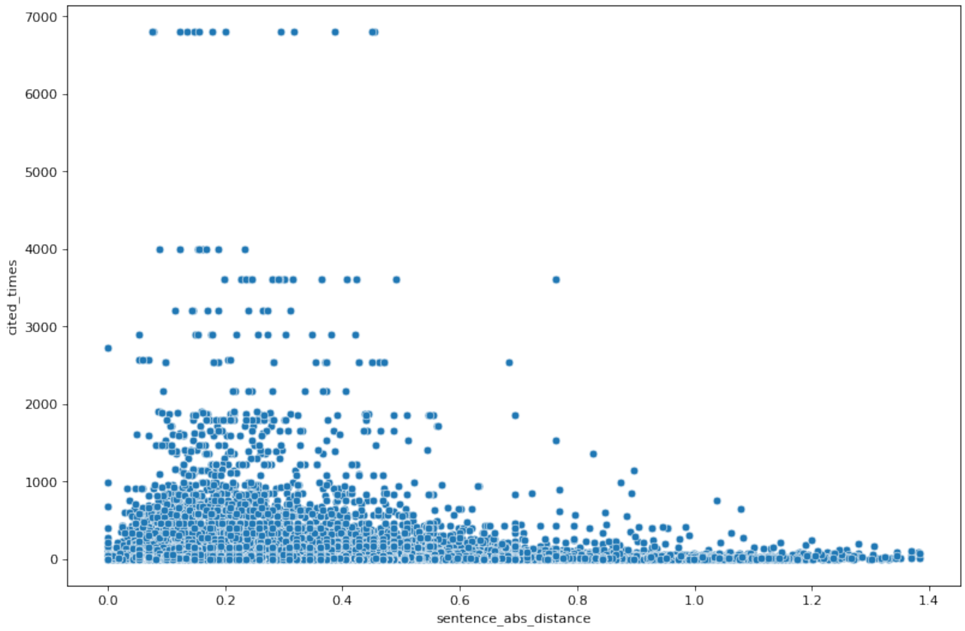

Figure 8: Citation Times and Abstract Cosine Distance of 135,404 Sentences

the more diverse the text is in terms of semantic information, the less likely this paper outranks its contemporaries regarding academic impact measured by citation. In short, the more diverse the text of a paper, the smaller its chances of being among the most influential papers in its field.

Following the linear regressions in the previous section, it appears that increasing text diversity may contribute to *advertising* a paper, but it is highly possible to prevent it from becoming a *peak* among the scholarly works in its field. While we have not found evidence of a significant positive relationship between the diversity of text in a data sentence and the rest of the content of a paper, we do pay attention to this potential relationship and should be able to test it with larger datasets.

6 Discussion & Conclusion

This project aims to provide a pipeline enabling academics to extract dataset information from scholarly publications concentrating on certain themes in the literature. As an example study, we investigated how to extract data sentences from epidemiological research about COVID-19 against clinical articles using LDA and adjusting the threshold of most significant topic probability based on the specific corpus in the study. We have a sufficient proportion of epidemiological articles among the anticipated 78,340 papers. Furthermore, although TF-IDF, singly or combined with or WordNet, fails to provide data-related vocabulary based on the following analysis, the LDA model gave us specific topics, especially when the data size grows. LDA outperformed in filtering epidemiological papers and providing reference to the possible elements contained in data sentences. The manual check revealed that there is a high proportion of epidemiological papers among the 78,340 titles filtered by LDA according to the topic probability distribution of the sentence. Nevertheless, we question whether TF-IDF fails due to the limited sample size, which is also a problem possibly in tasks that use only one discourse atom to identify the data sentence cluster in the discourse atom model. This may lead to the creation of more annotated samples with labels indicating whether they include dataset information. Our McFadden's pseudo-R-squared reached 0.558 using the combination of discourse atoms in our logistic regressions that predict whether a sentence is a data sentence or not.

While we chose LDA results as the primary measurement to filter epidemiological papers against word2vec and doc2vec embeddings, word2vec representations produced meaningful results in regression and classification tasks. With dimensions reduced from 100 to 30 using UMAP, word2vec centroid loadings of the annotated samples passed classification algorithms with five cross-validation folds in predicting whether the sentence is a data one or not with high precision and recall, where the random forest model's precision score reached 0.4545, and the naive Bayes' model's recall score reached 0.8847. When using the re-trained classification models obtained from approximately 1 million WoS papers' title and abstract, the performance of the "**pre-trained**" word2vec models improved obviously. In contrast, the logistic regression model reached 0.5185 in precision. The naive Bayes model achieved 0.9360 in recall even when the data size and data sentence number nearly doubled. We gave credit to the pre-trained model, which shed light on the possibility of using a word representation model with a larger corpus in the future. The satisfactory performances of classification models on annotated samples with pre-trained word representations gave us the confidence to generalize them to an unsupervised dataset. We will pay more attention to the random forest and the naive Bayes algorithms in future computation, considering their predominating fulfillment at the current stage.

In addition, we investigated the relationships between text diversity and whether a sentence is a data one, and between text diversity and the paper's impact, expecting text diversity to act as a *bridge* to link data sentences and the paper's impact. The motivation is that the paper's impact can be evaluated by its citation times and the information is easy to acquire thanks to existing databases like Microsoft Academic Graph Database. In the meantime, our intuition is that different datasets do affect whether a paper produces impressive and noteworthy results. Therefore, this section discussed the possibility of detecting data sentences with the paper's impact taken into account. In Section 5.2.3, we found that the more scattered words in a sentence, the less scattered words are in the rest of the content when we attempt to achieve higher precision in predicting whether a sentence contains data. In contrast, if we are seeking higher recall, more diverse data sentences will be accompanied by more diverse sentences in the rest of the content.

We see the same positive relationship when we use the actual labels for the independent and dependent variables. As a consequence of this finding, we investigated the correlation of the word diversity of a sentence to the content of a paper. In addition, we investigated its citation times with three measures of word diversity. No matter which metric we use, it is surprising that a larger diversity brings more citations, while it is difficult for papers to be cited above the top 10% level. In line with this, we considered the possibility of using the data sentences for each paper to calculate the likelihood that it will receive citations compared to other papers. By creating accurate techniques for sifting out datasets from such sentences, we might be able to test not only the entire sentences containing data, but also the words embedded within the sentences containing data related to the dataset. In addition, it appears that higher text diversity in the content of the paper increases its popularity. However, it does not make it stand out from its peers in terms of citations. Despite this fact, if the corpus can be expanded to include more citations, we may be able to be more confident in the results. Also, in order to obtain data statements in a particular academic field, we need to develop more sophisticated tools that take into account both the context and meaning of words. Relying on these tools, we will further investigate how to filter dataset information given target sentences. Once we can finalize the pipeline of filtering data sentences given specific academic topics, we aim to extract the exact (or most accurate) dataset information using these data sentences. With those datasets, we can create the network where the datasets are the nodes whose edges are verified hypotheses. These graphs are destined to empower us to explore further how various data combinations advance scientific research.

In summary, we recognized LDA to be the most successful algorithm in detecting papers of a certain academic focus (epidemiological papers about COVID-19 against clinical studies) in this case study, while we still retain the expectation for TF-IDF to make its contribution and we hope to test this assumption with larger datasets in the future. Utilizing word2vec vector representations, we discovered that data sentences exhibit several features, including the interaction between information about the data, text diversity, and the impact of the paper. We anticipate that datasets with a larger size will solidify our current findings and assumptions in line with the results shown in filtering epidemiological papers. The first step toward separating data sentences from a paper is to determine the relationship between data information, text diversity, and the paper's impact. This will allow us to extract dataset information from a paper and evaluate its contribution to scientific collaboration and progress.

References

[1] R. Agarwal and V. Dhar, "Big data, data science, and analytics: The opportunity and challenge for is research," 2014.

[2] S. Leonelli, "What difference does quantity make? on the epistemology of big data in biology," *Big data & society*, vol. 1, no. 1, p. 2053951714534395, 2014.

[3] M. Mitsuhiro and T. Hoshino, "Kernel canonical correlation analysis for data combination of multiple-source datasets," *Japanese Journal of Statistics and Data Science*, vol. 3, no. 2, pp. 651–668, 2020.

[4] N. C. D. S. for Adolescent Depression Trials Study Team including:, T. Perrino, G. Howe, A. Sperling, W. Beardslee, I. Sandler, D. Shern, H. Pantin, S. Kaupert, N. Cano *et al.*, "Advancing science through collaborative data sharing and synthesis," *Perspectives on Psychological Science*, vol. 8, no. 4, pp. 433–444, 2013.

[5] J. Brainard, "Scientists are drowning in covid-19 papers. can new tools keep them afloat," *Science*, vol. 13, no. 10.1126, 2020.

[6] E. Garfield, "Citation indexes for science," vol. 122, no. 3159, pp. 108–111, Jul. 1955. [Online]. Available: https://doi.org/10.1126/science.122.3159.108

[7] K. Wang, Z. Shen, C. Huang, C.-H. Wu, Y. Dong, and A. Kanakia, "Microsoft academic graph: When experts are not enough," *Quantitative Science Studies*, vol. 1, no. 1, pp. 396–413, 2020.

[8] K. L. Reed, "Citation analysis of faculty publication: beyond science citation index and social science citation index." *Bulletin of the Medical Library Association*, vol. 83, no. 4, p. 503, 1995.

[9] L. Hou, Y. Pan, and J. J. Zhu, "Impact of scientific, economic, geopolitical, and cultural factors on international research collaboration," *Journal of Informetrics*, vol. 15, no. 3, p. 101194, 2021.

[10] L. Wu, D. Wang, and J. A. Evans, "Large teams develop and small teams disrupt science and technology," *Nature*, vol. 566, no. 7744, pp. 378–382, 2019.

[11] B. Uzzi, S. Mukherjee, M. Stringer, and B. Jones, "Atypical combinations and scientific impact," *Science*, vol. 342, no. 6157, pp. 468–472, 2013.

[12] S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi *et al.*, "Science of science," *Science*, vol. 359, no. 6379, 2018.

[13] L. M. Bettencourt, D. I. Kaiser, and J. Kaur, "Scientific discovery and topological transitions in collaboration networks," *Journal of Informetrics*, vol. 3, no. 3, pp. 210–221, 2009.

[14] D. J. de Solla Price, "Is technology historically independent of science? a study in statistical historiography," *Technology and Culture*, pp. 553–568, 1965.

[15] Y. Lin, J. A. Evans, and L. Wu, "New directions in science emerge from disconnection and discord," 2021.

[16] D. Hicks, P. Wouters, L. Waltman, S. De Rijcke, and I. Rafols, "Bibliometrics: the leiden manifesto for research metrics," *Nature News*, vol. 520, no. 7548, p. 429, 2015.

[17] L. Waltman, C. Calero-Medina, J. Kosten, E. C. Noyons, R. J. Tijssen, N. J. van Eck, T. N. van Leeuwen, A. F. van Raan, M. S. Visser, and P. Wouters, "The leiden ranking 2011/2012: Data collection, indicators, and interpretation," *Journal of the American society for information science and technology*, vol. 63, no. 12, pp. 2419–2432, 2012.

[18] Q. Zhang, J. Abraham, and H.-Z. Fu, "Collaboration and its influence on retraction based on retracted publications during 1978–2017," *Scientometrics*, vol. 125, no. 1, pp. 213–232, 2020.

[19] L. E. Wool and T. I. B. Laboratory, “Knowledge across networks: how to build a global neuroscience collaboration,” *Current Opinion in Neurobiology*, vol. 65, pp. 100–107, 2020.

[20] U. Naseem, I. Razzak, S. K. Khan, and M. Prasad, “A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models,” *Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 5, pp. 1–35, 2021.

[21] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, “Learning word vectors for sentiment analysis,” in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 2011, pp. 142–150.

[22] E. Haddi, X. Liu, and Y. Shi, “The role of text pre-processing in sentiment analysis,” *Procedia Computer Science*, vol. 17, pp. 26–32, 2013.

[23] N. Babanejad, A. Agrawal, A. An, and M. Papagelis, “A comprehensive analysis of preprocessing for word representation learning in affective tasks,” in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 5799–5810.

[24] M. Naili, A. H. Chaibi, and H. H. B. Ghezala, “Comparative study of word embedding methods in topic segmentation,” *Procedia computer science*, vol. 112, pp. 340–349, 2017.

[25] S. Li, T.-S. Chua, J. Zhu, and C. Miao, “Generative topic embedding: a continuous representation of documents,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 666–675.

[26] C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma, “Topic modeling for short texts with auxiliary word embeddings,” in *Proceedings of the 39th International ACM*

SIGIR conference on Research and Development in Information Retrieval, 2016, pp. 165–174.

[27] Y. Ren, R. Wang, and D. Ji, "A topic-enhanced word embedding for twitter sentiment classification," *Information Sciences*, vol. 369, pp. 188–198, 2016.

[28] M. Bahgat, S. Wilson, and W. Magdy, "Towards using word embedding vector space for better cohort analysis," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 919–923.

[29] A. Greiner-Petter, A. Youssef, T. Ruas, B. R. Miller, M. Schubotz, A. Aizawa, and B. Gipp, "Math-word embedding in math search and semantic extraction," *Scientometrics*, vol. 125, no. 3, pp. 3017–3046, 2020.

[30] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[31] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. M. Kinney, Z. Liu, W. Merrill, P. Mooney, D. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, C. Wilhelm, B. Xie, D. A. Raymond, D. S. Weld, O. Etzioni, and S. Kohlmeier, "Cord-19: The covid-19 open research dataset," *ArXiv*, 2020.

[32] K. Kowsari, D. E. Brown, M. Heidarysafa, K. J. Meimandi, M. S. Gerber, and L. E. Barnes, "Hdltex: Hierarchical deep learning for text classification," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec 2017, pp. 364–371.

[33] C. Analytics, "Web of science," 2017.

[34] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. Hsu, and K. Wang, "An overview of microsoft academic service (mas) and applications," in *Proceedings of the 24th*

international conference on world wide web, 2015, pp. 243–246.

[35] K. Wang, Z. Shen, C. Huang, C.-H. Wu, D. Eide, Y. Dong, J. Qian, A. Kanakia, A. Chen, and R. Rogahn, "A review of microsoft academic services for science of science studies," *Frontiers in Big Data*, vol. 2, p. 45, 2019.

[36] C. Chen, "A glimpse of the first eight months of the covid-19 literature on microsoft academic graph: Themes, citation contexts, and uncertainties," *Frontiers in research metrics and analytics*, vol. 5, p. 24, 2020.

[37] H. Qin, J. Zeng, and X. Ma, "Trend analysis of research direction in computer science based on microsoft academic graph," in *The 2nd International Conference on Computing and Data Science*, 2021, pp. 1–4.

[38] S. Angioni, A. A. Salatino, F. Osborne, D. R. Recupero, and E. Motta, "Integrating knowledge graphs for analysing academia and industry dynamics," in *ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium.* Springer, 2020, pp. 219–225.

[39] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242, no. 1. Citeseer, 2003, pp. 29–48.

[40] S. Qaiser and R. Ali, "Text mining: use of tf-idf to examine the relevance of words to documents," *International Journal of Computer Applications*, vol. 181, no. 1, pp. 25–29, 2018.

[41] A. Aizawa, "An information-theoretic perspective of tf–idf measures," *Information Processing & Management*, vol. 39, no. 1, pp. 45–65, 2003.

[42] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of tf* idf, lsi and multi-words for text classification," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2758–2765, 2011.

[43] A. Arseniev-Koehler, S. Cochran, V. Mays, K.-W. Chang, and J. G. Foster, "Discourses of death: Extracting the semantic structure of lethal violence with machine learning," 2020.

[44] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[45] D. M. Blei and J. D. Lafferty, "A correlated topic model of science," *The annals of applied statistics*, vol. 1, no. 1, pp. 17–35, 2007.

[46] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[47] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 113–120.

[48] C. Wang, D. Blei, and D. Heckerman, "Continuous time dynamic topic models," *arXiv preprint arXiv:1206.3298*, 2012.

[49] T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum, "Integrating topics and syntax." in *NIPS*, vol. 4, 2004, pp. 537–544.

[50] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[51] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.

[52] J. Turian, L. Ratinov, and Y. Bengio, "Word representations: a simple and general method for semi-supervised learning," in *Proceedings of the 48th annual meeting of the association for computational linguistics*, 2010, pp. 384–394.

[53] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.* Valletta, Malta: ELRA, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[54] Y. Goldberg and O. Levy, "word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method," *arXiv preprint arXiv:1402.3722*, 2014.

[55] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[56] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning.* PMLR, 2014, pp. 1188–1196.

[57] J. H. Lau and T. Baldwin, "An empirical evaluation of doc2vec with practical insights into document embedding generation," *arXiv preprint arXiv:1607.05368*, 2016.

[58] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Science China Technological Sciences*, pp. 1–26, 2020.

[59] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[60] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural*

language processing (EMNLP), 2014, pp. 1532–1543.

[61] F. Chollet *et al.*, “Keras,” https://keras.io, 2015.

[62] Y. Qi, D. S. Sachan, M. Felix, S. J. Padmanabhan, and G. Neubig, “When and why are pre-trained word embeddings useful for neural machine translation?” *arXiv preprint arXiv:1804.06323*, 2018.

[63] I. Tenney, D. Das, and E. Pavlick, “Bert rediscovers the classical nlp pipeline,” *arXiv preprint arXiv:1905.05950*, 2019.

[64] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, “Flair: An easy-to-use framework for state-of-the-art nlp,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 2019, pp. 54–59.

[65] S. Weisberg, *Applied linear regression.* John Wiley & Sons, 2005, vol. 528.

[66] S. Russell and P. Norvig, “Artificial intelligence: a modern approach,” 2002.

[67] A. McCallum, K. Nigam *et al.*, “A comparison of event models for naive bayes text classification,” in *AAAI-98 workshop on learning for text categorization*, vol. 752, no. 1. Citeseer, 1998, pp. 41–48.

[68] J. Novakovic and A. Veljovic, “C-support vector classification: Selection of kernel and parameters in medical diagnosis,” in *2011 IEEE 9th International Symposium on Intelligent Systems and Informatics*, 2011, pp. 465–470.

[69] L. E. Peterson, “K-nearest neighbor,” *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.

[70] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression.* John Wiley & Sons, 2013, vol. 398.

[71] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE transactions on systems, man, and cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

[72] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[73] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the national academy of sciences*, vol. 79, no. 8, pp. 2554–2558, 1982.

[74] T. Hastie, R. Tibshirani, and J. Friedman, "Boosting and additive trees," in *The elements of statistical learning.* Springer, 2009, pp. 337–387.

[75] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[76] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[77] S. Seabold and J. Perktold, "statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, 2010.

[78] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation." *Encyclopedia of database systems*, vol. 5, pp. 532–538, 2009.

[79] M. A. Carreira-Perpinán, "A review of dimension reduction techniques," *Department of Computer Science. University of Sheffield. Tech. Rep. CS-96-09*, vol. 9, pp. 1–69, 1997.

[80] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *ArXiv e-prints*, Feb. 2018.

[81] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell, "Dimensionality reduction for visualizing single-cell data using umap," *Nature biotechnology*, vol. 37, no. 1, pp. 38–44, 2019.

[82] L. McInnes, J. Healy, N. Saul, and L. Grossberger, "Umap: Uniform manifold approximation and projection," *The Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018.

[83] C. Fellbaum, "Wordnet," in *Theory and applications of ontology: computer applications*. Springer, 2010, pp. 231–243.

[84] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[85] "Wordnet," https://wordnet.princeton.edu/, accessed: 2021-07-06.