PACIS 2022 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

7-4-2022

Understanding the Role of Document Representations in Similarity Measurement in Finance and Accounting

Steffen Bankamp
University of Goettingen, steffen.bankamp@uni-goettingen.de

Jan Muntermann
University of Augsburg, muntermann@wiwi.uni-goettingen.de

Understanding the Role of Document Representations in Similarity Measurement in Finance and Accounting

Completed Research Paper

Steffen Bankamp
University of Goettingen
Goettingen, Germany
steffen.bankamp@uni-goettingen.de

Jan Muntermann
University of Augsburg
Augsburg, Germany
jan.muntermann@uni-a.de

Abstract

Document similarity is an important concept for many research questions. It can be applied to trace information exchanged on the capital market. For similarity calculations, the document must be transformed into a vector (document representation). Researchers can choose from a variety of document representations. We review the finance and accounting literature and find many different practices for estimating document similarity but little guidance on how to choose the right approach. To address this gap, we propose a framework of three similarity dimensions (object, author, and time). Based on this framework, we conduct an experiment on a corpus of analyst reports to quantify the accuracy of the estimated similarity. Our results help researchers and practitioners to choose an appropriate document representation for their analysis. Doc2vec achieves the overall highest accuracy, while Latent Dirichlet Allocation performs well on the object dimension. Bag-of-words models achieve surprisingly promising results despite their simplicity.

Keywords: Document Representation, Similarity, Finance, Accounting

Introduction

Applying methods from natural language processing (NLP) in the finance and accounting domain is relatively new (Loughran and McDonald 2016). In contrast to standard statistical methods applied to structured data, hardly any conventions have been developed. At the same time, the analysis of numerous text documents in finance and accounting offers much potential to answer challenging and intriguing research questions (Loughran and McDonald 2016). The literature that has applied text mining methods has largely focused on sentiment analysis, while the application of document similarity remains underexplored in the finance and accounting domain (Loughran and McDonald 2016). However, document similarity can be a promising measure for important constructs, as it provides a direct measure of the informativeness of financial disclosures (e.g., Hanley and Hoberg 2010). It also allows for assessment of the time-varying association between companies (e.g., from annual reports) that extends beyond classical industry classification (Hoberg and Phillips 2016). Hence, overall document similarity is an important measure for constructs of interest and thus to test theories in the finance and accounting domain and beyond. Also for practitioners (e.g., asset managers or auditors), document similarity might help to filter for those documents containing new information and thus reduce the problem of information overload. However, before calculating these document similarities, the text must be transformed into a document representation, representing the document as a numeric vector of a specific length. Researchers and practitioners alike are faced with a large number of document representations from which to choose. Considering the early stage in the application of these methods in the finance and accounting domain, it is not surprising that hardly any conventions or best practices have been developed in the selection of document representations. Thus, little guidance is provided to authors faced with this decision. Only in Rawte et al. (2021), a comparison of different methods for text similarity estimations in the finance and accounting domain could be found. However, the authors only consider the temporal dimension of similarity and do not compare the accuracy of different representations for document similarity estimations. Knowing the accuracy of their methods is crucial for researchers and practitioners to make an informed decision. We address this issue and provide holistic guidance in this challenging selection process by answering the following research question:

RQ: *How should one choose document representations based on the dimension of similarity to capture?*

To answer this question, we first develop a framework encompassing three essential dimensions of similarity (*object*, *author,* and *time*). Based on previous research in the finance and accounting domain, we show that this framework is suitable for classifying the existing research. An experiment is conducted by analyzing over 200,000 financial documents (analyst reports) to answer the research question. We contribute to existing literature by providing methodological guidance to researchers and practitioners who want to calculate similarities between finance-related documents. This should support the selection of suitable methods. Our results suggest that irrespective of the utilized document representations, the object dimension is the most accurate and the temporal dimension the most difficult one to capture. Among the representations, doc2vec proves to be highly accurate across all dimensions. Surprisingly, simple methods such as term frequency (TF) and term frequency–inverse document frequency (TF-IDF) achieve promising results. In addition, the topic model Latent Dirichlet Allocation (LDA) provides high accuracy in detecting documents about similar companies (object dimension).

Theoretical Background

Document Similarity

The automated calculation of document similarity (sometimes also referred to as text similarity) is an important task within NLP (Shahmirzadi et al. 2019). The information generated from this calculation can be used in many applications, such as search engines (Pradhan et al. 2015) or question–answer matching (Tan et al. 2016). Document similarity might be applied as an operationalization for different constructs. Bär et al. (2011) argue that text similarity lacks a precise definition and that it is necessary to define what to measure with text similarity. The authors suggest three dimensions of text similarity: *structure*, *style,* and *content*. While we agree with Bär et al. (2011) that the question of similarity cannot be answered without a precise definition, the application of their dimensions requires manual coding to evaluate different methods of text similarity. To overcome this issue, we propose an approach where attributes for each dimension can be drawn from the metadata of many texts in the finance and accounting domain and beyond. We propose the dimensions *author*, *time,* and *object*, as illustrated in Figure 1. This allows for an evaluation of document representation for similarity measurement without manual labeling and thus an evaluation on a much larger dataset. In addition, the results are independent of human judgment, which would not be the case with a manually labeled dataset. It should be noted that the dimensions of Bär et al. (2011) and the three dimensions applied in this paper are related. The *author* dimension might capture elements of *style* and *structure*. This dimension can also be linked to the problem of author identification that is intensively discussed in the literature (e.g., Madigan et al. 2005). The *time* and *object* dimensions might relate to the *content* dimension of Bär et al. (2011). In our review of finance and accounting literature, we show that the dimensions we propose are actually present in this domain (see Table 1).

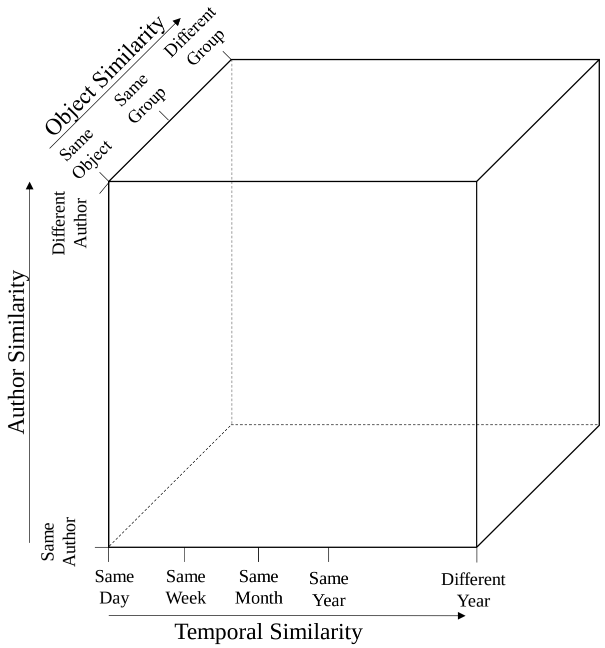

Figure 1. Similarity Cube

Our similarity framework is illustrated by a simple example. Consider the market for scientific textbooks. For each book, the three metadata characteristics discussed above should be available. If a retailer wants to suggest another book to a reader based on a book previously bought, the retailer could recommend books covering a similar object. In this case, the retailer must calculate the object similarity to all other books. However, the reader may not want to read any more books about this topic but likes the properties of the author (e.g., writing style) and would thus be interested in author similarity. Retailers that want to recommend the purchase of a book's new edition to a customer might be interested in temporal similarity. This example shows that not all similarities are the same, and substantial differences exist between the dimensions of similarity. At the same time, this also means that there is not necessarily a single operationalization to provide the best result for all these different constructs.

The technical process of calculating similarity can be divided into three steps, as illustrated in Figure 2. In the first step, the texts are pre-processed; this includes, for example, the transformation of the characters into lower case and stemming or lemmatization to reduce differently conjugated words or plurals to their basic form. The optimal extent of pre-processing can vary and depends on the dataset and the downstream analysis (Naseem et al. 2020). After pre-processing, the text is transformed into vectors, the so-called document representation. This transformation can be done using numerous different methods, which are discussed in detail later. Aside from their differences, the texts are always transformed into vectors of the same length. TF, for example, produces large and sparse vectors (several thousand elements), while document embeddings are dense and relatively small (usually 300–600 elements). The uniform shape of the vectors is necessary to compare the vectors of two documents. For this comparison, a similarity measure is applied that calculates the distance between the two vectors. Many different distance/similarity measures exist, some of which are shown in Figure 2. The result of this process is a numeric value that indicates how similar two texts are. The focus of this work is not on the whole process but on the choice of document representation. We apply the cosine similarity in our experiment, which is the predominantly used measure for text similarity (Singhal 2001). It should be noted that the choice of the similarity measure is less important, as Huang (2008) found similar results for a variety of similarity measures except for the Euclidian distance, which is less suitable for text mining.

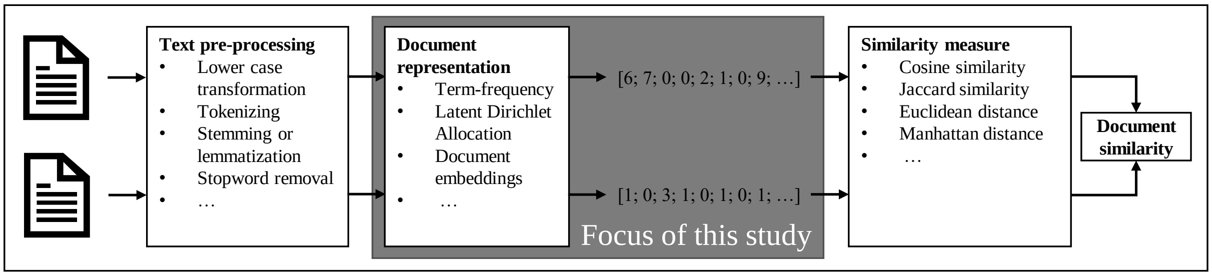

Figure 2. Pipeline for Similarity Calculation of Document Pairs

Document Representations

Bag-of-Words Models

Bag-of-words models are the simplest document representations. The document is represented only by the words it contains, and the order of the words is not considered. The length of the vector corresponds to the size of the corpus' vocabulary. As a result, the vectors can grow to a considerable size. Bag-of-words models also lead to sparse vectors (many elements containing zeros). Each feature represents an individual word. Features can be binary and thus only indicate whether the specific word occurs in the document or they can be a numeric feature (e.g., TF) and thus indicate the number of occurrences of the word within the document (Jurafsky and Martin 2009). An extension to this representation is TF-IDF. With this representation, the TF is normalized by the document frequency, which assigns more weight to words that occur only in a few documents and thus to more document-specific words (Sparck Jones 1972).

Topic Models

Topic models apply dimension reduction of the term-document matrix (TDM). Thus, as in bag-of-words models, the sequence of words is not considered (Blei et al. 2003). One approach to dimension reduction is latent semantic indexing (LSI) or latent semantic analysis (LSA) (often used synonymously). This approach was developed by Deerwester et al. (1990). The feature reduction is conducted by applying singular-value decomposition on the TDM. Words with semantic similarity are placed close to one another in the semantic space and are represented by the same topic (Deerwester et al. 1990). LDA is a probabilistic model that was developed by Blei et al. (2003). Each document is considered as a distribution of topics and each topic in turn as a distribution of words. The advantage of LDA compared to LSA is higher generalizability, easier application to new documents, and less overfitting (Blei et al. 2003). The number of topics and thus the size of the vector constitute a central hyperparameter in both topic models discussed.

Word and Document Embeddings

Word embeddings represent a single word in a dense vector of defined size. The embeddings are derived from large, unlabeled text corpora. The word2vec embeddings developed by Mikolov et al. (2013a) are built either by predicting a current word from its surrounding words (continuous bag-of-words model) or by predicting surrounding words from the current word (continuous skip-gram model). This task is only performed to obtain the weights learned within the neural network for each word in the corpus. These weights represent the word embeddings. Global Vectors for Word Representation (GloVe) is a commonly used word embedding but uses a slightly different architecture. These embeddings are trained on the co-occurrence of words (Pennington et al. 2014). To convert word embeddings into a document representation, they must be aggregated to the document level. One possibility is to average all word embeddings of the words within the document to obtain a document representation (e.g., Mauritz et al. 2021). Kusner et al. (2015) suggest a word mover distance, which defines the sum of the shortest distance between the embeddings of all words from one document to another document. Alternatively, the user can directly apply document embeddings instead of word embeddings. Doc2vec (also known as paragraph vector), developed by Le and Mikolov (2014), is based on word2vec and uses a similar architecture. This method generalizes the word2vec architecture to the document level and directly provides document embeddings.

Transformer-Based Models

The invention of the transformer architecture (Vaswani et al. 2017) has enabled the development of many new language models. These include the bidirectional encoder representations from transformers (BERT) (Devlin et al. 2019) or the universal sentence encoder (USE) (Cer et al. 2018a). These models are much more complex than the previously discussed models and consider the position of words and their context within the document. In addition, these models are typically trained on several tasks, which leads to higher generalizability and state-of-the-art performance (Cer et al. 2018a; Devlin et al. 2019). However, the training of these models is much more complex than the training of the models discussed above.

Document Representation for Similarity Measurement in the Finance and Accounting Literature

Document representations can be applied for various downstream analyses beyond text similarity calculations. Yan et al. (2018), for example, use averaged word embeddings to build a classifier for loan project recommendations on social lending platforms. However, we only focus on literature that specifically applies document similarities. To determine how document representations are used in finance and accounting research for similarity calculations between documents, we analyzed the related literature. The results can be found in Table 1.

The use of document similarity as a variable in accounting and finance research is relatively new and not as established as sentiment analysis (Loughran and McDonald 2016). In their literature review, Loughran and McDonald (2016) identified three papers that use TF or TF-IDF as document representation for similarity calculations. They had already highlighted LSA as a promising document representation but could not identify any paper in the domain at that time. This is in line with our analysis of the literature, as the oldest paper we could identify that uses a document representation other than TF or TF-IDF for text similarity was published in 2018 (see Table 1). It took some time before these more advanced methods found their way into finance and accounting research. Apart from the application for text similarity, some document representations have been used earlier (e.g., Eickhoff and Muntermann 2016).

	Object	Author	Temporal
TF	(Hanley and Hoberg 2010) (Lang and Stice-Lawrence 2015) (Hoberg and Phillips 2016)	(Hanley and Hoberg 2010)	(Hanley and Hoberg 2010)
TF-IDF	(Brown and Knechel 2016)	(Mauritz et al. 2021)	(Brown and Tucker 2011)
LSA/LSI		(Beaupain and Girard 2020)	(Beaupain and Girard 2020)
LDA		(Palmer et al. 2018) (Mauritz et al. 2021)	
Word Embeddings	(Liu et al. 2020)	(Mauritz et al. 2021)	(Liu et al. 2020) (Adosoglou et al. 2021)
Document Embeddings			(Adosoglou et al. 2021)
Transformer	(Chen and Sarkar 2020)		(Chen and Sarkar 2020)

Table 1. Document Representation in Research of the Finance and Accounting Domain

The work of Hanley and Hoberg (2010) is a milestone in the field of financial and accounting document comparisons. The authors examine initial public offering (IPO) prospectuses to determine whether a higher proportion of informative content reduces IPO underpricing. To distinguish informative content from standard phrases, they compare the document similarity between different IPO prospectuses. They demonstrate that IPO prospectuses from the same underwriter (author dimension), from companies in the same industry (object dimension), and that are published in a close temporal context (temporal dimension) are more similar than pairs of different underwriters, industries, or time frames. Thus, all three dimensions of similarity are addressed in this paper. The work of Brown and Tucker (2011) is also among the first in this field. They focus on the temporal dimension to show how the information value of the management discussion and analysis (MD&A) section of an annual report changes over time. Therefore, they compare the similarity between an MD&A section and the same company's MD&A section of the previous year. A high degree of similarity corresponds to low information value. The paper of Mauritz et al. (2021) is interesting, as the authors explicitly choose and justify different document representations for different constructs. They examine the auditor's role in the preparation process of financial reports and compare the financial reports of different companies that are audited by the same auditor. They found a higher degree of document similarity between annual reports audited by the same auditor compared to pairs of reports audited by different auditors. Even though the auditor is not the official author of the annual report, it is clear that Mauritz et al. (2021) focus on the author dimension. Interestingly, Mauritz et al. (2021) use three different document representations to measure different aspects of similarity (TF-IDF for wording similarity and LDA/word embeddings for content similarity). We could not find more complex document representations, such as document embeddings or transformer-based representations, used for similarity measurement in the classical accounting and finance literature. The papers using these advanced methods that we could identify tend to originate from computer science but address a problem from the accounting or finance domain. Adosoglou et al. (2021), for example, propose a system that uses document embeddings (doc2vec) to identify companies with little change in the semantic of their annual report compared to that from the previous year (temporal dimension) and show that abnormal returns can be achieved by investing in these companies. Overall, it can be concluded from the analysis that many different document representations are used in the finance and accounting literature. Our literature review also suggests that researchers measure different dimensions of similarity, sometimes even within the same study. The fact that all papers could be easily grouped into the three similarity dimensions proves the suitability of our proposed framework from Figure 1.

Data

To experimentally investigate the role of document representations for the similarity calculation in the finance and accounting context, we use a large sample of analyst reports from the Thomson ONE database. These are documents published by brokerage houses that analyze a company based on different approaches. Analysts usually provide price or earnings forecasts and make recommendations on whether investors should buy the company's stock (Huang et al. 2018). Analyst reports are suitable for the study because many different authors conduct analyses on the same object. In addition, the reports are published throughout the year rather than only once a year, as is the case with 10-K reports. This allows for the similarity and its dimensions (see Figure 1) to be analyzed as comprehensively as possible.

We choose the constituents of the S&P100 index as a sample. By using a sample of large companies, we ensure sufficient analyst coverage. We choose a relatively long time horizon of 13 years ranging from 01/01/2007 to 12/31/2019. With this time horizon, we can cover many analyst reports and are able to make more generalizable statements. To build our sample, we start with all analyst reports available from Thomson ONE in this period on companies that have been a constituent of the S&P100 within our observation period. We remove automatically generated analyst reports, extremely short reports of less than 300 words, and reports with more than 50 pages (as they are usually industry reports) from the sample. Reports that are written in a language other than English are also dropped from the sample. We then eliminate duplicates, as these would distort the experimental analysis of similarity. This leaves a total of 207,445 analyst reports on 137 companies published by 367 brokerage houses. The analyst reports are available in PDF format, and we perform standard pre-processing steps to make them usable for further analysis – these steps include the elimination of boilerplate, disclaimers, graphs, and tables. We transform the remaining text to lower case, remove stop words, punctuation, and numbers. In addition, we use lemmatization to reduce the words to their original form.

Experimental Design

The experimental design is built upon the similarity cube (see Figure 1). To evaluate how well the different document representations are suited to capture the three similarity dimensions, we keep two dimensions constant while varying the third. We do this by walking along the 12 edges of the similarity cube. The experimental design is illustrated in Figure 3.

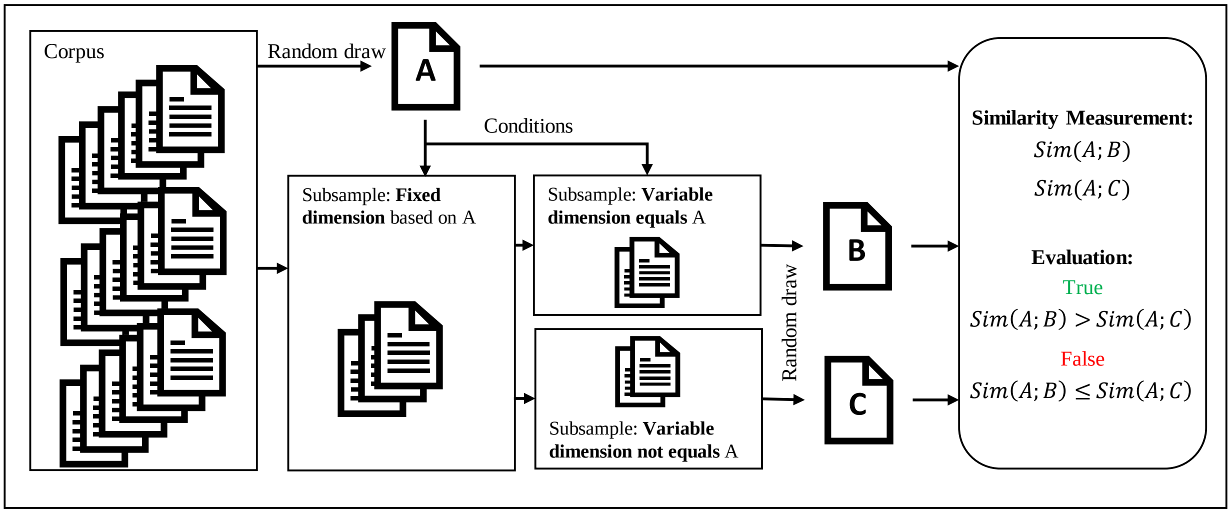

Figure 3. Experimental Design

As mentioned earlier, we avoid manual labeling by deriving the “true” similarity values from the metadata of the analyst reports. For each dimension, we categorize the similarity of report pairs into binary labels (similar vs. not similar). The categorization is presented in Table 2. The cutoffs for the temporal dimensions are derived from the analysts’ observed publication structure. As analyst reports are strongly clustered on few days surrounding the earnings announcement (EA) (Huang et al. 2018), the bound of 14 days, on the one hand, ensures that these reports that are all linked to this date are actually recognized as temporally similar. On the other hand, the bound of 180 days ensures that the reports are at least two quarterly reporting periods apart and thus relate to a completely different point in time. We match groups consisting of three analyst reports. To assess how well the representations capture the object dimension under the condition of different authors (low author similarity) and the same publication time frame (high temporal similarity), we proceed as follows. We first draw a random analyst report from the entire corpus, which we call Report A. Subsequently, another report is drawn from a subset of reports about the same company as that of A (high object similarity) and are published by a different author at the same time relative to A. We call this Report B. Now we draw another report from a subset of reports about different companies than that of A (low object similarity) but with the same relationship to A in the other dimensions as Report B has to A (different object similarity and same temporal similarity). This report is called C. Now we calculate the similarity measures based on the different representations (see Table 3) between A and B and between A and C. If the similarity between A and B is higher than that between A and C, the representation has correctly captured the object similarity and the group is considered to be correctly classified. We form a total of 5,000 of these groups to obtain statically robust results. Reports that have been drawn are not reclined during the repeated drawings for the same combination of dimensions. The process is conducted for all 12 edges of the cube.

Dimension	Similar	Not similar
Object	Same company	Different company
Author	Same broker	Different broker
Temporal	Less than a 14-day difference	More than a 180-day difference

Table 2. Configuration of the Binary Experiment for Similarity Detection

Given the large number of different document representations and the even larger number of variations of these representations, a selection of representations must be considered for the experiment. The selection is based on the four overall classes of representations introduced in the theory section. For each of these classes, we choose at least one representation. The final selection within these classes is then based on the findings of our literature review (see Table 1). Thus, we predominantly select representations that are currently applied in the finance and accounting domain and that are popular among researchers and practitioners.

In total, we use eight different document representations, which are shown in Table 3, including their configuration. This selection covers a broad range of representations. For the transformer-based representation, the BERT model might be the most popular representation, and it is also the model applied in the paper of Chen and Sarkar (2020) (see literature review). However, we only consider representations that are actually capable of fully processing the data used in the experiment, otherwise it would not be possible to determine whether the measured effects are due to the representation itself or due to a partial recognition of data. As the BERT model has a limited input length (Sun et al. 2019), which many analyst reports exceed, we use the universal sentence encoder to represent the class of transformer-based models. This model does not have a constraint on the length of input data. We use a pre-trained model, since the training of such transformer-based models requires extensive computational resources and large text corpora. Furthermore, many researchers who want to apply these models might opt for the pre-trained models because of these demanding requirements. For the word2vec representation, we consider a model that is pre-trained on news articles and a model that we trained on our corpus of analyst reports. We also found both options in the literature we analyzed. Mauritz et al. (2021) use a pre-trained model, while Liu et al. (2020) train the model on their own corpus. All other representations are created based on our corpus of 207,445 analyst reports. The hyperparameters correspond to the typical values found in the literature or the default values of the software packages (see Table 3). Pre-processing is not applied to USE, as the required pre-processing is already built into the model's implementation (Cer et al. 2018a).

Representation	Configuration
TF	max doc frequency: 50% (Şaşmaz and Tek 2021) min doc frequency: 1% (González et al. 2015)
TF-IDF	max doc frequency: 50% (Şaşmaz and Tek 2021) min doc frequency: 1% (González et al. 2015)
LSA	max doc frequency: 50% (Şaşmaz and Tek 2021) min doc frequency: 1% (González et al. 2015) vector size: 100 (Deerwester et al. 1990)
LDA	hyperparameter: default from *mallet* vector size: 100 (Niraula et al. 2013)
Word2vec (news)	trained on news articles vector size: 300 (Jatnika et al. 2019) averaging of word vectors (Mauritz et al. 2021)
Word2vec (own)	hyperparameter: default from *gensim* trained on the corpus of the study vector size: 300 (Jatnika et al. 2019) averaging of word vectors (Mauritz et al. 2021)
Doc2vec	hyperparameter: default from *gensim* trained on the corpus of the study vector size: 300 (Trieu et al. 2017)
USE	TF2.0 Model (v4) no pre-processing (Cer et al. 2018a) vector size: 512 (Cer et al. 2018b)

Table 3. Configuration of Document Representation

Experimental Results

Three Dimensions of Similarity

As illustrated in Figure 4, significant differences exist in the accuracy of similarity recognition depending on the dimension combination and document representation. The x-axes of the bar plots show the configuration of fixed dimensions, and the y-axes indicate the accuracy of similarity recognition – the proportion of groups where $\mathrm{Sim}(A; B) > \mathrm{Sim}(A; C)$. The horizontal dashed-dotted line indicates the expected value for a random estimation, and the error bars show the accuracy's 95% confidence interval.

The object dimension is easily detectable by most of the representations (upper left plot). This is hardly surprising, since the names of companies, products, or the management team provide features that should differentiate well between report pairs of the same vs. different companies. LDA, doc2vec, and TF-IDF perform particularly well. By contrast, word embeddings and the universal sentence encoder perform the worst. These types of models are designed to capture semantics and to generalize very well. However, these properties could also lead to difficulties in capturing the proper nouns described above, resulting in low performance of word embeddings and the universal sentence encoder on this dimension. This is especially the case for the pre-trained models, which may never have made contact with these proper nouns in pre-training.

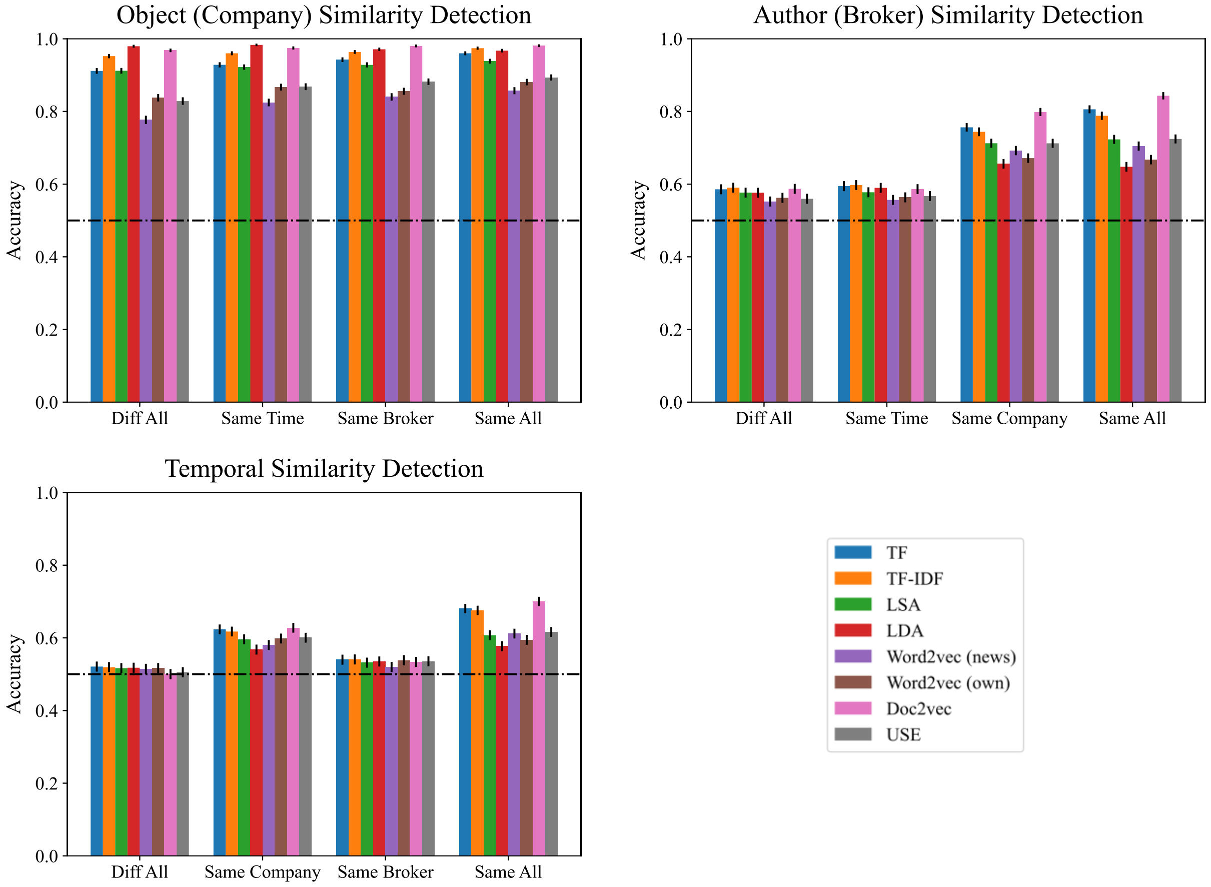

Figure 4. Accuracy of Similarity Estimations Based on Various Document Representations

The author dimension is more difficult to recognize (upper right plot of Figure 4) than the object dimension. Unlike the object dimension, where named entities provide features that contribute to an easier identification (e.g., the name of the company or its representatives that the report was written about), stylistic similarities play a more important role in the author dimension. These stylistic features are harder for the applied representations to capture. The identification of the author dimension is particularly difficult if different companies (objects) are involved. One explanation for this is that analysts working for the brokerage house are typically assigned to specific industries or companies. When examining reports from the same broker regarding the same company, it is more likely that the same analyst has actually written the article compared with report pairs that only share their broker. Doc2vec and the simple bag-of-words models (TF and TF-IDF) perform particularly well on the author dimension.

The temporal similarity dimension is the most difficult to capture. In the case of report pairs from different companies and brokers, the representations are hardly superior to a random estimation. This is not surprising, as only features such as the general economic environment or political decisions could indicate proximity. However, if the same company is considered, an estimation becomes significantly more accurate.

Features that relate to the company's strategic actions are informative here. The sparse representations (TF and TF-IDF) and doc2vec perform well on this task.

Overall, the simplest document representations (TF and TF-IDF) perform relatively well. Doc2vec is one of the best-performing representations across all combinations. However, average word embeddings do not capture similarities particularly well in this experiment. It is also unclear whether a word2vec model trained on the data performs better than a word2vec model pre-trained on a general news corpus. Surprisingly, the modern USE method performs relatively poorly. However, this can be attributed to the fact that it was developed for short texts (especially sentences). Moreover, it was not trained on the corpus. The preceding analysis provides an initial overview of the similarity dimensions and representations. However, the dimensions have only been considered in a binary way (see Table 2). We extend this analysis and investigate the object and temporal dimension in more detail. Therefore, we examine the extent to which gradations within these dimensions can be recognized. For the broker dimension, we refrain from this analysis because no meaningful gradation for the similarity of the authors can be drawn from the available metadata, as a pair of analyst reports is published either by the same or by different brokers.

Detailed Analysis of Object Similarity

Most representations could identify relatively well whether a text pair concerned the same or a different company (see Figure 4). We attribute this to company names and other company-specific proper nouns. For practical applications, however, it is also important to recognize how similar the two companies are. To investigate this, we use The Refinitiv Business Classification (TRBC) to obtain the "true" values for company similarity. This classification consists of five levels: the highest level is the economic sector (e.g., Energy), and the lowest level is the activity (e.g., Wind Systems & Equipment) (Refinitiv 2022). Figure 5 illustrates how the average cosine similarity of pairs changes depending on whether the pairs share the company, the activity, or belong even to different economic sectors. The actual cosine similarity is shown in the left plot, and the *z*-transformed cosine similarity is depicted in the right plot. The *z*-transformation removes the different levels in mean and variance from the cosine similarity of the eight document representations. Each pair consists of reports from two different authors that were published within a time frame of 30 days.

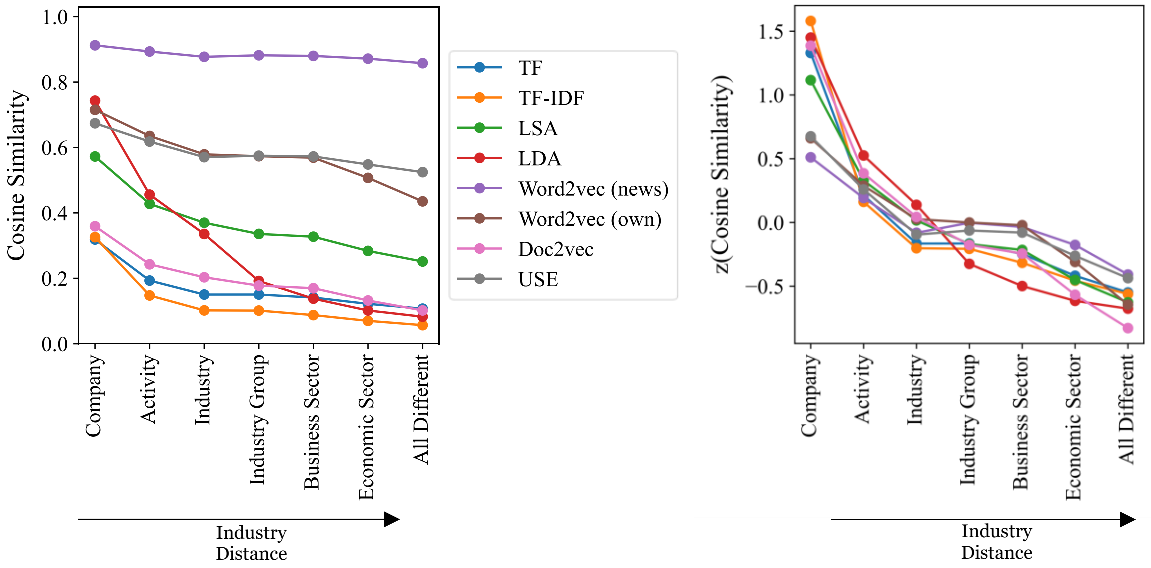

Figure 5. Document Similarity and Industry Distance

For all document representations, we observe a clear decrease in document similarity when the distance of the companies according to the TRBC business classification increases. The sharp increase from report pairs of the same activity to report pairs of the same company can, at least to some extent, be allocated to the proper nouns discussed above. This is also confirmed by the fact that the increase is much smaller in models such as USE or word2vec. LDA shows a strong differentiation between the industry, activity, and company level. This is also a reason why, in some studies, topic models are created for every single industry (Huang et al. 2018) or even for every company (Palmer et al. 2018). This prevents LDA from only representing industries or companies.

To quantify the representations' usefulness for the object dimension, 15,000 groups are formed following the experimental design (see Figure 3). Reports A and B now share their business activity but originate from different companies and the same time frame, and Reports A and C come from different business sectors. The calculation of accuracy is identical to those applied in Figure 4. Furthermore, Figure 6 indicates that LDA and doc2vec are still delivering the best results on the object dimension; however, the level is significantly reduced. Whereas LDA detected reports of the same vs. different companies (fixed dimensions: different author; same time frame) in 98% of cases (see Figure 4), report pairs on companies with the same activity are only detected correctly in 85.55% of cases (see Figure 6). Doc2vec achieves slightly but significantly better results (87.39%). Finally, TF-IDF is the third-best representation on this task, achieving an accuracy of 82.02%.

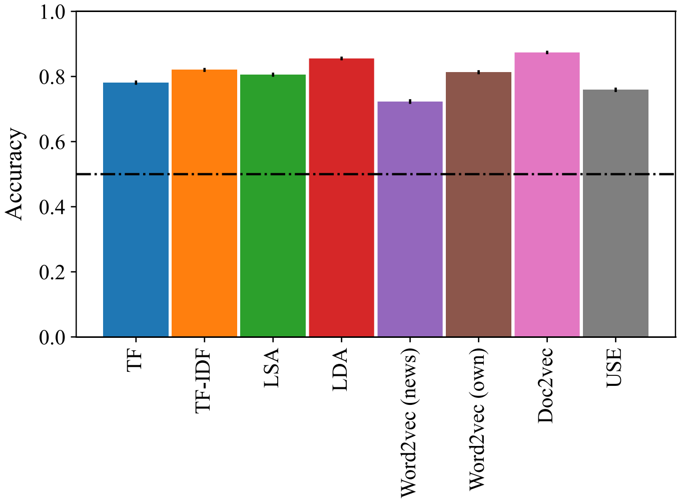

Figure 6. Accuracy of Different Document Representations with Respect to Object Similarity

Detailed Analysis of Temporal Similarity

The finance and accounting literature has especially focused on the temporal dimension of document similarity and applied a variety of representations to this task (see Table 1). To refine the binary analysis carried out earlier, report pairs are formed with different publication intervals ranging from 0 to 52 weeks. All pairs are built from reports on identical companies and different authors. The line plot (see Figure 7) shows how the temporal distance of pairs is related to document similarity. Unlike with object similarity investigated in Figure 5, we do not observe monotonously falling curves but a striking pattern of waves with a wavelength of approximately 13 weeks. This can be attributed to the quarterly earnings releases, which are important events for financial analysts and are discussed intensively in their reports (Huang et al. 2018). A report published shortly after the EA is more likely to be similar to a report published in 13 weeks than to a report published in two weeks and thus between two EAs. Such seasonal aspects are not specific to analyst reports but can be applied to many other domains. It is likely that seasonal fluctuations in the document similarity of news articles due to factors such as weather, holidays, or annual events are also observable. Since such a confounding factor can distort analyses, researchers should be aware of this problem and control for it. We do this by calculating the relative distance to the closest EA date for each report. We then form pairs where Reports A and B are published within five days, and Reports A and C are at least 70 days apart. However, both pairs must be published at the same time relative to the next earnings date. This procedure controls for the problem of seasonality. The report pairs are also about the same company and from different brokers. The results are depicted in Figure 8.

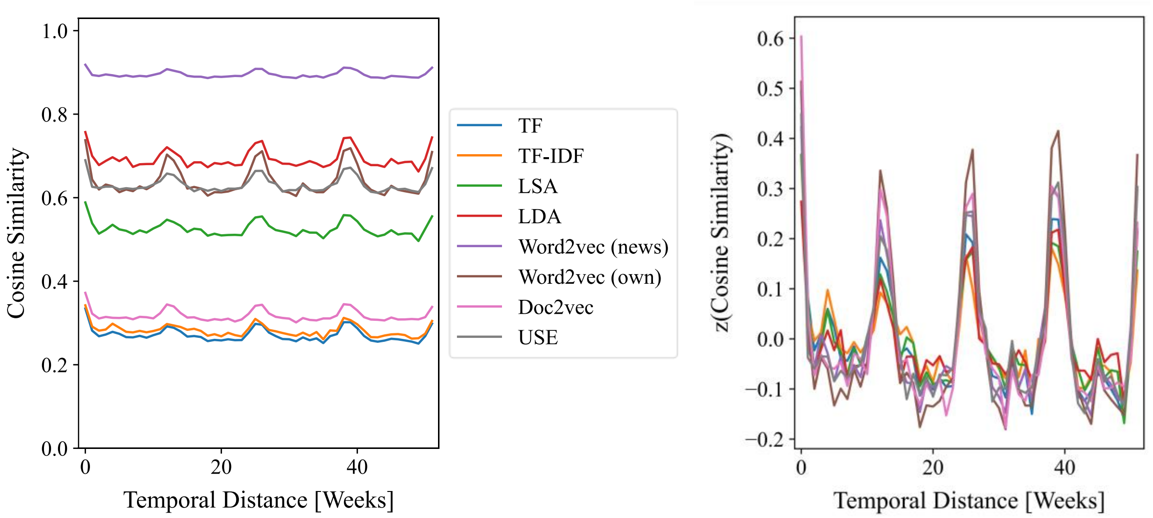

Figure 7. Document Similarity and Temporal Distance

It is revealed that doc2vec is again the most precise representation with an accuracy of 62.32%, followed by TF with 61.89%. The results are on the same level as those from the comparable analysis in Figure 4 without considering seasonality. LDA performs worst, with only 57.64% correctly identified pairs on their temporal proximity.

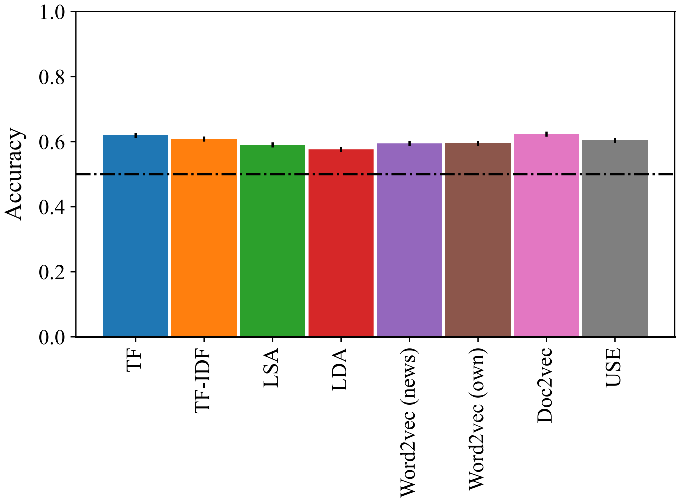

Figure 8. Accuracy of Different Document Representations with Respect to the Temporal Similarity

Discussion

Our paper provides numerous insights into the computation of document similarities in the finance and accounting domain. The similarity cube provides users with a useful framework to define the similarity dimension of interest. The analysis of the existing literature further suggests that these dimensions are well chosen and can be easily applied to the problems and research questions found in the literature. However, it seems plausible that this framework can also be applied to other domains, since the dimensions *author*, *object,* and *time* are generic. We also demonstrate that existing literature often applies certain representations without discussing how the representation fits the underlying research question. Only in Mauritz et al. (2021), we find that different representations were chosen for certain aspects of similarity.

Thus, a systematic investigation of document representations for the estimation of document similarity seems appropriate.

Doc2vec has proven to be a prime general-purpose solution across all dimensions. Therefore, especially those who are not interested in measuring a specific dimension should consider using it. Moreover, LDA has proven to be particularly useful for recognizing the object dimension. To create text-based company clusters or peer groups, LDA could be a promising approach. Considering the other dimensions, however, LDA is not suitable. The dense and widely used document representations TF and TF-IDF also perform generally well on all three dimensions. This finding can help researchers who must balance accuracy on the one hand with simplicity and replicability on the other. The relatively poor performance of the universal sentence encoder in this study shows that the most complex models do not always lead to the most accurate results. This analysis can also help researchers and practitioners to detect and analyze seasonal effects in the studied document streams. Since these effects are likely to be present in many areas and are not limited to financial documents or analyst reports, they can distort analyses.

Our analysis comes with some limitations. First, as the study relies on a corpus of analyst reports, only one type of financial documents is used. This limits the generalizability of our analysis. In addition, the author dimension is based on the broker level. In fact, many analysts work for a single brokerage house, which is why relying on the analyst level would be more accurate. However, this is not used because analysts are industry or even company experts, which would result in a large overlap between the object and author dimensions. This problem is avoided by using the broker level. In addition, information about the analysts is not available in the dataset's metadata. Another limitation is the isolated consideration of the document representation. In real-world research projects, the choice of pre-processing, document representation, and similarity measure is not made independently. We make an exception in the case of USE and do not perform pre-processing for this representation (Cer et al. 2018a), but it would also be possible to improve the results of the other representations by adjusting pre-processing steps. For example, certain expressions could be concatenated by phrase detection (e.g., cash flow → cash_flow) (Mikolov et al. 2013b). Another possibility to improve performance, especially for LDA on the temporal and author dimensions, could be to create industry- or company-specific models (Huang et al. 2018; Palmer et al. 2018). For the USE representation, fine-tuning based on a classification task could be performed to potentially increase the performance. However, this would require researchers who want to apply the same approach to have a labeled dataset at hand. It would also mean that we compare a supervised learning method with unsupervised representations. Thus, we refrained from all these possible optimizations to avoid complicating the analysis by adding further combinations. Finally, it should be noted that neither the similarity dimensions nor the representations included in the analysis are exhaustive.

This paper offers multiple starting points for future research. On a conceptual level, the similarity cube can be supplemented by additional dimensions. An in-depth analysis of the interactions between the proposed dimensions would be an interesting task for future research. This could involve simultaneously changing several dimensions. At the same time, further representations should be evaluated using the framework and experimental design developed in this study. Since the calculation of the document similarity as shown in Figure 2 is complex and consists of several process steps, the neighboring process steps (text pre-processing and similarity measure) should be evaluated. The analysis of different combinations of the individual sub-steps could also be an interesting task for future research but will probably lead to high complexity. Furthermore, the document representation itself offers interesting opportunities for extension. These include enrichment with information from named-entity recognition (Friburger et al. 2002), which could be of particular interest in the finance and accounting domain. Table 3 shows that document representations require a comprehensive set of hyperparameters that can be tuned to achieve more accurate results than in our experiment. Future studies might guide researchers and practitioners to find an appropriate hyperparameter configuration depending on their problem. Future research should also address the issue of document length. In this study, we used very long documents, as they are common in the finance and accounting domain (e.g., 10-K or sustainability reports). However, this prevents the usage of models such as BERT, whose input length is limited (Sun et al. 2019). It is important to investigate how well the different models perform with short texts.

However, particular emphasis should be placed on future empirical research in the finance and accounting domain that applies document similarity to close research gaps. Whenever associations between objects, authors, or the temporal dimension are to be captured on a textual level, or when information flows are studied, the use of document similarity might be a methodological approach to consider.

Conclusion

Researchers and practitioners face the challenge of choosing from many available document representations and justifying their choice when calculating document similarity. In the finance and accounting domain, and probably beyond, researchers aim to capture many different constructs by using the similarity between documents. A review of the literature has revealed that there are no generally accepted best practices for choosing document representations. This applies to the overall level as well as to individual similarity dimensions. Our results suggest that, on the one hand, the use of doc2vec provides accurate results across all dimensions. The topic model LDA, on the other hand, accurately captures the object dimension but does not provide satisfactory results for the other dimensions. Furthermore, the simple and understandable bag-of-words models perform surprisingly well. In addition, we show that seasonality plays an important role when investigating the temporal similarity dimension and that it should be controlled for, where appropriate.

This paper is intended to help researchers from the finance and accounting domain when deciding which methodologies to use for answering interesting research questions that require a quantification of document similarity. The aim of this study is also to stimulate this kind of research, as the use of NLP in the finance and accounting domain is still largely focused on sentiment analysis and topic modeling (Kang et al. 2020; Loughran and McDonald 2016).

Document similarity provides an exciting playground for future methodological research. Since the measurement of document similarity is a highly complex task, there are many parameters whose effects on the results and accuracy should be further explored.

References

Adosoglou, G., Lombardo, G., and Pardalos, P. M. 2021. "Neural Network Embeddings on Corporate Annual Filings for Portfolio Selection," *Expert Systems with Applications* (164).

Bär, D., Zesch, T., and Gurevych, I. 2011. "A Reflective View on Text Similarity," *Proceedings of the International Conference Recent Advances in Natural Language Processing*, Hissar, Bulgaria, pp. 515-520.

Beaupain, R., and Girard, A. 2020. "The Value of Understanding Central Bank Communication," *Economic Modelling* (85), pp. 154-165.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. "Latent Dirichlet Allocation," *Journal of Machine Learning Research* (3), pp. 993-1022.

Brown, S. V., and Knechel, W. R. 2016. "Auditor-Client Compatibility and Audit Firm Selection," *Journal of Accounting Research* (54:3), pp. 725-775.

Brown, S. V., and Tucker, J. W. 2011. "Large-Sample Evidence on Firms' Year-over-Year MD&A Modifications," *Journal of Accounting Research* (49:2), pp. 309-346.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., and Tar, C. 2018a. "Universal Sentence Encoder for English," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Brussels, Belgium, pp. 169-174.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Céspedes, M., Yuan, S., and Tar, C. 2018b. "Universal Sentence Encoder." Available at arXiv:1803.11175.

Chen, J., and Sarkar, S. 2020. "A Semantic Approach to Financial Fundamentals," *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, Kyoto, Japan, pp. 22-26.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. 1990. "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science* (41:6), pp. 391-407.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of NAACL-HLT*, Minneapolis, United States, pp. 4171-4186.

Eickhoff, M., and Muntermann, J. 2016. "How to Conquer Information Overload? Supporting Financial Decisions by Identifying Relevant Conference Call Topics," *Proceedings of the 20th Pacific Asia Conference On Information Systems*, Chiayi, China.

Friburger, N., Maurel, D., and Giacometti, A. 2002. "Textual Similarity Based on Proper Names," *Proceedings of Workshop on Mathematical Formal Methods in Information Retrieval at th 25th ACM SIGIR Conference*, Tampere, Finland, pp. 155-167.

González, J. R. C., Romero, J. J. F., Guerrero, M. G., and Calderón, F. 2015. "Multi-Class Multi-Tag Classifier System for Stackoverflow Questions," *2015 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)*, Ixtapa, Mexico, pp. 1-6.

Hanley, K. W., and Hoberg, G. 2010. "The Information Content of IPO Prospectuses," *The Review of Financial Studies* (23:7), pp. 2821-2864.

Hoberg, G., and Phillips, G. 2016. "Text-Based Network Industries and Endogenous Product Differentiation," *Journal of Political Economy* (124:5), pp. 1423-1465.

Huang, A. 2008. "Similarity Measures for Text Document Clustering," *Proceedings of the sixth New Zealand Computer Science Research Student Conference*, Christchurch, New Zealand, pp. 9-56.

Huang, A. H., Lehavy, R., Zang, A. Y., and Zheng, R. 2018. "Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach," *Management Science* (64:6), pp. 2833-2855.

Jatnika, D., Bijaksana, M. A., and Suryani, A. A. 2019. "Word2Vec Model Analysis for Semantic Similarities in English Words," *Procedia Computer Science* (157), pp. 160-167.

Jurafsky, D., and Martin, J. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, (2nd ed.). Upper Saddle River, United States: Prentice-Hall.

Kang, Y., Cai, Z., Tan, C.-W., Huang, Q., and Liu, H. 2020. "Natural Language Processing (NLP) in Management Research: A Literature Review," *Journal of Management Analytics* (7:2), pp. 139-172.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. 2015. "From Word Embeddings to Document Distances," *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, pp. 957-966.

Lang, M., and Stice-Lawrence, L. 2015. "Textual Analysis and International Financial Reporting: Large Sample Evidence," *Journal of Accounting and Economics* (60:2), pp. 110-135.

Le, Q., and Mikolov, T. 2014. "Distributed Representations of Sentences and Documents," *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China: PMLR, pp. 1188-1196.

Liu, R., Mai, F., Shan, Z., and Wu, Y. 2020. "Predicting Shareholder Litigation on Insider Trading from Financial Text: An Interpretable Deep Learning Approach," *Information & Management* (57:8).

Loughran, T., and McDonald, B. 2016. "Textual Analysis in Accounting and Finance: A Survey," *Journal of Accounting Research* (54:4), pp. 1187-1230.

Madigan, D., Genkin, A., Lewis, D. D., Argamon, S., Fradkin, D., and Ye, L. 2005. "Author Identification on the Large Scale," *Proceedings of the 2005 Meeting of the Classification Society of North America*, St. Louis, United States.

Mauritz, C., Nienhaus, M., and Oehler, C. 2021. "The Role of Individual Audit Partners for Narrative Disclosures," *Review of Accounting Studies*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013a. "Efficient Estimation of Word Representations in Vector Space." Available at arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. 2013b. "Distributed Representations of Words and Phrases and Their Compositionality," *Proceedings of the 26th International Conference on Neural Information Processing Systems*, Lake Tahoe, United States, pp. 3111-3119.

Naseem, U., Razzak, I., and Eklund, P. W. 2020. "A Survey of Pre-Processing Techniques to Improve Short-Text Quality: A Case Study on Hate Speech Detection on Twitter," *Multimedia Tools and Applications* (80:28), pp. 35239-35266.

Niraula, N., Banjade, R., Ştefănescu, D., and Rus, V. 2013. "Experiments with Semantic Similarity Measures Based on LDA and LSA," in *Statistical Language and Speech Processing,* A.-H. Dediu, C. Martín-Vide, R. Mitkov and B. Truthe (eds.). Berlin, Heidelberg, Germany: Springer, pp. 188-199.

Palmer, M., Eickhoff, M., and Muntermann, J. 2018. "Detecting Herding Behavior Using Topic Mining: The Case of Financial Analysts," *Proceedings of the 26th European Conference on Information Systems*, Portsmouth, United Kingdom.

Pennington, J., Socher, R., and Manning, C. D. 2014. "GloVe: Global Vectors for Word Representation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, pp. 1532-1543.

Pradhan, N., Gyanchandani, M., and Wadhvani, R. 2015. "A Review on Text Similarity Technique Used in IR and Its Application," *International Journal of Computer Applications* (120:9), pp. 29-34.

Rawte, V., Gupta, A., and Zaki, M. J. 2021. "A Comparative Analysis of Temporal Long Text Similarity: Application to Financial Documents," in *Mining Data for Financial Applications. MIDAS 2020. Lecture Notes in Computer Science,* V. Bitetta, I. Bordino, A. Ferretti, F. Gullo, G. Ponti and L. Severini (eds.). Cham, Switzerland: Springer International Publishing, pp. 77-91.

Refinitiv. 2022. "The Refinitiv Business Classification Methodology." Retrieved 05-19-2022, from https://www.refinitiv.com/content/dam/marketing/en_us/documents/methodology/trbc-business-classifcation-methodology.pdf

Şaşmaz, E., and Tek, F. B. 2021. "Tweet Sentiment Analysis for Cryptocurrencies," *6th International Conference on Computer Science and Engineering (UBMK)*, Ankara, Turkey, pp. 613-618.

Shahmirzadi, O., Lugowski, A., and Younge, K. 2019. "Text Similarity in Vector Space Models: A Comparative Study," *2019 18th IEEE International Conference On Machine Learning And Applications*, Boca Raton, United States, pp. 659-666.

Singhal, A. 2001. "Modern Information Retrieval: A Brief Overview," *IEEE Data Engineering Bulletin* (24:4), pp. 35-43.

Sparck Jones, K. 1972. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval," *Journal of Documentation* (28:1), pp. 11-21.

Sun, C., Qiu, X., Xu, Y., and Huang, X. 2019. "How to Fine-Tune BERT for Text Classification?," *China National Conference on Chinese Computational Linguistics*, Kunming, China: Springer, pp. 194-206.

Tan, M., Dos Santos, C., Xiang, B., and Zhou, B. 2016. "Improved Representation Learning for Question Answer Matching," *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, pp. 464-473.

Trieu, L. Q., Tran, H. Q., and Tran, M.-T. 2017. "News Classification from Social Media Using Twitter-Based Doc2Vec Model and Automatic Query Expansion," in: *Proceedings of the Eighth International Symposium on Information and Communication Technology*. Nha Trang City, Vietnam: Association for Computing Machinery, pp. 460–467.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. 2017. "Attention Is All You Need," *31st Conference on Neural Information Processing Systems*, Long Beach, United States, pp. 5998-6008.

Yan, J., Wang, K., Liu, Y., Xu, K., Kang, L., Chen, X., and Zhu, H. 2018. "Mining Social Lending Motivations for Loan Project Recommendations," *Expert Systems with Applications* (111), pp. 100-106.