# Università degli Studi di Palermo

Dottorato in Scienze Economiche e Statistiche

Dipartimento di Scienze Economiche, Aziendali e Statistiche

SECS-S/06 - Metodi matematici dell'economia e delle scienze attuariali e finanziarie


# Development of statistical methods for the analysis of textual data


IL DOTTORE
**Andrea Simonetti**

IL COORDINATORE
**Andrea Consiglio**

IL TUTOR
**Michele Tumminello**

IL CO-TUTOR
**Andrea Consiglio**



CICLO XXXV

ANNO CONSEGUIMENTO TITOLO 2022

# Contents

# List of Tables

# List of Figures

# Introduction

The amount of text data is increasing, and the development of hardware and software platforms for the web and social networks enabled the rapid creation of large repositories. So, there is a need to explore methods and algorithms that can face various text applications and tasks. Text Mining or knowledge discovery in text (KDT), first introduced by Feldman and Dagan (1995) [82], refers to the techniques and algorithms to extract meaningful information from the textual data in a dynamic and scalable way. It spans many research areas, including computational linguistics, information retrieval, data mining, and machine learning. Indeed, many of the text mining algorithms make use of Natural Language Processing (NLP) tools, such as Part-Of-Speech tagging (POS-tag), syntactic parsing, Named Entity Recognition (NER), and Word Sense Disambiguation (WSD). Computational linguistics, involving computer science, artificial intelligence and linguistics, aims to understand human language using computers [131, 142]. It employs techniques to learn, understand, and produce human language contents. Early computational language approaches focused on automating the analysis of the linguistic structure of language and developing essential technologies such as machine translation, speech recognition, and speech synthesis.

In the last decades, researchers implemented lexical databases, such as WordNet[1] and BabelNet [2], and grammar rules of human languages. These tasks present many difficulties due to the variability, ambiguity and context-dependent interpretation of human language. To state the meaning of a word is a tricky concept because words may have different meanings in different contexts. Many algorithms extract textual information representing documents as a set of words without regarding the disclosure structure and meaning. The "bag-of-words" (BoW) is a representation used in Information Retrieval (IR), where the text is represented as the bag (multi-set) of its words, disregarding grammar and even word order but keeping multiplicity. This representation is commonly used in document classification methods where the word's occurrences are used as features for training a model.

[1] `https://wordnet.princeton.edu`
[2] `https://babelnet.org`

Therefore, De Roeck and Sarkar (2004) [64] showed that frequent words are not distributed homogeneously over a text and provided evidence that the bag of words assumption is invalid. For example, *homonymy* and *polysemy* are common issues in linguistics. *Homonym* refers to words that are entirely unrelated but spelt the same way. *Polysemy* concerns words with different meanings (also called word senses). For example, the word *interest* can have several meanings indicating curiosity (interest in football), a stake (a 5% interest in Google), or the fee paid for a loan (interest rate of 4.9%). In semantics, the definition of word sense is still challenging, and the task of determining the proper word sense for a word in a given context is called Word Sense Disambiguation (WSD). Lexical semantics, the study of the meaning of words, shows that the context, such as words in a close neighbour or words in a larger window, is a good indicator of understanding the sense of a word.

Indeed, if words are the atomic units of meaning, then the sequence of words is the next step. An $n$-gram is defined either as a textual sequence of length $n$ or, similarly, as a sequence of $n$ adjacent "textual units". $N$-grams are examples of the local context of words and can be defined as co-occurrences in short windows of length $n$. The use of $n$-grams provides advantages in many applications, such as spelling error detection and correction, query expansion, text compression, language identification and text generation. The latter refers to the analysis of patterns in languages and is based on statistics of $n$-grams. It is closely related to the concept of stylometry. Stylometry is the application of the study of linguistic style, usually written language, and it is often used to attribute authorship to anonymous or disputed documents. Authorship attribution is the task of identifying the author of a given document. The authorship attribution problem has recently gained more importance due to new applications in forensic analysis and humanities scholarship [204]. However, many works rely on invalid assumptions [191], and researchers focus the analysis on attribution techniques rather than extracting new style markers that are more precise and based on less strong assumptions. Some studies [108, 178] point out the importance of exploiting sequential information in text and prove that word sequences can be relevant to understanding patterns that may be characteristic of a specific type of text.

In Chapter 1, we propose an application of Language Models (LMs) to the financial and accounting domain. The Language Models regard the analysis of language features, such as $n$-grams, to determine the probability of a given sequence of words occurring in a sen-

tence. So, we transpose the task of authorship identification into bankruptcy prediction. We aim to verify the hypothesis that the reports' disclosure style of public companies is representative of their financial conditions and the challenge concerns discovering the existence of a bankruptcy language. Moreover, we want to show that the use of dictionaries' sentiment of words is limited and falls into the Word Sense Disambiguation problem. Instead, our approach analyses how sentences are composed, trying to understand how the language changes from healthy to bankrupt companies, from those companies who need nothing to hide or to justify to those who do not. In this respect, we can consider our work as an exploration of methodologies to identify linguistic patterns, retrieve features in the framework of suspicious language.

Furthermore, one of the main recent results of NLP was the introduction of the semantic vector space model. Distributional semantics is a research area in linguistics that aims to quantify and categorize semantic similarities between words, studying their distributional properties. The distributional hypothesis states that words that occur in similar contexts share the same meaning [100]. Semantic space relies on the computational challenges of retrieving distributional characteristics to measure the similarity among words, sentences, or entire documents. The first proposal was the "bag-of-words" vector space model, used to represent a document as a vector where each dimension corresponds to a separate word. Recent works have focused more on the vector representation of words instead of documents. Words are mapped in a vector space of real values, encoding their meanings. Mikolov et al. (2013) [154] proposed a word embedding representation through neural networks. Global Vectors for Word Representation (GloVe), developed and described by Pennington et al. (2014) [167], is another method for word vector representation. The results show how relationships among words can be performed by simple algebraic operations through vector representation of words. Popular examples are:

- vec(*"king"*) - vec(*"man"*) + vec(*"woman"*) = vec(*"queen"*)
- vec(*"biggest"*) - vec(*"big"*) + vec(*"small"*) = vec(*"smallest"*)
- vec(*"german"*) + vec(*"capital"*) = vec(*"berlin"*)

These methods generate word vectors in terms of the context in which words appear. The assumption is that the statistics of word co-occurrences is the primary source of information to implement unsupervised methods for learning word representations. GloVe can be

considered a count-based method: it uses the co-occurrence word-word matrix and reduces it to a co-occurrence word-feature matrix.

Topic Model is one of the main applications in text mining that involves distributional semantics. In particular, probabilistic topic modeling is one of the most popular probabilistic clustering algorithms. It aims to process extensive collections of texts that are useful for tasks such as classification and summarization. However, topic discovering is an unsupervised process that does not guarantee the interpretability of its output. These models do not automatically provide a way to interpret their outputs. One of the fundamental challenges in topic detection models is assessing the semantic coherence of estimated topics in terms of human interpretability. However, state-of-the-art coherence measures focus on the marginal probabilities of words and their co-occurrence and none of them takes into account the randomness of co-occurrences.

In Chapter 2, we study word co-occurrences to face the task of interpretation. We propose a method to measure the coherence of a set of words to evaluate probabilistic topic models. So, we move our attention from local co-occurrences ($n$-grams) to global co-occurrences of words, considering the context of words as the entire sentence in which they occur. We propose a new quality evaluation method based on Statistically Validated Networks (SVNs) [214]. The method represents each topic as a weighted network of its most probable words. The presence of a link between each pair of words is assessed by statistically validating their co-occurrence in sentences against the null hypothesis of random co-occurrence. The statistically significant pairwise associations of words represented by the links in the SVN might reasonably be expected to be strictly related to a topic's semantic coherence and interpretability. However, the understanding and comprehension of language are not deterministic or absolute. Instead, it is subjective for each person, so there is no panacea for organizing words and text into pre-defined categories So, we needed human judgment as a benchmark to prove the efficacy of our method. In doing so, we set up a survey among all the PhD students at the University of Palermo.

The work presents an approach to represent the semantic relationships among words through a weighted network. The results of the application motivate us to go ahead on exploring the potential of the methodology proposed. Then, we explore the ability of

the SVN method to manage the word co-occurrences in a collection of documents. The third work, described in Chapter 3, aims to use the SVN method to face the task of document clustering and topic extraction. We construct a co-occurrence network in which we apply community detection algorithms to find the latent thematic structure of a collection of texts. Generally, it is difficult to deal with co-occurrence networks [179] due to the high density of links. However, we introduce a statistical test to filter out less informative semantic relations among words. We analyse a collection of abstracts used to conduct systematic literature reviews (SLR), which aim to summarize and discern what we should know about a specific theme. We present an approach based on network analysis and Natural Language Processing (NLP) that allows extracting textual features to (i) select relevant studies on a specific theme; (ii) discern the main topics around the theme. Furthermore, we want to propose a method that tries to solve some issues of probabilistic topic models. The most popular topic model is the Latent Dirichlet Allocation (LDA) model [33]. However, the model suffers when applied to short texts, and the number of topics must be fixed in advance. Other methods, such as hierarchical bayesian model [163], Stochastic Block Model (SBM) [88], and hierarchical SBM (hSBM) [111, 166] solve the problem of setting the number of topics providing a hierarchical structure of topics. However, these models are stochastic and need to fine-tune the hyper-parameters. Our proposal overcomes these issues and supports academic research from the perspective of literature reviews. In SLR, the research regards a collection of papers regarding a unique general topic. In Appendix C, we apply our method in more heterogeneous collections of documents. Gerlach et al. (2018) [88] show the connections between probabilistic topic models and SBM, linking the task of topic extraction with community detection in a network of words. We apply the SVN method to a collection of Wikipedia articles representing documents in a network. We compare the results with Hyland et al. (2021) [111] and prove the method's efficacy in capturing the key semantic relationships among words. Indeed, the network representation of documents allows studying its topology to discriminate more methodological and interdisciplinary articles.

Finally, Chapter 4 is more methodological than the previous. We introduce a new discrete probability distribution that aims to describe the concentration of word attributes in short sentences. Following the results proposed in Chapter 1, we analyse the text of companies' reports testing if word attribute distribution highlights linguistic differences

between healthy and bankrupt companies. The research aims to compare two corpora, as in Chapter 1, to extract features of bankruptcy language and examine subtle choices in grammar use [58]. Furthermore, we study other datasets: twins affected by neuronal disorders, and children in the household, with respect to their status as NEET, student, or worker, and with respect to gender.

# Chapter 1

# Bankruptcy prediction: analysis of word sequences and words meaning in different contexts

**Abstract**

*Academics and practitioners searched for reliable indicators of companies' failure focusing only on quantitative data such as financial ratios and market variables. However, recent literature aims to quantify textual information of financial reports studying features such as topics and words' co-occurrences, confirming their usefulness in predicting company bankruptcy. In this work, we propose a new approach to analysing texts that focuses on sentences interpreted as ordered sequences of words. We propose a new approach, based on Language Model, to predict the company's bankruptcy that was released in the next year. Given the high predictive power of the model, we investigate the sentences of texts to gain insights into how failing companies' language differs from the non-failing one. Our approach allows us to move away from fixed wordlists, exploring linguistic features to understand how a word is used in different contexts. Therefore, contexts give words a certain degree of association with failure or non-failure. The results of our analysis lead us to observe that the concept of bankruptcy can take on different meanings arising from the different legitimisation strategies that companies facing bankruptcy may use.*

## 1.1 Introduction

Corporate bankruptcy is one of the most important credit risk factors and attracts the attention of creditors and investors. Therefore, an accurate bankruptcy forecasting model is essential for practitioners, regulators, and academic researchers who can use it to supervise the financial health of individual institutions, contain systemic risks and predict default probability to price corporate debt [195]. Given the massive costs of bankruptcy [38, 43], academics and practitioners searched for reliable indicators of companies' failure to allow investors to reduce their risk of investing in such companies. First, accounting and finance studies monitored credit risk and bankruptcy prediction focusing only on quantitative data such as financial ratios and market variables [10, 198, 11]. Recent Business Failure Prediction (BFP) literature focused on incorporating textual content extracted from annual disclosures to the respective regulatory authorities [34]. The business failure event has been defined in many studies and often refers to circumstances leading to a discontinuity of a company's operation, including filing for bankruptcy or insolvency. It regards a binary classification task that aims to predict future performance based on all known about that company at a given moment in time. Predicting such events has a dual purpose: detecting companies that are at risk and helping to understand why firms are at risk. Therefore, predictive performance is not the only requirement, as it should also provide output that can help decision-makers to understand why firms are at risk of failure [76].

In quantitative finance, such a problem is related to modelling the credit risk of companies. The most popular Altman's Z-Score Model [9] is still used as a benchmark for new models. It concerns applying discriminant analysis to combine financial statement measures and the equity market value to predict corporate defaults over one year or more. Statistical and neural network models have shown that accounting-based ratios (e.g., profitability and liability ratios) and stock market data (e.g., stock market returns and volatility) offer helpful information on whether a firm is financially healthy or may file for bankruptcy [75, 211]. However, it has been highlighted that these models based only on financial ratios and market variables can be subject to severe limitations. These limitations are related to factors such as the decline in the explanatory power of financial ratios [22] and the paradox of relying on accounting-based measures prepared with a going-concern assumption to predict failure [129]. On this basis, Lopatta et al. (2017) [133] posit that

no versions of accounting measures-based models constitute a comprehensive bankruptcy prediction model because of their backwards-looking nature.

The digitisation and the online availability of corporate reports' narratives and other companies' textual sources provide relevant textual data to include as language variables for credit risk modelling to improve bankruptcy investigation [133]. Indeed, according to accounting literature, narratives in corporate reports contain incremental information to accounting statements and financial ratios [152]. The information in these textual sources complements the traditional financial reporting model overcoming its limitations by providing more forward-looking and non-financial insights into companies' value creation process [21].

Recent works on text analysis of financial reports aim to quantify textual information as significant variables to predict future company financial performance. These works concern predicting stock returns [134], finding the role of investment analyst reports [110], analysing manager sentiment tone [233] or investigating phenomena such as financial distress and bankruptcy [47, 147, 133, 229, 4]. Some of these studies on qualitative corporate filings confirmed text data's explanatory power and usefulness in predictive tasks, such as discriminating between bankruptcy and non-bankruptcy companies. In other words, these studies seem to prove that a "bankrupt language" exists. However, most of these works on textual analysis rely on descriptive text statistics techniques such as sentiment word count, length, spelling errors, tones and readability [73, 3, 128, 127]. These approaches present some limitations related to the pitfalls of investigating the complexity and fuzziness of natural language. Moreover, management has incentives to hide bearish information or to use vague language in their disclosure. So, predictive models based on textual data require rethinking the entire modelling process.

Dictionary methods based on expert knowledge are perhaps the simplest form of feature extraction, whereby general or domain-specific words are extracted from the text and treated as feature inputs. In contrast to these approaches, neural network models have also been explored to extract textual features from annual disclosures [139, 34, 205, 226]. These models have the advantage of working with vector representation of words (word embedding) but also some limitations, such as the large amount of manually annotated

data needed for training the classifier and the difficulty in exploring the linguistic features learned. Therefore, we propose a new approach, based on Language Model, to investigate corporate narratives by combining the analysis of the sequence of words and word co-occurrences. In particular, we want to move our focus from a fixed word list to words' contexts and meanings.

## 1.2 Literature review

The forms of textual analysis employed in accounting and finance vary in a continuum between qualitative to quantitative methods [153]. Qualitative methods are usually conducted manually, while quantitative methods are performed automatically.

As regards qualitative methods, scholars often use the term "narrative" to indicate the mainly European critical/interpretive branch of studies that rely on these methods [20]. These studies are grounded on the narrative turn concerning searching for narratives' meaning through hermeneutic methods such as interpretive content analysis. Moreover, content analysis has some limits since its studies focus on "what" is disclosed [99] and the text is just seen as a representation or reflection of social reality [151]. Instead, scholars suggested more in-depth analyses of "why" and "how" the message is disclosed. In this regard, several researchers used discourse[170, 213], rhetoric [104] and narrative analysis [162] to interpret the meaning of written narratives. These methods can be more precise and tailored to the specific research setting. However, they lead to the impossibility of dealing with large sample sizes, limiting the empirical results' generalisability.

As regards quantitative methods, academics distinguish a mainly North American and positivistic branch of studies, characterised by widespread usage of these computer-assisted textual analyses, from the European interpretive one, by using the term "disclosure" studies [19]. These more quantitative and automated methods are grounded in a positivist paradigm. They have been related to the development of computer science, the exploitation of big data and sophisticated computational methods of text mining and natural language processing (NLP). One of their main objectives is to extract incremental information to solely financial statements to predict the performance of companies [19, 153]. The investigation of written narratives through quantitative automated methods has the advantage of being economical in terms of time and effort. This advantage allows researchers

to investigate large samples and draw inferences from texts. Indeed, these computer-based approaches improve the generalisability of the empirical results and lead to more follow-up research.

Sentiment analysis is a concept taken from natural language processing (NLP). It aims to measure the positive or negative orientation of the text. The application in the financial domain is a challenging task due to the specialized language. In the case of corporate reports, positive texts are understood as information that has a positive impact on the company's value. Negative texts are those that contain information that has a negative impact on the company's value. Two main areas of research explore methods used to determine the sentiment measures in financial texts: methods based on dictionaries and methods based on machine learning (ML). The dictionary approach concerns "bag-of-words" models, which treat documents as a set of words. The words are disconnected from their context and have predefined sentiment categories. The sentiment of a text involves the calculation of indicators, usually based on the number of words belonging to each category.

One of the first studies used the information content of accounting narratives to explain bankruptcy using textual analysis [209]. The researchers investigated whether there are differences in the narratives of companies approaching bankruptcy and companies that are not and if these differences permit classifying companies in bankruptcy and non-bankruptcy. Shirata et al. (2011) [197] investigated differences between the languages of bankrupt and non-bankrupt companies. The researchers analysed word co-occurrences and found that some words appearing together in the same annual report section could help to recognise the company as bankrupt or not. These studies provide evidence regarding significant differences in the language of bankrupt and non-bankrupt companies. According to these findings, it seems possible to assume that companies' language could contain a predictive power regarding their financial distress.

In their pioneering work, Loughran and McDonald (2011) [135] built a sentiment dictionary which assigns a more accurate tone to words compared with the traditional Harvard's General Inquirer word list in the context of the financial text. For example, Gandhi et al. (2019) [86] use the sentiment tags of words in the dictionary to extract linguistic indicators to examine the financial distress of U.S. banks, suggesting that a higher probability

of distress is related to a higher frequency of negative words in the reports.

Other works examine the 10-K financial reports, also known as Form 10-K, that must be submitted by most public companies to the United States Securities and Exchange Commission (SEC) every year. They are essential for investors to evaluate companies since they contain more detailed and accurate information due to the laws and commandments of the SEC, which forbids misleading and false information. For example, relying on a dictionary approach, Lopatta et al. (2017) [133] investigated companies' language used in 10-K filings focusing on litigious, positive, and negative terms by relying on a dictionary approach. Their findings highlight a significant relationship between the use of negative and litigious terms and the risk of bankruptcy. Yang et al. (2018) [229] analysed textual differences in high-frequency word occurrences between bankrupt and non-bankrupt companies founding that some high-frequency words suggest differences between bankrupt and non-bankrupt companies' ongoing status.

In these forms, the Management Discussion & Analysis (MD&A) section is a forward-looking statement in which the executives examine their company's performance, address the compliance and risks, and express their views on future company projects. In this vein, Cecchini et al. (2010) [47] created dictionaries from 10-k filings' MD&A section to discriminate between bankrupt and non-bankrupt and fraudulent and non-fraudulent companies using a vector space model, a modified TF-IDF and ontologies. Their dictionaries alone performed better in predicting bankruptcy and fraud than models based on quantitative measures (i.e., [9],[25]) and even better when combined with them. Their findings suggest that textual data contains relevant information complementary to quantitative measures. Although dictionary methods of textual analysis can solve the problem of disregarding information by relying on accounting statements, financial ratios, and market variables to study the company's results and behaviour, this approach is not free from drawbacks. They omit the meaning of words facing the problem of ambiguity due to their different contexts, which is sometimes crucial for understanding the tone of given sentences. Several accounting and finance researchers have already highlighted the limitations of this approach, identified as "context-related limitations", [136, 101]. For example, the word "decrease" may have a positive tone regarding the company's debt but a negative tone in regarding profit. So, The mutual contextual connections between words in sentences

are not considered [226, 101]. In computational linguistics, it is known as Word-sense disambiguation (WSD), which is still an open problem. Indeed, it is acknowledged that analysing phrases or $n$-grams rather than words would provide a better unit of analysis [94]. The introduction of neural network models partially resolves these issues. These new approaches allow considering words in terms of their contexts. Techniques, such as Word Embedding [154], provide vector representation of words based on word co-occurrences. Convolutional Neural Networks (CNN) [231] and Recurrent Neural Networks (RNN), such as long short-term memory (LSTM)[207] and Bidirectional Encoder Representations from Transformers (BERT)[69], consider the local context and the entire sequences of words, respectively. The evolution of such techniques has significantly improved the predictive capabilities in the Natural Language Processing (NLP) domain and has shown their effectiveness and power in extracting textual features. Araci (2019) [15] implemented a BERT architecture, pre-trained with text in the financial domain and further fine-tuned for sentiment analysis with the annotated sentiment dataset Financial PhraseBank [1] created by [141]. The authors address the problem of having a domain-specific model to retrieve better and more related word representations. For example, Mai et al. (2019) [139] applied different neural network architectures to predict bankruptcy. The authors used 10-K annual reports of U.S. public companies. It has been found that neural networks improve predictive accuracy when using text alongside accounting and market-based data. Similarly, Matin et al. (2019) [145] predict bankruptcy using segments of text from annual reports for Danish non-financial and non-holding private limited and stock-based firms finding that the text combined with financial features leads to improved prediction. Borchert et al. (2022) [34] tested deep neural network architectures, such as CNN and BERT, on a database of $13,571$ European companies. Recent studies [83, 121] also show the benefit of long short-term memory (LSTM) models to predict the financial market. In the framework of Micro, Small and Medium Enterprise (mSME) credit risk modelling, the work of Stevenson et al. (2021) [205] exploits different neural network models. These studies are significant as they demonstrate that deep learning models can improve bankruptcy predictions using financial and textual variables. However, they require a large amount of training data, manually annotated, as a prerequisite for obtaining a well-performing sentiment classification model. Unfortunately, due to the lack of large sets of labelled financial datasets, which are costly to get in the financial domain, it is challenging to use

[1] The dataset can be found here: https://www.researchgate.net/publication/251231364_FinancialPhraseBank-v10

neural networks to their full potential for sentiment analysis. Even when the model uses pre-trained values (word embedding), the rest of the model still needs to learn complex relations with a relatively small amount of labelled data. Moreover, these are black-box models, and and it isn't straightforward to understand their feature-constructing process.

Our analysis focuses attention on segments (sentences) of MD&A section of annual reports to investigate the relationships of the textual features with the probability of default, demonstrating the competitive predictive power of those textual features. Authorship attribution refers to identifying authors from texts by their unique textual features and in this context. Statistical Language model achieved great performance in authorship attribution by building $n$-gram models from a text produced by each author, and these models serve the role of author profiles [113, 114]. In our work, we rely on a statistical approach to analyse the local context of words ($n$-grams) through Language Model. However, to our knowledge, no works related to applying Statistical Language Models to explore the different linguistic styles of healthy and bankruptcy reports. Moreover, we want to discover the categories of meaning underlying the narration and investigate the words' system within the reports. In doing so, we propose an approach to test the context in which words are embedded and capable of detecting predefined categories of meaning underlying the narration (i.e., the "bankruptcy category").

## 1.3 Data collection

As mentioned above, several studies investigated US companies' 10-K filings. These reports are required by the Securities and Exchange Commission (SEC), are publicly available and their content is highly structured and digitalised. These characteristics make them particularly suitable for automated textual analyses, allowing researchers to analyse their content with minimal pre-processing. To create our 10-Ks' sample, we relied on two different sources of data: the UCLA-LoPucki Bankruptcy Research Database (UBRD)[2] and the Loughran and McDonald texts repository[3]. Bankrupt companies' CIK codes and information are obtained from the UBRD. We matched every bankruptcy in the UBRD with a non-bankrupt company. In particular, we matched companies by SIC codes (Industrial Sector), year of reports and Total Assets (control company within 10% of total

[2] LoPucki, L. M. (2015). UCLA-LoPucki bankruptcy research database user's manual. Unpublished manuscript. `http://lopucki.law.ucla.edu/index.htm`, retrieved October 01, 2020

[3] `https://sraf.nd.edu/sec-edgar-data/cleaned-10x-files/`

assets of the match) as in previous works [47]. In Appendix A, Table A1 provides the description of SIC codes, and Table A2 reports the list of matched pairs of companies providing information about the name, CIK code, SIC code, State and the date of failure for the bankrupt companies. On the ground of the aforementioned criteria, we obtained a sample of 82 bankruptcies and 82 non-bankruptcies. For each of these bankrupt companies, we collected the 10-Ks related to the year before failure, and 10-Ks were collected for the matched healthy companies during the same period. We collected the 10-Ks in .txt format from the Loughran and McDonald repository. Overall, we collected 164 10-Ks for the analysis. Indeed, we decided to test our approach by focusing on this particular section of the 10-k. We extracted the MD&A section using a Python script. Our choice was motivated by the fact that, among the 15 items composing the 10-K, the relevance of MD&A in terms of incremental information content is widely acknowledged. Indeed, it concerns the discussion of the company's financial condition – covering liquidity and capital resources – and the discussion of the results of operations and forward-looking information. In other words, this section allows us to deal with the soft information we are interested in, avoiding coping with too much noise and highly consuming computational efforts. In this respect, as described above, most of the studies using automated textual analyses have specifically addressed the MD&A [47, 147, 133], demonstrating its informative content.

### 1.3.1 Preprocessing

We first performed standard text preprocessing tasks, such as removing stop words, numbers, and punctuation. Then, we split the documents in sentences with the help of the package Spacy in Python and we stemmed the words (removing the inflectional endings from words) using the package NLTK in Python. Specifically the preprocessing steps are described as follow:

- retrieve MD&A section from the 10-K report, as described in Anand et al.(2020)[13]
- split in sentences with Spacy
- remove numbers, punctuations and stop words
- stemming words with NLTK

## 1.4 Methods

### 1.4.1 Language Model

Language model is an important method widely used in many applications in computational linguistics to face tasks such as speech recognition, spelling correction, machine translation, natural language generation, part-of-speech tagging and information retrieval for pattern recognition. It uses probabilities to estimate how likely any given sequence of words belongs to a language. The model can learn the rules of a language as a probability distribution of words, and it can predict the probability of a sequence of $t$ words. It attempts to reflect the frequency with which each sequence of length $t$ occurs as a sentence in natural text. It means that it approximates how a text is written.

The heuristics behind is:

*"ordinary" word sequences occur more often in text and speech than "weird" word sequences*

So, we want to answer the question:

*How likely is a sentence to appear in a language?*

In particular, our focus is on assigning probabilities to sentences. Considering the joint probability of a sentence of $t$ words:

$$\begin{aligned} P(w_1, \ldots, w_t) &= P(w_1) \cdot P(w_2|w_1) \cdot P(w_3|w_1, w_2) \ldots P(w_t|w_1, \ldots, w_{t-1}) \\ &= \prod_i^t P(w_i|w_{j<i}) \end{aligned} \tag{1.1}$$

we applied the chain rule property in order to manage the joint probability of words. Then, the straightforward way to compute the conditional probabilities is to consider the number of occurrences of the words' sequences:

$$P(w_t|w_1, \ldots, w_{t-1}) = \frac{count(w_1, \ldots, w_t)/N}{\sum_{w \in V} count(w_1, \ldots, w_{t-1}, w)/N} \tag{1.2}$$

where $N$ is the number of sequences of length $t$, and $V$ is the size of the vocabulary.
This is called the maximum likelihood (ML) estimate for $P(w_t|w_1, \ldots, w_{t-1})$, that is sim-

ply the number of times the sequence appears divided by the total number of sequences of length $t$.

The problem with the MLE estimation arises when the sequence is not seen in the held-out data, and no matter how large the corpus is, it's impossible for it to contain all possible sequences. Therefore, it is not easy to work with these probabilities since longer sequences may not occur in a corpus and, therefore, the resulting probability will be zero.

### N-gram Language Models and Markov assumption

Statistical Language Model is constructed by calculating $n$-gram probabilities, where an $n$-gram is defined as a sequence of $n$ words. In such a model, the Markov assumption has a central role in defining the time dependencies of the sequence of words. It means that the probability of observing a word only depends on a fixed number of previous words. Formally:

$$P(w_t|w_{t-n}, \ldots, w_{t-1})$$

where $n$ represents the order of the Markov Process. $N$-gram model decomposes the probability of a sequence of words into conditional probabilities of each word given the previous context. The Markov assumption comes to overcome the issue of computing conditional probabilities. The assumption states that only the previous local context, the last few words, affects the next word. So, the probability of a word only depends on a fixed number $n$ of previous words.

$$P(w_t|w_1, \ldots, w_{t-1}) = P(w_t|w_{t-n+1}, \ldots, w_{t-1}) \tag{1.3}$$

Although the introduction of independence from the past with the Markov assumption tries to avoid zero occurrences of a long sequence of words, the problem of zero division is not yet solved. As the order of the Markov Process increases, the chance that all $(n-1)$-grams are present in the training corpus is slight. However, providing a more extensive corpus for training does not solve the problem due to Zipf's law of the frequencies of words. In a given corpus, there are few widespread words but very infrequent words. So, $n$-gram models need a smoothing technique to face the problem of assigning non-zero probabilities to sequences that may never be seen in the training corpus. The solution is to "smooth" the language model to move some probability towards unknown $n$-grams. The term smooth-

ing describes techniques for adjusting the maximum likelihood estimate to produce more accurate probabilities hopefully. In literature, there are many smoothing techniques such as Additive smoothing, Witten-Bell smoothing, Jelinek-Mercer smoothing (interpolation), Katz smoothing (backoff) [142, 79, 51]. One of the most effective smoothing methods is the *Kneser-Ney Smoothing* [119] because of its use of absolute-discounting interpolation, which consists in subtracting a fixed value from the probability's lower order terms to leave out $n$-grams with lower frequencies. The method uses both higher- and lower-order $n$-grams, reallocating some probability mass to simpler unigram models. Empirical evidence points to Kneser-Ney smoothing as the state-of-art for $n$-gram language modelling [91]. The method is described below:

Let $w_{i-n+1}^{i-1}$ be the $n-1$ words before $w_i$ and $c(w, w')$ be the number of occurrences of the word $w$ followed by the word $w'$.

$$P_{KN}(w_n|w_1, \ldots, w_{n-1}) = \frac{max(c(w_{i-1}w_i) - \delta, 0)}{\sum_{w' \in V} c(w_{i-n+1}^{i-1}, w')} + \gamma_{w_{i-1}} \cdot P_{KN}(w_i|w_{i-n+2}^{i-1}) \tag{1.4}$$

where $\delta$ is called absolute discount factor, and $\gamma$ is the back-off weight, the amount of probability mass we left for the next lower-order mode:

$$\gamma(w_{i-1}) = \frac{\delta}{\sum_{w' \in V} c(w_{i-1}, w')} \cdot |\{w' : c(w_{i-1}, w') > 0\}| \tag{1.5}$$

The recursion stops at the unigram model:

$$P_{KN}(w) = \frac{c(w)}{\sum_{w' \in V} c(w')} \tag{1.6}$$

where $c(w)$ represents the occurrence of word $w$.

So, interpolating the probabilities, if a sequence has any k-gram suffix present in the corpus, it will give a non-zero probability. When we calculate the probability of a word given a context, we want to consider the current context and the number of contexts in which the word appears. This is the contribution of Kneser-Ney smoothing. For example, if a word appears after a small number of contexts, it should be less likely to occur in a novel context. Modifications of this method also exist, particularly the use of multiple discount values, as described in Chen et al. (1999) [51]. This approach is once used for Google Translate under a MapReduce implementation. KenLM is a performant open-

source implementation [102]. Heafield et al. (2013) [102] propose an efficient algorithm to estimate modified Kneser-Ney models, including interpolation.

### 1.4.2 Statistical test on word context

Words are fuzzy, and their meanings vary with the context, narration, and stories in which they are embedded. The discursive concept of bankruptcy can take on different meanings arising from the different legitimisation strategies that companies facing bankruptcy may use. So, we faced the problem of Word-sense disambiguation (WSD), the process of identifying which sense of a word is meant in a sentence or other context segment.

Given the high predictive power of the model, we move forward and investigate language patterns in the narration of bankruptcy and healthy companies and gain insights into how failing companies' language differs from non-failing ones. In particular, we focus on understanding how a word is used in the "fail" or "health" context. So, we explore linguistic features, particularly the use of words in their different contexts. We assume that bankruptcy companies should use some words in a significantly different way than healthy ones. In this respect, we can assign a specific word to the bankruptcy or healthy category by performing a statistical test. So, for each company, we test if its words have context more similar to the bankruptcy or healthy cases. For each word, we collect two lists of cosine similarities from the pairwise comparison of documents related to the two corpora described above. Then, we perform the Bootstrap t-test to compare the means of these two independent samples of word similarities. We set the significant level $\alpha$ equal to 0.01. Then, we assign a label "negative" or "positive" if the mean of the similarities scores is higher in the bankruptcy or healthy corpora, respectively. Tibshirani and Efron (1993) [212] proposed the following test to compare the means of two independent samples. For each document, we represent words in a vector space where each dimension corresponds to a word in the Vocabulary. We defined these vectors for each word as "word's profiles". Then, we account for the contexts surrounding that word considering the windows of length $\pm 2$. Then, to bind all these vectors in a unique vector representing the meaning of that word in the document – the "word profile" – we use the sum of these occurrences. So, for each document, we build the word-word co-occurrence matrix. Once we obtain the word profiles for every word for a specific company, we measure the profile of the same word in the population of bankrupt documents and healthy documents. Then, we

measure the similarity between two vectors with the cosine similarity, as typically used in literature to compare vectors representing words, sentences or documents. The cosine similarity measures the extent to which the same term used in two different documents (i.e., the word in the document we are testing compared with that word in bankruptcy and healthy documents) are similar.

Let $d_k$ be the vector's profile of word $k$ for the document to test. So, we have two lists of cosine similarity scores as follow:

$$S_{d_k,b} = \{x_1, \ldots, x_i, \ldots, x_n\}$$

that represents the similarity scores for the $k$-th word between the document to test and the documents in the bankruptcy corpora.

$$S_{d_k,h} = \{y_1, \ldots, y_j, \ldots, y_n\}$$

that represents the similarity scores for the word $k$ between the document to test and the documents in the healthy corpora. Then, we perform the test for each word for each company. Each time the result of the test suggests that a specific word is significantly more similar to those of bankrupt companies, we append that word to a bankruptcy word list for the company we are testing. In doing so, we can create specific dictionaries (one for each report we are interested in testing) containing terms that indicate "bankrupt words" for each bankrupt case considered.

## 1.5 Results

This work aims to analyze the text of the MD&A section and predict the bankruptcy of the company that released it in the following year. As mentioned before, we focus on the predictive power of sentences, as sequences of words, in the MD&A and use them to predict company bankruptcy. Moreover, we explore the different disclosure languages between failed and healthy companies in terms of using words with different contexts.

### 1.5.1 Classification performance

We face the task of prediction one year ahed the possible failure of a company analysing the language 's style of the text. To this end, we create two corpora: one related to companies that went bankrupt the year following one of the analyzed MD&A and the other to companies that remained solvent the next year. To extract the texrual features we use the `Sequence prediction model` [4] based on the modified Kneser-Ney smoothing algorithm proposed by Heafield et al. (2013) [102] to estimate these probabilities. Then, we train two Language Models, one for each corpus. The trained models estimate two probabilities, respectively, for all sentences of an unseen document, assigning the probability of how likely a sentence comes from the corpus of bankruptcy and the probability that it comes from the healthy corpus. We apply a leave-one-out cross-validation to the 82 pairs of matched companies' documents to evaluate the model's performance. So, we have two documents in the test set. For each document on the test set and all sentences in the document, we collect the logarithm of the probabilities predicted by the two models (one from the corpus of "will-fail" companies and one from the corpus of "healthy" companies) separately. Then, we compute a score for a document $d_j$ according to the following formula:

$$score_M(d_j) = exp\{\frac{1}{N}\sum_{i=1}^{N} p_M(s_i)\} \tag{1.7}$$

where $N$ represents the number of sentences in the document $d_j$, $M \in \{F, H\}$ identifies from which model, and the log probabilities $p_M(\cdot)$ come from ("will-fail" or "healthy"). So, we collect 82 pairs of scores and we assign a prediction label, $H$ or $F$, according to the following condition:

$$\text{Prediction Label}(d_j) = \begin{cases} F, & \frac{score_F}{score_F + score_H} > \tau \\ H, & else \end{cases} \tag{1.8}$$

Then, we need to find a suitable threshold $\tau$ for each of the 82 iterations to achieve the best accuracy. We select the threshold $\tau$ as follows. In the $i$-th iteration of the cross-validation, we perform a second leave-one-out cross-validation among the remaining 81 pairs. So, we collect 81 $\tau_i$'s that discriminate the 81 out-of-sample matched pairs of inner cross-validation iterations. Then, we verify the prediction power on the $i$-th out-of-sample pair

[4] software Mathematica [1]

of the first cross-validation computing $\bar{\tau}_i$, as the mean of the 81 $\tau_i$'s. Finally, we calculate $\bar{\tau}$ for each of the 82 iterations by which we predict the 82 out-of-sample pairs, achieving an accuracy of 90%. The resulting $\bar{\tau}$ values are higher than 0.514. So, according to Eq.1.8, we need a $\tau$ value higher than 0.514 to discriminate between bankruptcy and a healthy language. Moreover, we try different orders of the Markov Model, and among various trials, since we achieve similar performances, we select the order equal to 2, according to Occam's razor. As shown in the Table 1.1 , the model's accuracy is the highest in the literature concerning bankruptcy prediction through the analysis of MD&A texts.

**Table 1.1:** Prediction of bankruptcy: accuracy performance comparison

| Method | Accuracy | Variables | Study |
|---|---|---|---|
| Language Model | **90**% | Textual | our study |
| State-of-the-art | | | |
| Support Vector Machine | 66% | Altman | Cecchini et al. (2010) [47] |
| | 80% | Textual | |
| | 83% | Textual & Altman | |
| Deep Learning Models | 63% | Accounting & Market | Mai et al. (2019)[139] |
| | 57% | Textual | |
| | 71% | Textual, Accounting & Market | |

In Table 1.1, we compare our model with the others proposed in the literature with the goal of bankruptcy prediction of public companies in the U.S., including text variables retrieved from the MD&A section of 10-K reports.

Cecchini et al. (2010) [47] analyze a balanced sample with 78 companies that went bankrupt between 1994 to 1999 and 78 other healthy companies and test the accuracy of a Support Vector Machine classifier with a leave-one-out cross-validation. They extract textual features and represent documents as a vector of concepts. They map words into concepts through Word Net, a lexical database of semantic relations between words, and assign a score to each concept based on its ability to help discriminate between two corpora made by bankruptcy and healthy companies. They prove the informative power of the text information showing that the use of textual variables solely improves the accuracy of the

classifier based on Altman variables. The Altman's bankruptcy discrimination function is as follows:

$$AltmanZ_{score} = \beta_1 \frac{WorkingCapital}{TotalAssets} + \beta_2 \frac{RetainedEarnings}{TotalAssets} + \beta_3 \frac{EBIT}{TotalAssets} + \beta_4 \frac{MarketValueofEquity}{TotalLiabilities} + \beta_5 \frac{Sales}{TotalAssets} \quad (1.9)$$

Moreover, they show that combining the variables enhances the accuracy. In the second work, Mai et al. (2019) [139] used a sample of $11,827$ companies, among which 477 are bankrupt cases, from 1994 to 2014. They use word embedding representation [154] as textual variables and combine them with accounting-based and market-based predictor variables provided by Compustat North America. Then, they test different neural network architectures, randomly splitting the dataset by selecting 80% as the training set and the remaining 20% as the test set. They show that the model with only financial variables performs better than the model with only textual variables. However, they point out that the model achieves better accuracy when combining textual and numerical variables, showing the informative power of textual variables.

### 1.5.2 Language of Bankruptcy

Then, we focus our attention on linguistic features. In particular, we want to understand how a word's context influences its association with failure or non-failure. Specifically, to perform the bootstrap t-test described above, we build two corpora where we consider the normalized difference between the scores that the two models provide for each sentence of a given MD&A. So, for each company, we can divide all sentences into two categories, which we call "negative" and "positive", according to the higher log probability between the two different trained models. Then, we select the corpora made by negative sentences of bankrupt companies and the other one by the positive sentences of healthy companies. We consider these two corpora to investigate the meaning of words in two opposite cases. As described in the previous section, we represent each word in a document as a vector of co-occurrences and measure its similarities to the words of other documents. Once we obtain the cosine similarities, we search for the bankruptcy language category. We test if the mean of the two samples is different. Then, for each company, we have a list of words with profiles more similar to the ones related to bankrupt companies than healthy companies. Here, we focus our attention on the sentiment word list of Loughran and McDonald [135]. In particular, we find that some words tagged as "positive" could

have a different meaning in different contexts. The Table 1.2 below lists some representative sentences for each corpora. These results show the efficacy of the Language Model in managing the local context of words and understanding where the sentences are more likely to come from.

**Table 1.2:** Different use of positive words in sentences of the two corpora

| **Stemmed words** | Negative sentences & bankrupt companies | Positive sentences & healthy companies |
|---|---|---|
| **advantag** | We may also be prevented from taking **advantage** of business opportunities that arise because of the limitations imposed on us by such restrictive covenants | Subject to financing alternatives, we may also increase our capital expenditures significantly to take **advantage** of opportunities we consider to be attractive |
| **improv** | As we continue the exploitation and development drilling in the Mid-Continent, we expect to show **improvement** in our operating results | Average net price **improved** from the prior year due to a two percent increase in average selling prices and a more favorable product mix |
| **satisfi** | Our inability to generate sufficient cash flow to **satisfy** our debt obligations, including obligations under the notes, or to obtain alternative financing, could materially and adversely affect our business, financial condition, results of operations and prospects | Allocate the transaction price to the performance obligation in the contract, and recognize revenue as the entity **satisfies** performance obligations |
| **success** | We provide no assurances that we will be able to **successfully** consummate the Restructuring or other alternatives to restructure our existing indebtedness, in which case we may need to restructure under the Bankruptcy Code | Our production continues to grow through drilling **success** as we place new wells on production and through additions from acquisitions partially offset by the natural decline of our natural gas and oil reserves through production and asset sales |
| **opportun** | The Company's debt agreements impose significant operating and financial restrictions which may prevent the Company from executing certain business **opportunities**, such as making acquisitions or paying dividends, among other things | Consistent with our history of growth, we intend to continue to expand our store base in existing markets and penetrate new markets when suitable **opportunities** can be found |
| **suffici** | Additional sources of liquidity in the future as a result of our inability to generate **sufficient** cash flow from operations to service our long-term capital needs | Cash flow from operations and available borrowings under our revolving credit facility will be **sufficient** to meet our liquidity needs in the coming twelve months |

## 1.6 Conclusions

Our work contributes to previous literature investigating bankruptcy using textual variables by confirming past findings on the incremental information of textual data and demonstrate that a bankruptcy language exists. Our work intends to offer theoretically and methodologically contributions to literature in text analysis of corporate narratives.

From a theoretical perspective, we highlight the fuzziness of words contained in written narratives. The meaning of words varies with the context, narration and stories in which they are embedded. So, a word may have a negative tone/sentiment or not. The results show that there is neither a single dictionary indicative of bankruptcy nor a unique polarization for a specific word. Therefore, each bankruptcy has its own story and narration, and so does every word used in a bankrupt document. The discursive concept of bankruptcy can take different meanings arising from the different legitimization strategies that companies facing bankruptcy may use. Our proposal overcomes the issue of word sense disambiguation inherent in the classic dictionary approach. We demonstrated that a bankrupt language category exists, and it is characterized by the contexts in which words are used.

From a methodological point of view, we propose an approach with high potentiality in written narratives investigation. Our method deals with n-grams and improves automatic coding on a statistical basis, without human involvement in annotating sentiments in sentences. Indeed, our proposal provides good prediction performance and interpretable outputs that could give insight into why a company went into bankruptcy, allowing the investigation of the sentences classified as “negative" and moving away from fixed word lists [135]. Indeed, our study extends the bankruptcy prediction literature investigating the predictive power of a company's textual disclosure in annual reports through a stylometric analysis of language, providing deeper insight into the ”language of default". To our knowledge, this study is the first analysis of bankruptcy prediction using the Language model. This approach allows us to select specific “red flag patterns" (one for each report) regarding negative sentences or words with a bankruptcy profile. Moreover, Language Models are non-parametric models, at least for the choice of the order $n$, and don't need fine-tuning during the training phase as the neural network models. Our approach aims to retrieve the linguistic style of companies' reports instead of computing sentiment

analysis. So, our approach is free from a possible bias in the dataset used to train the sentiment classifier. Indeed, it is acknowledged that analysing phrases or n-grams rather than words would provide a better unit of analysis for investigating corporate narratives through textual analysis [94]. Finally, our results demonstrate that we could effectively construct a statistical language model for predicting the corporate default and bridge the performance gap between the deep learning and dictionary-based approaches.

Indeed, our proposal could have practical implications for practitioners and regulators in monitoring credit risk. Moreover, nowadays, audit companies – such as Ernst & Young, PWC and KPMG – and government regulatory agencies – such as the security and exchange commission (SEC) – are deepening the potentiality of text analysis and natural language processing.

## 1.7 Limitations and Future results

Despite the encouraging results, this work is not without limitations. In particular, one of the main limitations is the sample size. We used a smaller sample than other studies that have coped with textual analysis in finance and accounting research [133, 139, 34]. Nonetheless, other studies have used a similar sample size and we relied upon the same testing method used in these studies [47]. However, an improvement could be expanding our sample and testing our approach using an unbalanced one following sampling procedures suggested in methodological accounting and finance studies [95, 218].

Moreover, our approach meets the expectation of further development in improving automated textual analysis in accounting and finance. In this respect, further, development could be related to the investigation of topics. Indeed, according to managerial literature, there are three stages of the crisis that lead to bankruptcy: strategy, performance, and liquidity crisis [210]. It could be possible to search for these topics by exploring the time evolution of the bankruptcy language. Furthermore, we want to compare our results with well-known financial indicators of default risk, such as Value-at-Risk, expected shortfall and volatility, and with the indicators of the company's status, such as stock prices or returns. Indeed, with a larger sample, we can explore the performance of the proposed method in different market regimes. Finally, we plan to combine linguistic features with financial variables to provide a more integrated analysis of annual reports. Other further

developments could concern implementing this approach to investigate other parts of the report (like the accounting policy section) or different kinds of reports (such as sustainability and integrated reports) and phenomena (such as fraud and integrated thinking).

# Chapter 2

# Ranking coherence in Topic Models using Statistically Validated Networks

**Abstract**

*Probabilistic topic models have become one of the most widespread machine learning techniques in textual analysis. Topic discovering is an unsupervised process that does not guarantee the interpretability of its output. Hence, the automatic evaluation of topic coherence has attracted the interest of many researchers over the last decade, and it is an open research area. The present article offers a new quality evaluation method based on Statistically Validated Networks (SVNs). The proposed probabilistic approach consists of representing each topic as a weighted network of its most probable words. The presence of a link between each pair of words is assessed by statistically validating their co-occurrence in sentences against the null hypothesis of random co-occurrence. The proposed method allows one to distinguish between high-quality and low-quality topics, by making use of a battery of statistical tests. The statistically significant pairwise associations of words represented by the links in the SVN might reasonably be expected to be strictly related to the semantic coherence and interpretability of a topic. Therefore, the more connected the network, the more coherent the topic in question. We demonstrate the effectiveness of the method through an analysis of a real text corpus, which shows that the proposed measure is more correlated with human judgement than the state-of-the-art coherence measures.*

## 2.1 Introduction

The scientific interest in automatic textual analysis has grown dramatically over the last decade. The task of extracting meaningful information from texts has become more important due to the increase in available digital textual data. Indeed, researchers from several disciplines have become increasingly interested in incorporating textual data in their works. One of the most critical goals of text mining is the clustering task [7], studied in different research domains such as data mining [30], machine learning [149], and information retrieval [225]. Topic modeling [33] is one of the most popular probabilistic clustering algorithms, since it aims to process extensive collections of texts that are useful for tasks such as classification, novelty detection, summarisation, similarity and relevance judgments.

These models learn topics automatically, from unlabeled documents in an unsupervised way. These topics are called **hidden thematic structure** or latent topics and are typically represented as sets of essential words. Documents are considered as a mixture of topics, where each topic is represented by a probability distribution of words [32]. Thus, these models build latent topics as multinomial distributions of words and the models assume that each document can be described as a mixture of these topics. [48]. Once the models are trained, they provide a framework for humans to understand document collections both directly by "reading" models or indirectly by using topics as input variables for further analysis [37]. The Latent Dirichlet Allocation (LDA) is one of the most popular topic models and the state-of-the-art unsupervised machine learning technique for extracting thematic information (topics) from a collection of documents. Indeed as highlighted by Boyd-Graber et al. (2017) [37], LDA plays an essential role in the analysis of historical documents, scientific documents, fiction, poetry and literature. The main obstacle in topic detection models is that not all the estimated topics are of equal importance and not all correspond to genuine domain themes. Some of the topics can be a collection of irrelevant words or unchained words representing insignificant themes. Often, in qualitative studies, the goal is to find meaningful and interpretable topics. Researchers usually use top-N words with the highest probability given a topic [124, 159, 6, 182], and employ humans to obtain an interpretability score. Indeed, topic discovering algorithms do not automatically provide a way to interpret their output. For instance, Chang et al. (2019) [48] state that "Although there appears to be a longstanding assumption that the latent space discovered by topic models is meaningful and useful, evaluating such assump-

tions is difficult because discovering topics is an unsupervised process". Moreover, Hoyle et al. (2021) [109] highlight that automated evaluation metrics often suffer from inconsistency. Therefore, it would be desirable to fully automize the process by introducing a metric that automatically ranks learned topics closely matching human judgments. This challenge motivated recent research on topic quality metrics that closely match human judgement. Within this framework, quantifying the coherence of a set of words plays a central role [8, 124, 159, 6, 160, 182, 189]. In topic models, a topic can be viewed as a set of words that frequently co-occur in the same documents, which is very similar to latent word groups (or communities) [235] in the word network. Since words that frequently co-occur in the same sentences are closely connected in the semantic space, they tend to appear in the same document.

This paper proposes a new topic coherence measure based on the construction and analysis of Statistically Validated Networks (SVNs) of words [214]. Specifically, the method builds a co-occurrence network for each topic whose most probable words are the nodes. We set a link between two nodes (words) in each network if their co-occurrence in sentences is statistically significant. We claim that these links carry relevant information about the structure of the topic, i.e., the more connected the network, the more semantically coherent the corresponding topic. Therefore, we propose to use connectivity measures on the SVN of words to build a metric of topic coherence.

The main contributions of this paper are: i) to define a new coherence measure ($Coh_{SVN}$) based on a rigorous statistical model that approximates human ratings better than state-of-the-art methods; ii) to filter out marginal associations of words and to facilitate the graphical representation and interpretation of the obtained topics through Statistically Validated Networks (SVNs) [214].

## 2.2 Background and related works

The main idea of topic modeling is to create a probabilistic generative model for a corpus of text documents. A probabilistic topic model is a type of generative model that aims to learn the latent semantic structure of a corpus. Probabilistic topic models reduce the complex process of document generation to a small number of probabilistic steps by assuming exchangeability, because only word occurrence information (i.e., fre-

quencies) is considered. The first probabilistic topic model was the Probabilistic Latent Semantic Analysis (pLSA), introduced by Hofmann (1999) [105]; unfortunately, the model does not provide any probabilistic model at the document level. Then, Blei et al. (2003) [33] proposed the Latent Dirichlet Allocation (LDA) model as an extension of the pLSA, introducing a Dirichlet prior on mixture weights of topics per document. The name of the model incorporates its main features. Specifically, the term *Latent* indicates that the model involves probabilistic inferences for extrapolating missing probabilistic pieces of the generative story from texts. The term *Dirichlet* recalls that the model uses Dirichlet parameters to encode sparsity. Finally, the name includes the word *Allocation* since the Dirichlet distribution encodes the prior probability for each document's allocation of the topics [37]. In these models, documents are described as random mixtures over latent topics, where a distribution of words characterizes each topic [33]. The words of the documents are the observed variables, whereas the topic structures are the hidden variables. The problem of inferring the hidden topic structure from the documents consists in computing the posterior distribution of topic structures, that is, the conditional distribution of the hidden variables given the documents [32].

Recently, many other probabilistic topic models that consider topic correlations were proposed, such as the Correlated Topic Model (CTM) [31], the Pachinko Allocation Model (PAM) [130]. Other works extend probabilistic topic models focusing on the evolution of topics over time, such as the Dynamic Topic Model (DTM) [71], or introducing word embedding representation - the Embedded Topic Model (ETM) by Dieng et al. (2020) [72].

Finally, neural topic models represent a broader set of related models. These mainly focus on improving topic modeling inference through deep neural networks [203]. Finally, Blei et al. (2012) [32] and Boyd-Graber et al. (2017) [37] provide comprehensive reviews of probabilistic topic models. Among these models, we applied our coherence measure to the LDA model, since it represents a benchmark in the topic modelling community, for comparison with its various extensions. However, it is worth highlighting that the proposed measure applies to any topic model.

### 2.2.1 Literature review

Evaluating the quality of the latent spaces provided by topic models is a difficult challenge because discovering topics is an unsupervised process that gives no guarantees on the interpretability of its output. In text mining, the problem of semantic evaluation has attracted much interest breaking down the research into coherence measures [189]. There is no gold-standard list of topics to compare against for every corpus. Thus, a technique for evaluating the outputs of topic models could be employed on gathering exogenous data. In this section, we discuss previous work on the topic evaluation. For many years, the primary way to evaluate the quality of a topic model was to measure the log-likelihood of a held-out test set [33, 217]. The held-out likelihood consists in density estimation on a collection of unseen documents given a training set. The most commonly used measure based on the held-out method is the perplexity, a monotonically decreasing function of likelihood:

$$perplexity(D) = exp\Big\{ - \frac{\sum_{d=1}^{M} \log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d} \Big\},$$

where $D$ is the collection of documents, $N_d$ is the number of words in document $d$, and $p(\mathbf{w}_d)$ is the marginal distribution of document $d$, following the notation used in previous section. A lower perplexity score indicates better generalization performance. However, Chang et al. (2009) [48] showed that the perplexity on held-out test set emphasizes *complexity* rather than *interpretability*, which is the property users are mostly interested in. In their work, they fit three different topic models to two corpora and demonstrated that the perplexity scores are negatively correlated with human ratings. In other words, such measure is useful for evaluating the predictive performance of the model, but it do not address the more explanatory goals of topic modeling. Indeed, topic models are mainly used to organize, summarize and help users to explore large corpora, while evaluating the predictive performance of the model is a completely different task. Therefore, there is no technical reason to suppose that held-out accuracy corresponds to better organization or easier interpretation. In recent years, many methods have been proposed for assessing topic coherence. The approaches can be split into two categories: qualitative methods and quantitative methods. Qualitative methods are less common than quantitative, since require the use of human resources for topic assessment, and are time-consuming. Quantitative approaches, on the other hand, seek to automate the whole evaluation process

trying to replicate human judgment.

### 2.2.2 Qualitative methods

Chang et al. (2009) [48] proposed the task of *word intrusion* to create a formal setting where humans can evaluate the latent space of a topic model. This task allows for an evaluation of whether a topic has human-identifiable semantic coherence or not. In the *word intrusion* task, the subject is presented with six randomly ordered words, and the task of the user is to find the word which is out of place or which does not belong with the others, i.e., the *intruder*. In 2018, Morstatter and Liu [157] proposed a modified version of the word intrusion task, named *Model Precision Choose Two*. As in the word intrusion task, they propose to form a list with the top (most likely) five words from a topic and to inject one low-probability word from the same topic into the list. The critical difference with word intrusion is that they ask the annotators to select *two* intruded words from the six. The intuition behind this experiment is that the annotators' first choice will be the intruded word, just as in Chang et al. (2009) [48]. However, their second choice is what makes the topic's quality clear. In a coherent topic, the annotator will not be able to distinguish a second word as all of the words will appear similarly coherent.

### 2.2.3 Quantitative methods

The qualitative methods are time consuming since they require the manual annotations of humans. In the last decade, researchers have proposed to fully automating the process by introducing a metric that allows to automatically rank learned topics. One of the first automated measure was proposed by Alsumait et al. (2009)[8]. They introduced an approach to **automatically** rank the LDA topics based on their semantic importance and, eventually, to identify junk and insignificant topics. Their idea is to measure the amount of "*insignificance*" that an inferred topic carries in its distribution by measuring how "different" the topic distribution is from a "*junk*" distribution. In the same work, Al-Sumait et al. proposed three definitions of Junk and Insignificant (J/I) topic distribution, namely: i) the Uniform Distribution Over Words *(W-Uniform)*, ii) the Vacuous Semantic Distribution *(W-Vacuous)* and iii) the Background Distribution *(D-BGround)*. Finally, to quantify the difference between an estimated topic and a J/I distribution, three different distance measures are employed, namely: Kullback-Leibler (KL) Divergence; Cosine

Dissimilarity; and Correlation Coefficient. Later, Wang et al. (2011) [219] proposed a re-ranking algorithm to select "significant" topics by topic similarity calculation. Specifically, each topic is represented as a probability distribution $p(w_i|z_j)$ over words. To compute the distance between word-topic distributions they employed the Jensen-Shannon distance (a symmetrised extension of the KL divergence):

$$Dist(z_i, z_j) = \frac{1}{2}[KL(z_i||z_j) + KL(z_j||z_i)].$$

Finally, for each topic $i$, they computed the average distance between $i$ and all the other topics, and they sort the average distance for each topic in a queue. The last element in the queue is ranked the highest. In the framework of topic quality evaluation, many relevant works make use of the top-N most probable words (rather than using the entire word-topic distribution), and they assess pairwise semantic cohesion among them through their co-occurrences provided by the dataset or external sources. The general idea is to compute the mean of the sum of the pairwise scores of the top-N words that most contribute to describing the topic:

$$Coherence = \frac{2}{N \cdot (N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} score(w_i, w_j) \tag{2.1}$$

One of the best-known topic quality measures based on the top-N words was proposed by Newman et al. (2009) [158]. They introduced for the first time, a model that uses external text data sources, such as Wikipedia and Google hits, to predict human judgements. Specifically, Newman et al. (2009) [158] measured co-occurrence of word pairs, taken from the list of the ten most probable words in a given topic, using two huge external text datasets: all articles from English Wikipedia and the Google n-grams data set. Specifically, they identify a co-occurrence of words $w_i$ and $w_j$ if they occurred together in a 10-word window of any Wikipedia article. Similarly, they identify a co-occurrence of the two words according to Google n-grams if they both appear in any of the existing 5-grams. Finally, they measure the score of association between word pairs through the Pointwise Mutual Information (PMI) [35]:

$$PMI(w_i, w_j) = \log \frac{p(w_i, w_j) + \epsilon}{p(w_i)p(w_j)}, \tag{2.2}$$

where $p(\cdot)$ is the relative frequency of a word and $p(\cdot,\cdot)$ is the relative frequency of the co-occurrence of two words, while $\epsilon$ is a smoothing term. This measure is also called *UCI*. Minmo et al. (2011) [156] pointed out that "bad" topics can be categorized into three definitions:

- *Chained*: every word is connected to every other word through some pairwise word chain, but not all word pairs make sense.
- *Intruded*: either two or more unrelated sets of related words, joined arbitrarily, or an otherwise good topic with a few "intruder" words.
- *Random*: no clear, reasonable connections between more than a few pairs of words.

In their work, the authors suggest that these poor-quality topics could be detected using metrics based on word co-occurrences within the documents. They proposed to use an asymmetrical confirmation measure, $UMass$, between top word pairs (smoothed conditional probability), where the estimations of word probabilities are based on their frequencies in the original documents used to train the algorithm on the topics:

$$UMass(w_i, w_j) = \log \frac{D(w_i, w_j) + 1}{D(w_j)}, \tag{2.3}$$

where $D(w_i)$ is the *document frequency of word*, (i.e., the number of documents that contains $w_i$, and $D(w_i, w_j)$ is *co-document frequency* (i.e., the number of documents containing both words). Note that Eq. 2.3 is equal to the empirical conditional log-probability $\log p(w_i|w_j) = \log \frac{p(w_i,w_j)}{p(w_j)}$ smoothed by adding one to $D(w_i, w_j)$, where $p(w_i) = \frac{D(w_i)}{M}$. Therefore, the score function is not symmetric as it is an increasing function of the empirical probability $p(w_j|w_i)$, where the probability of $w_i$ is higher than the word $w_j$, given a topic. Therefore, this score measures how much (within the words used to describe a topic) a common word, $w_i$, is on average a good predictor for a less common word, $w_j$. Another important contribution was given by Lau et al. (2014) [124] who proposed to use the Normalized Pointwise Mutual Information (NPMI) [35] of word pairs in the automated methods of word intrusion and observed coherence:

$$\text{NPMI}(w_i, w_j) = \frac{\text{PMI}(w_i, w_j)}{-\log\left[p(w_i, w_j) + \epsilon\right]}, \tag{2.4}$$

where $p(\cdot)$ and $p(\cdot,\cdot)$ are defined as for PMI. The NPMI ranges between (-1,+1) resulting in -1 (in the limit) for never occurring together, 0 when they are distributed as expected under

independence, and +1 (in the limit) for complete co-occurrence. Aletras and Stevenson (2013) [6] proposed a method for determining topic coherence using the distributional similarity between the $n$ most likely words of the topic. Representing each word as a vector, let $\overrightarrow{w_1}$,$\overrightarrow{w_2}$,..., $\overrightarrow{w}_n$ denote the vectors of the top $n$ most probable words in a topic. The authors also assume that each vector consists of $N$ elements (the size of the Vocabulary) and $\overrightarrow{w}_{ij}$ is the $j$th element of vector $\overrightarrow{w_i}$. The semantic space was created using Wikipedia as a reference corpus and a window of ±5 words. Then they compute the similarity between words using three measures:

- Cosine similarity:

$$Cos(\overrightarrow{w_i}, \overrightarrow{w_j}) = \frac{\overrightarrow{w_i} \cdot \overrightarrow{w_j}}{\|\overrightarrow{w_i}\|\|\overrightarrow{w_j}\|}$$

- Dice coefficient:

$$Dice(\overrightarrow{w_i}, \overrightarrow{w_j}) = \frac{2\sum_{k=1}^{N} \min(\overrightarrow{w_{ik}}, \overrightarrow{w_{jk}})}{\sum_{k=1}^{N}(\overrightarrow{w_{ik}} + \overrightarrow{w_{jk}})}$$

- Jaccard coefficient:

$$Jaccard(\overrightarrow{w_i}, \overrightarrow{w_j}) = \frac{\sum_{k=1}^{N} \min(\overrightarrow{w_{ik}}, \overrightarrow{w_{jk}})}{\sum_{k=1}^{N} \max(\overrightarrow{w_{ik}} + \overrightarrow{w_{jk}})}.$$

Then, the coherence of topics is constructed by the mean of all pairwise scores. Each of these measures estimates the distance between a pair of words in a topic and produce a topic cohesion measure based on distributional semantics. Roder et al. (2015) [189] proposed a framework that allows for the construction of existing word-based coherence measures as well as new ones, by combining elementary components. They conducted a systematic search of the space of coherence measures for the evaluation and they identified a complex combinations (named $CV$) as the best performers on their test corpora. Omar et al. (2015) [161] quantitatively describe topics via normalized mean values of pair-wise word similarities. They used two types of word similarities, namely, thesaurus and local corpus-based as the descriptive features of a topic, and performed topic classification by using the represented topics as input and a binary 0-1 human ratings. Some of the latest work in the field was produced by Nikolenko et al. (2017) [160]: they highlighted that the topic coherence defined by Minmo et al. (2011) [156] is able to consistently identify bad

topics (i.e., topics with poor coherence) but does not perform well in identifying good ones (i.e., topics with a high degree of coherence). To cope with this problem, Nikolenko et al. (2017) [160] proposed *tf-idf* (term frequency - inverse document frequency) coherence as a modification of Mimno's coherence metric that accounts for the informative content of the topics. Their idea is to introduce *tf-idf* scores instead of the number of co-occurrences in order to construct their measure. The *tf-idf* value, as defined by Salton and Buckley (1988) [192], increases proportionally to the number of times a word appears in a document and is inversely proportional to the number of documents in the corpus that contain that word. This measure privileges the words that not only frequently occur in a given text, but that also occur rarely in other texts. Thus, a coherence metric with *tf-idf* scores penalizes co-occurrence of common words that have low discriminative power. The measure for a given topic is defined as follow:

$$C_{tf\text{-}idf}(w_i, w_j) = log \frac{\sum\limits_{d:w_i,w_j \in d} tf\text{-}idf(w_i, d) tf\text{-}idf(w_j, d) + \epsilon}{\sum\limits_{d:w_i \in d} tf\text{-}idf(w_i, d)}, \tag{2.5}$$

where $\epsilon$ is a smoothing count usually set to either 1 or 0.01, while the *tf-idf* metric is computed with augmented frequency:

$$tf\text{-}idf = tf(w, d) \cdot idf(w, d),$$

where

$$tf(w, d) = \Big( \frac{1}{2} + \frac{f(w, d)}{\max_{w^* \in d} f(w^*, d)} \Big),$$

$$idf(w, d) = log \frac{|D|}{|\{d^* \in D : w \in d^*\}|}.$$

## 2.3 Method

In this section, we propose a new coherence measure to evaluate the interpretability of the top words of a topic. Our method consists in building a co-occurrence network for each topic whose most probable words (according to the estimated topic model) are the nodes. The weights of links are calculated as the number of sentences in which the connected words co-occur. In each network, we identify the links whose weight is statistically significant, i.e.,

those that cannot be explained in terms of random co-occurrences of words in the sentences. Although several measures in the literature have already considered co-occurrence between words as a measure of association, none has undertaken a statistical approach based on hypotheses testing to assess whether the co-occurrence obtained between two words can be attributed to chance or whether these links carry relevant information about the structure of topics. To do this, we exploit Statistically Validated Networks.

### 2.3.1 Statistically Validated Networks

In recent years, many complex systems have been represented by bipartite networks [87, 176, 112]. The Statistically Validated Network, introduced by Tumminello et al. (2011) [214], is an unsupervised method to statistically test the significance of each link of a projected weighted network as obtained from a multipartite network. It is an unsupervised method that introduces a system of hypotheses for link testing when a multipartite network is projected into a set of nodes. The idea is to represent text data as a bipartite network, Figure 2.1, in which the set of nodes $S$ is made by the sentences of a corpus and the other set of nodes $W$ is made by a list of words associated with a given topic. A link is set between a word and a sentence if the word belongs to that sentence. Therefore, projecting the set of words, the resulting network is a word-co-occurrence network [235, 164].

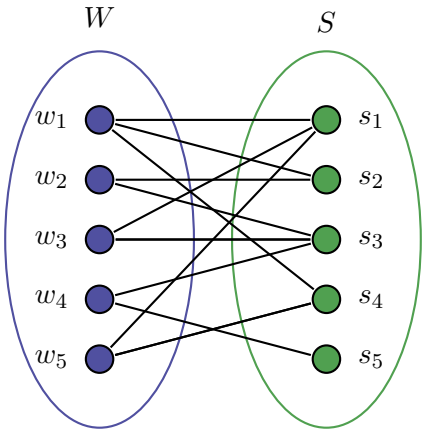



**Figure 2.1:** Bipartite network

To take into account the heterogeneity of the set of sentences, a suitable system of hypotheses is introduced. The hypothesis test is constructed as follows. Let us consider a corpus made of $N$ sentences, then consider two words, say, $w_i$ and $w_j$, and indicate with $X_{ij}$ the times they appear in the same sentences. We are interested in validating the co-occurrences of the words $w_i$ and $w_j$ statistically against a null hypothesis of random

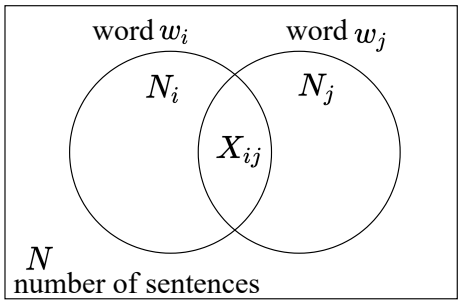



**Figure 2.2:** Venn Diagram showing the overlap of two words

co-occurrence that accounts for the heterogeneity of the considered words, that is, the total number of times they appear individually in the text, $N_i$ and $N_j$, respectively. The probability distribution that describes the random co-occurrence is the hypergeometric distribution, according to which, the probability of observing $X_{ij}$ co-occurrences is given by

$$\mathrm{pmf}_H(X_{ij}|N, N_i, N_j) = \frac{\binom{N_i}{X_{ij}}\binom{N-N_i}{N_j-X_{ij}}}{\binom{N}{N_j}} \tag{2.6}$$

where parameters $N_i$ and $N_j$ naturally allow for the incorporation of the aforementioned heterogeneity of words in the null hypothesis. The Hypergeometric distribution describes the probability mass function under the null hypothesis in which the probability of co-occurrence between words is conditioned by their marginals, i.e., their individual occurrences. The distribution introduced can be used to test the presence of an excess of co-occurrence between any pair of words, $w_i$ and $w_j$. Indeed, assuming that the actual co-occurrences of these words is $N_{ij}$, then the probability that a value larger than or equal to $N_{ij}$ is observed by chance, according to the null hypothesis, is:

$$p_v(N_{ij}|N_i, N_j, N) = \sum_{X=N_{ij}}^{min(N_i,N_j)} \frac{\binom{N_i}{X}\binom{N-N_i}{N_j-X}}{\binom{N}{N_j}}. \tag{2.7}$$

To claim that the number of co-occurrences, $N_{ij}$, between words is too large to be consistent with the null hypothesis of random co-occurrences, we shall set a threshold $\alpha$ of statistical significance. However, since we are facing multiple and dependent comparisons, errors of the first kind are a real issue. Therefore, we use the conservative Bonferroni correction [155] for multiple hypothesis testing. The correction states that given a univariate threshold of statistical significance, $\alpha$, then the threshold corrected for multiple hypothesis testing is $\alpha_T = \frac{\alpha}{T}$, where $T$ is the total number of performed tests, be they dependent or otherwise. The Family Wise error rate (FWER) is the probability of rejecting at least one

true hypothesis, that is, of making at least one type I error. The advantage of the Bonferroni correction is that it provides a very strict control of the FWER, even when tests are dependent, as they are in this case, since the same word appears in many tests. Moreover, since a co-occurrence between two words indicates a semantic relation, we focus more on controlling false positives than false negatives because we are interested in selecting the strongest semantic relationships among words.

### 2.3.2 Coherence based on SVNs

In this section, we describe how to construct the new coherence measure, $Coh_{SVN}$, which makes use of Statistically Validated Networks as combined with different word similarity indices. Specifically, our algorithm can be summarised in the following 5 steps, also sketched in the diagram reported in Fig. 2.3:

(A) Estimate a topic model, and extract the top-$m$ words from each estimated topic;

(B) Represent each topic as a Statistically Validated Network of words;

(C) Evaluate each link's importance, $Imp(w_i, w_j|z_k)$ by considering the strength of the association between word pairs and the relative relevance of each word in the topic;

(D) Compute a global measure of coherence, $Coh_{SVN}$, for each topic network;

(E) Produce the final ranked list of topics, by sorting them in decreasing order of coherence.

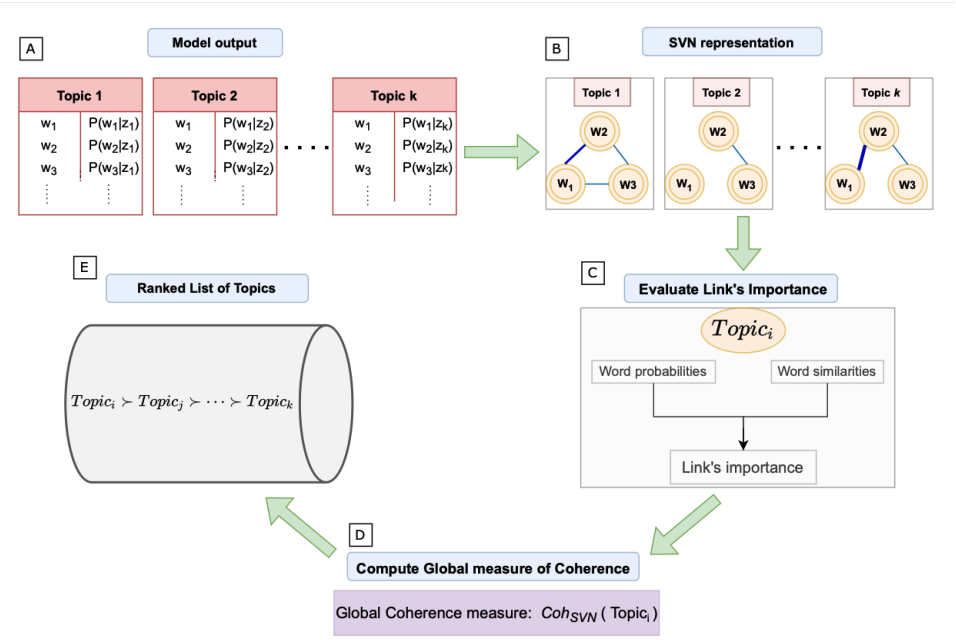



**Figure 2.3:** Diagram describing the 5 steps of the algorithm.

Regarding the first step, the specific topic model used, the parameter tuning and the choice of the optimal number of topics lay outside the scope of this paper. Relevant insights on these subjects can be found in references [16, 120, 193, 56]. The estimation of the LDA model provides a list of $K$ latent topics, each one described by an ordered list of words. So, to conclude the first step, we select the $m$ most probable words[1]. To build the SVN of a given topic, we perform $\frac{m(m-1)}{2}$ statistical tests (against the null hypothesis of random co-occurrence), one for each pair of words, and we set the value of $\alpha$ of Bonferroni correction to 0.01. The results are $K$ weighted Statistically Validated Networks with $m$ nodes and a number of links equal to the number tests that rejects the null hypothesis of random co-occurrence at a given level, $\alpha$, of statistical significance, after the Bonferroni correction for multiple hypothesis testing. An example is shown in Figure 2.4.

[1] In the present application, we follow the standard approach of setting $m = 10$.

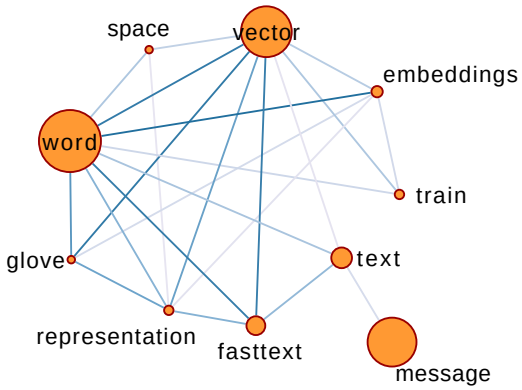



**Figure 2.4:** Statistically Validated Network of an artificial topic.

The size of each node $i$ in Figure 2.4 is proportional to the probability $P(w_i|z_k)$ that the corresponding word $w_i$ appears in the topic $z_k$, while the opacity of each link is proportional to the strength of the association between the linked words.
To compute the strength of each validated link, we use corpus-based word similarities within distributional contexts. Specifically, let $N$ denote the total number of sentences in the corpus, $N_i$ and $N_j$ the occurrences of words $w_i$ and $w_j$, respectively, in the sentences of the corpus, and $N_{ij}$ their co-occurrence. To calculate word similarities we use four metrics already used in other studies. Specifically:

- $S_1$: Jaccard similarity index [184]

$$J(w_i, w_j) = \frac{N_{ij}}{N_i + N_j - N_{ij}} \tag{2.8}$$

- $S_2$: Dice-Sorensen coefficient [70] [2]

$$Dc(w_i, w_j) = \frac{2N_{ij}}{N_i + N_j} \tag{2.9}$$

- $S_3$: Sokal and Sneath coefficient [202]

$$SS(w_i, w_j) = \frac{N_{ij}}{2N_i + 2N_j - 3N_{ij}} \tag{2.10}$$

[2] Notice that it is equivalent to F1 score.

- $S_4$: Fowlkes–Mallows index [84]

$$FM(w_i, w_j) = \sqrt{\frac{N_{i,j}^2}{N_i \cdot N_j}}. \tag{2.11}$$

Furthermore, we also consider three metrics that are tightly related to the SVN method. These metrics are:

- $S_5$: Similarity based on the Pearson's correlation coefficient $\rho(w_i, w_j)$:

$$D_\rho(w_i, w_j) = \frac{1}{2}\left[1 + \rho(w_i, w_j)\right] \tag{2.12}$$

where

$$\rho(w_i, w_j) = \frac{N_{ij} - \frac{N_i N_j}{N}}{\sqrt{N_i(1 - \frac{N_i}{N})N_j(1 - \frac{N_j}{N})}} \tag{2.13}$$

Since the expected value of the Hypergeometric distribution $H(X|N, N_i, N_j)$ is $\frac{N_i N_j}{N}$ and the variance $\mathbb{V}[X] = \sigma_H^2 = \frac{N_i N_j}{N}\frac{N-N_i}{N}\frac{N-N_j}{N}$, it turns out that $\rho(w_i, w_j)$ is proportional to the Z-score of $N_{ij}$ under the null hypothesis[3].

- $S_6$: Normalized logarithmic robustness $\tilde{R}$

$$\tilde{R}(w_i, w_j) = \frac{\log_{10}(N) - \log_{10}(N^*|w_i, w_j)}{\log_{10}(N) - \log_{10}(n^*|w_i, w_j)}, \tag{2.14}$$

where

$$N^* = \min\{N : p_v(N_{ij}) < \frac{\alpha}{T}\},$$

is defined as the minimum number of sentences needed in the corpus to validate the co-occurrence between $w_i$ and $w_j$. While,

$$n^* = \min\{N : p_v(N_{ij}^*) < \frac{\alpha}{T}\}$$

is the minimum value of sentences needed to validate the co-occurrence between $w_i$ and $w_j$ assuming a perfect co-occurrence, $N_{ij}^* = \min(N_i, N_j)$.

[3] The constant of proportionality is $N^{-\frac{1}{2}}$.

- $S_7$: Similarity based on the normalized p-value $\tilde{p}_v$

$$\tilde{p}_v(w_i, w_j) = 1 - \frac{p_v(N_{ij}|N_i, N_j, N)}{\alpha/T}, \tag{2.15}$$

  where $p_v(N_{ij}|N_i, N_j, N)$ is computed following Eq. 2.7.

All of the proposed similarity measures, $\{S_1, \ldots, S_7\}$, take values in the range $[0, 1]$ where 0 indicates two totally unrelated words, while 1 indicates two perfectly associated words. Given a validated link between two words, say $w_i$ and $w_j$, belonging to the topic $z_k$, we define the link's importance $Imp(w_i, w_j|z_k)$:

$$Imp(w_i, w_j|z_k) = \sqrt{P(w_i|z_k)P(w_j|z_k)}\; S_h(w_i, w_j), \tag{2.16}$$

where $S_h$ is one of the similarity function described above: $\{D_\rho, \tilde{R}, \tilde{p}_v, J, Dc, SS, FM\}$. The importance of a validated link (Eq. 2.16), between $w_i$ and $w_j$ give a topic $z_k$, takes into account two components:

- the relative relevance of $w_i$ and $w_j$ within $z_k$:

$$\sqrt{P(w_i|z_k)P(w_j|z_k)};$$

- the strength of the association between $w_i$ and $w_j$:

$$S_h(w_i, w_j),\ h = 1, \ldots, 7.$$

The conditional probabilities $P(w_i|z_k)$ and $P(w_j|z_k)$ reflect the relevance of words $w_i$ and $w_j$, respectively, within the topic $z_k$. That is to say, words with a higher probability are more relevant within a topic. Therefore, the more relevant two terms, the more important the validated link between them. We decided to use the geometric mean of $P(w_i|z_k)$ and $P(w_j|z_k)$ as aggregating function to reduce the impact of the distribution's tails. As regards to $S_h(w_i, w_j)$, it measures the association between $w_i$ and $w_j$. Intuitively, the higher the association between two words, the greater the importance of the link between them. Note that, if $w_i$ and $w_j$ exhibit a "perfect" co-occurence, i.e., $N_i = N_j = N_{ij}$, then $S_h(w_i, w_j) = 1$ and the link's importance reduces to $Imp(w_i, w_j|z_k) = \sqrt{P(w_i|z_k)P(w_j|z_k)}$, that is, the geometric mean of the words probabilities, given the topic, provided by the

model.

Finally, we define the **global coherence** measure of a topic, $z_k$, as:

$$Coh_{SVN}(z_k) = \frac{\sum\limits_{w_i \neq w_j, \in \mathcal{L}} Imp(w_i, w_j | z_k)}{\sum\limits_{w_i \neq w_j, \in \Omega_k} \sqrt{P(w_i|z_k)P(w_j|z_k)}}, \tag{2.17}$$

where $\mathcal{L}$ is the set of word pairs linked in the SVN, while $\Omega_k$ is the set of all possible $m \cdot (m-1)/2$ word pairs for topic $z_k$.

In Eq.2.17, the denominator represents the coherence of a perfectly coherent topic, that is a fully connected network where all the pairwise word similarities are maximized, i.e. $S_h(w_i, w_j) = 1 \ \forall w_i, w_j \in \Omega_k$. Thus, $Coh_{SVN}(z_k)$ ranges in the set $[0, 1]$, where the minimum value indicates a totally incoherent and unintelligible topic, while a value of 1 represents a perfectly coherent topic.

Measure $Coh_{SVN}(z_k)$ allows us to rank topics in decreasing order of coherence, which completes the fifth (and final) step of the procedure presented in this section.

## 2.4 Experimental evaluation

### 2.4.1 Dataset and pre-processing

We evaluated our estimator of topic quality on a dataset of articles extracted from the *New York Times*, which was already analysed by Xing et al. (2019) [227]. The dataset (NYTd from now on) consists of 8,764 articles of the *New York Times*, which appeared between April and July 2016[4]. In particular, we decided to consider a reduced version of this dataset, obtained by removing all the articles with fewer than 20 total words (Hong and Davison (2010) [107] discuss how short documents can confuse topic modeling algorithms), and taking a random sample of size 1,000 out of those remaining. The following step is to perform data preprocessing in order to reduce noise from the data. The preprocessing usually consists of tasks such as filtering meaningless parts of text, and either lemmatization or stemming words. Lemmatization and stemming are two text normalization techniques for Natural Language Processing. The first one is the process of finding the base or dictionary form of a word, called *lemma*, with the aim to remove only inflectional endings considering morphological analysis as meaning and context. Instead, stemming is a method to convert words into their root form by cutting the suffix or

[4] https://www.kaggle.com/nzalake52/new-york-times-articles

prefix from the word. Comparing the lemmatization and stemming methods, we opted for the lemmatization. Stemmed words, in general, are very complicated to interpret, since roots of words were insufficient to discriminate among alternative meanings [196]. We removed urls, mails, punctuation and numbers from the texts through the `Python regex` function. Furthermore, we used the `gensim` library to construct compound words, such as *United_States* or *North_Korea*, and `spaCY`, an open-source natural language processing library for Python, to split up sentences. Finally, we removed i) infrequently used words (i.e. appearing only once per document); and ii) redundant words (a rule of thumb is to remove terms appearing in more than 80% of the documents). As a matter of facts, infrequently used terms will not contribute much information about topics, while discovery and removing them may greatly reduce the size of the vocabulary [67]. Equally, it has been shown that redundant words appearing frequently do not convey any meaningful message for topic modeling [18]. The original corpus dictionary, as directly obtained from the 1,000 articles, consisted of $28,104$ tokens, whereas the final corpus (after data preprocessing) included $8,770$ tokens. The LDA model was trained in `R` setting 50 topics [216], then we randomly extracted 30 of them for human judgment evaluation. We have chosen to use only part of the group of estimated topics due to time constraints. Indeed, we structured the questionnaire so that each annotator took, on average, 15 minutes to complete their task, assuming an average response time of about 30 seconds per topic. This issue is crucial for maximising the quality of the answers obtained; in fact, a questionnaire which takes too long to be completed entails the risk of receiving unreliable answers as the respondent's focus drops. Finally, we prepared graphical representations of the networks of topics using `Cytoscape` software.[5]

### 2.4.2 Coherence-based topic annotations

To obtain high-quality ratings, the survey was structured in two steps. During the first step, which we call "pilot", 23 PhD students from the Department of Economics, Business and Statistics at the University of Palermo, Italy, were brought in. We provided them with 32 topics (consisting of 10 words each) to be evaluated on a 5-point scale where 5="coherent" and 1="not coherent". Among topics, 30 were genuine topics according to the LDA model as applied to the New York Times dataset, and the remaining two were synthetic (control) topics. The first synthetic topic included a group of unrelated words that formed

[5] https://cytoscape.org

a meaningless and incoherent topic, $z_{31}$ = {Lasagna; Finance; Jeans; Buddhist; Pokemon; Drive; Molecule; Sound; Chess; Revolver}. Instead, the second synthetic topic included perfectly coherent words that formed a strongly coherent topic, $z_{32}$ = {Black; White; Red; Green; Pink; Purple; Brown; Yellow; Grey; Blue}.

We also provided textual guidelines on how to judge whether a topic was coherent or incoherent. In addition to showing several examples of such topics we provided the following preliminary instructions to the respondents.

**Guidelines**

*Topic modeling* consists of the automatic extraction of groups of words, called *topics*, from a collection of texts. For a topic to be "coherent", it must make sense and be interpretable. This means that the topic's words must:

1. be related to each other
2. belong to the same theme

An automatic procedure for the identification and evaluation of topics is reliable if the topics identified are coherent and interpretable for humans. This is why we are asking you to be part of a benchmark sample of individuals to test the effectiveness of a new topic modeling algorithm we are working on. Therefore, we ask you to rate the coherence of specific topics on a scale of 1 to 5. For example, you can give a topic a low mark if you find few links between the words in it, the mark increases as the number of linked words increases.

It is not always easy to evaluate a list of words, especially if some of them are unfamiliar or belong to a language other than yours (in this case, English). We ask you, *PLEASE*, we ask you to translate any words or nouns you do not know to give as informed a mark as possible.

You will notice that some topics share one or more words; this is not a problem! The topics are not related to each other, so each topic must be evaluated individually. There is no right or wrong answer, since we aim to collect your subjective opinion.

The role of the pilot was to assess the topic annotators' ability in understanding their assigned task. We also investigated which improvements were necessary in letting annotators deepen their comprehension of the meaning of "coherence". The most critical issue in

the pilot was to investigate whether an odd scale was appropriate. Thus, we studied the relationship between the percentage of neutral answers given by an annotator (i.e. giving a grade of 3) and their probability of failing at least one control topic evaluation.

**Table 2.1:** Relationship between giving neutral answers and failing at least one control topic evaluation

| | *Fail control* | | |
|---|---|---|---|
| *Neutral responses* | No | Yes | Total |
| $\leq 30\%$ | 14 | 1 | **15** |
| $> 30\%$ | 2 | 6 | **8** |
| Total | **16** | **7** | **23** |

Table (2.1) shows that these two features are strongly related since the odds ratio [194] is equal to $\frac{14\times6}{2\times1} = 42$. As a matter of fact, many studies [175, 208] showed that some respondents quickly select the midpoint on the 5-point scale as a dumping ground [57]. Such attitude can be explained in psychological terms:*"choosing a minimally acceptable response as soon as it is found, instead of putting effort to find an optimal response"*[57]. Therefore, we could easily identify "unreliable annotators" that do not produce reliable judgments, by looking at the respondents who fail the control topics. The results of the pilot survey informed our decision to provide the final survey annotators with the same guidelines, but we asked them to evaluate the coherence of topics on a scale from 1 to 4 to discourage annotators from expressing neutral responses. The final survey was designed to obtain human judgments to be used as ground truth for comparing our method with state-of-the-art coherence measures. The annotators of the final survey were 222 PhD students from various departments of the University of Palermo; in this way, we employed highly educated judges with heterogeneous knowledge within the sample. The 222 judges were asked to assess the coherence of 32 topics (30 genuine and 2 artificial topics) on a Google Form[6]. Table 2.2 reports the control topics' scores manual assigned by the 222 annotators. Overall, about 90% of the total (202 out of 222 annotators) succeeded in evaluating both control topics. In the case of the highly coherent topic $z_{32}$, we considered the ratings equal to 4 to "be successful" since a group of words containing only colours should receive the maximum rating. At the same time, we regarded ratings of 1 or 2 as a success for the incoherent coherent topic $z_{31}$.

[6] https://docs.google.com/forms/d/e/1FAIpQLSdoWQsO3MLMcQZDatkCkrSWaThuuj2D-Wm7sR18cy3x8XiRhw/viewform

**Table 2.2:** Control topics' scores assigned by annotators. Annotators are highlighted in red.

| | | Topic $z_{32}$ scores | | | | |
|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** | Tot |
| Topic $z_{31}$ scores | **1** | 1 | 1 | 2 | 192 | 196 |
| | **2** | 0 | 1 | 4 | 10 | 15 |
| | **3** | 0 | 0 | 2 | 4 | 6 |
| | **4** | 0 | 2 | 0 | 3 | 5 |
| | Tot | 1 | 4 | 8 | 209 | 222 |

Fig 2.5 reports the frequency distributions of the scores assigned by the annotators to the 30 genuine topics, removing the annotators who failed at least one control topic evaluation.

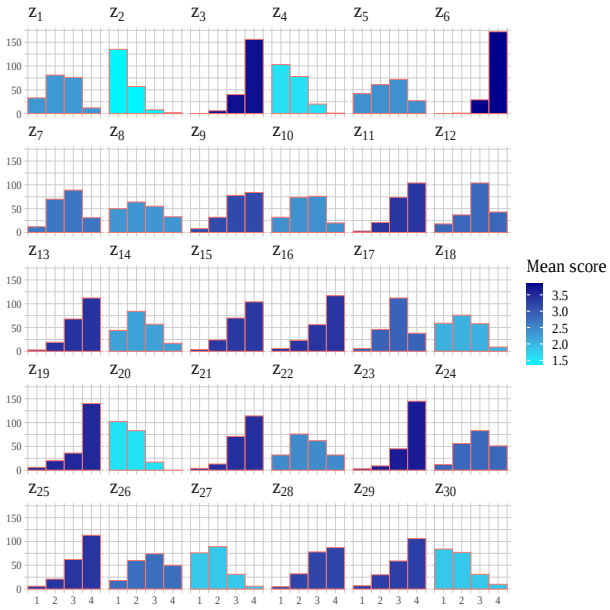



**Figure 2.5:** Annotators' coherence evaluations

The final dataset contains: i) the list of the most probable words, ii) the coherence ratings given by evaluators, and iii) the document term matrix used in our study. It is available upon request from the authors.

### 2.4.3 Data analysis and results

To compare the effectiveness of the proposed method in replicating human judgment with respect to the other coherence measures proposed in the literature, we collected the results of the survey and re-arranged them in the form of ranking data. Specifically, a

ranking $\pi$ is a mapping function from the set of topics $\{z_1, \ldots, z_{30}\}$ to the set of ranks $\{1, ..., 30\}$, endowed with the natural ordering of integers; $\pi = (\pi(1), \pi(2), ..., \pi(m))$ where $\pi(z_j)$ is the rank given to topic $z_j$. In our setting, conditioning to a specific coherence metric, the topic with the highest coherence score will be ranked 1 and the topic with the lowest coherence score will be ranked 30. Therefore, we build two matrices:

- the matrix of scores $\mathbf{S}_{30\times13}$, where the generic $s_{ij}$ element represent the coherence score of the $z_j$-topic assigned by the $i^{\text{th}}$ metric. As regards the last column, i.e. human judgment, the $z_j$-topic is given the average coherence score assigned by human evaluators. (see Table 2.3 for a reduced version of the matrix, and Table B1 for the full matrix);
- the matrix of rankings $\mathbf{R}_{30\times13}$, where the generic $r_{ij}$ element represent the relative rank of the $z_j$-topic assigned by the $i^{\text{th}}$ metric. In this matrix, the estimated topic coherences are compared with each other, in order to establish a preference ordering: from the most coherent to the least coherent topic. (see Table 2.4 for a reduced version of the matrix, and Table B2 for the full matrix).

**Table 2.3:** Coherence scores: the **S** matrix

| Topic | $D_\rho$ | $\tilde{p}_v$ | $\cdots$ | HumanJ |
|---|---|---|---|---|
| $z_1$ | 0.076 | 0.133 | $\cdots$ | 2.332 |
| $z_2$ | 0.049 | 0.084 | $\cdots$ | 1.391 |
| $z_3$ | 0.159 | 0.265 | $\cdots$ | 3.743 |
| $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ |
| $z_{30}$ | 0.100 | 0.150 | $\cdots$ | 1.837 |

**Table 2.4:** Ranking coherence scores: the **R** matrix

| Topic | $D_\rho$ | $\tilde{p}_v$ | $\cdots$ | HumanJ |
|---|---|---|---|---|
| $z_1$ | 26 | 26 | 25 | 23 |
| $z_2$ | 29 | 29 | 30 | 30 |
| $z_3$ | 18 | 18 | 20 | 2 |
| $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ |
| $z_{30}$ | 22 | 23 | 28 | 26 |

To evaluate the correlation between human judgments and the topic quality scores predicted by all the automatic metrics, we use the Emond and Mason's rank correlation coefficient, $\tau_x$ [81] (which reduces to Kendal correlation coefficient $\tau_b$ if there are no ties, see [172, 5] for an in depth discussion on the correlation measures focusing on the rankings). The higher the $\tau_x$, the better the metric is at measuring topic quality. In addition,

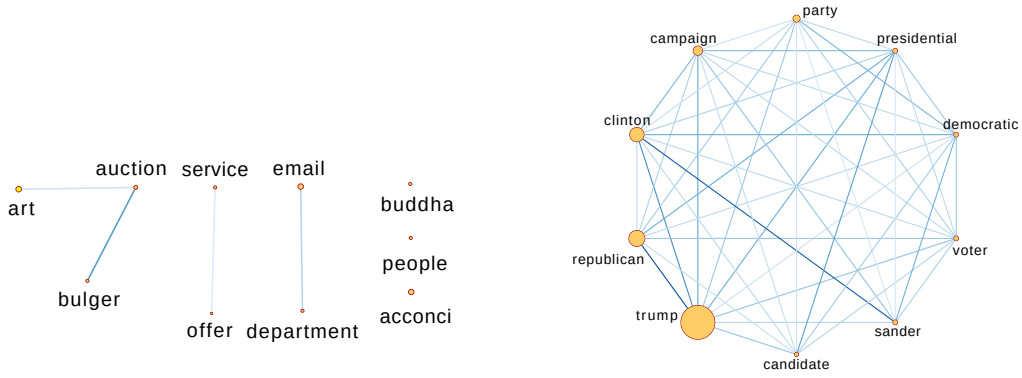
to conforming our comparison procedure to the literature standard, we also computed the Pearson's linear correlation coefficient [124, 189] and the Spearman's rank correlation coefficient [159, 6, 157], see table B3 in Appendix B. Although these two measures have been frequently used in the literature, we argue that they are not particularly suitable in this framework. On the one hand, the Pearson's correlation coefficient only considers the linear correlation between two vectors, which is undoubtedly restrictive for our purpose, and its value may be seriously affected by only one outlier [61]. On the other hand, the Spearman rank correlation suffers from the so-called *sensitivity to irrelevant alternatives*, that is: adding extra irrelevant objects to the ranking exercise could change the maximum agreement solution. This issue has been identified by Emond and Mason (2000) [80], it is due to the fact that Spearman's correlation estimator treats the ranks as numerical values instead of categorical ordered values. Moreover, Croux and Dehon (2010) [61] highlighted that the Spearman rank correlation has a smaller gross error sensitivity (GES) (low robustness) and a greater asymptotic variance (AV) (low efficiency) compared to the Kendall $\tau_b$ and $\tau_x$. These features make Spearman coefficient a less preferable estimator from both perspectives. Table 2.5 reports the $\tau_x$ rank correlation between human judgments and all the considered metrics. We compared the correlations obtained either by keeping ("with noise" column of Table 2.5) or removing ("without noise" column of Table 2.5) the unreliable annotators. The results show that the proposed SVN Coherence measure, based on $D_\rho$, outperforms all the baselines. We also set the value of $\alpha$ to 0.05, but the correlation with human judgments is lower. Indeed, we compare the rankings of $D_\rho$, with $\alpha = 0.01$ and $\alpha = 0.05$, and the resulting correlation coefficient is 0.91, proving the robustness of our method.

**Table 2.5:** Emond and Mason $\tau_x$ rank correlation coefficient with human judgments for metrics.

| Method | *Correlation with human judgement* $\tau_x$ with noise | $\tau_x$ without noise |
|---|---|---|
| $Coh_{SVN}$ | | |
| $J$ | 0.621 | 0.632 |
| $Dc$ | 0.616 | 0.627 |
| $SS$ | 0.616 | 0.627 |
| $FM$ | 0.708 | 0.714 |
| $D_\rho$ | **0.721** | **0.728** |
| $\tilde{R}$ | 0.579 | 0.586 |
| $\tilde{p}_v$ | 0.698 | 0.705 |
| *State-of-the-art* | | |
| PMI [158] | 0.616 | 0.618 |
| $UMass$ [156] | 0.565 | 0.563 |
| NPMI [124] | **0.685** | **0.687** |
| $CV$ [189] | 0.570 | 0.572 |
| $tf\text{-}idf$ [160] | 0.629 | 0.636 |

### 2.4.4 Interpretation of the resulting topics

In this section, we report a comparison between $Coh_{SVN}$ and human judgment in evaluating the coherence of some estimated topics. In Figure 2.6, topics for which there is high concordance between human judgement and $Coh_{SVN}$ are reported.

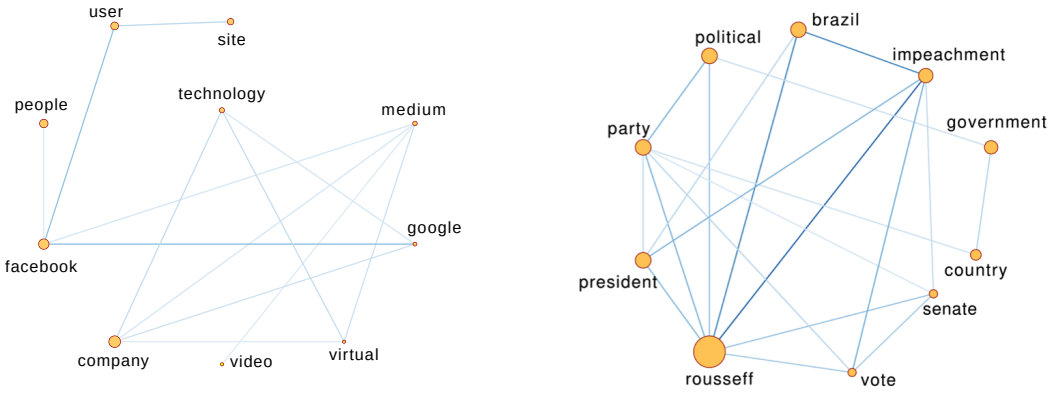



**(a)** *Topic* $z_2$ **(b)** *Topic* $z_6$

**Figure 2.6:** SVN representation of Topic $z_2$ and Topic $z_6$

In particular, Figure 2.6(a) represents topic $z_6$, which is the most coherent topic. It has been assigned an average score equal to 3.84 (first in the rank) by the annotators. Likewise, $Coh_{SVN}$ scores it 0.545, which make it the most coherent in the final ranking. As a matter of fact, topic $z_6$ can be considered a genuine theme of the domain, i.e., a politically themed topic where all the top words can be associated with US politics.

Therefore, the annotators quickly recognized that the words are strongly related, and the co-occurrences in the corpus reflect their solid semantic association. Topic $z_2$, in figure 2.6(b), is one of the least coherent topics. Annotators rated it with an average score of 1.37 (last position in the ranking). Besides, the topic's $Coh_{SVN}$ score is equal to 0.049, which corresponds to the second-to-last position in the ranking. The SVN constructed on topic $z_2$ reveals that the words composing it are mostly unrelated; therefore, there are few statistically validated links. Figure 2.7 report topics whose scores (and, consequently, the ranking) assigned by the annotators are not consistent with our coherence measure.




**(a)** *Topic* $z_3$ **(b)** *Topic* $z_{28}$

**Figure 2.7:** SVN representation of Topic $z_3$ and Topic $z_{28}$

Topic $z_3$ (figure 2.7(a)) has been positively evaluated by the annotators; the average score is equal to 3.74 (second in the ranking). Instead, $Coh_{SVN}$ places it 18th in the ranking, with a score equal to 0.159. The annotators considered the words in topic $z_3$ to be related to each other, but the semantic associations detected by humans are not reflected by the co-occurrences in the reference corpus. For example, the words *Facebook* and *company* are not linked in the resulting Statistically Validated Network. This issue could be due to the structure of the corpus used in the analysis. As a matter of fact, the statistical significance of word pairs' co-occurrences can also be validated including external text data sources, such as Wikipedia or Google hits, rather than using only the corpus sentences. Alternatively, one could use paragraphs instead of sentences to count co-occurrences, but if the text is not properly formatted it might prove difficult to identify the paragraphs. Finally, topic $z_{28}$ is reported in figure 2.7(b). The corresponding $Coh_{SVN}$ score is equal to 0.264, the 7th in ranking. While, according to the survey, it has an average score equal to 3.22, and it is 12th in ranking. In this case, the topic is considered

to be more coherent by $Coh_{SVN}$ than by humans; however, the discrepancy between the automatic measure and the human judgement is less relevant than in the previous case. Overall, about 20% of the annotators did not recognise a central theme and rated it with a low score (1 or 2). This issue could be since topic $z_{28}$ refers to a specific political event that took place in Brazil between 2015 and 2016. Moreover, it contains "hard-to-interpret" terms such as *Rousseff*, a little-known proper name, and *impeachment*, a technical term referring to the political sphere. Indeed, the evaluation of the topic is more complex than the other ones and requires respondents to carry out in-depth research.

### 2.4.5 Summary of main findings

In summary, according to the presented analysis, $Coh_{SVN}$ represents a new topic coherence measure that:

- follows a rigorous statistical model of co-occurrence based on multiple hypotheses testing, while state-of-the-art measures pass over the randomness of co-occurrence;
- ranges between $[0, 1]$, providing a more readable framework for evaluating the coherence of the topics;
- approximates human ratings better than state-of-the-art methods (see Table 2.5);
- allows the graphical representation and interpretation of the obtained topics through Statistically Validated Networks (SVNs)[214];
- is less sensitive to the text preparation since it considers co-occurrences of word pairs in sentences. Instead, most of the measures proposed in the literature, as summarised in the paper by Röder et al. (2015) [189], use a sliding window to calculate the co-occurrences, which makes these methods very sensitive to the preprocessing steps.

## 2.5 Conclusions

One of the fundamental challenges in topic detection models is assessing the semantic *coherence* of estimated topics in terms of human interpretability. State-of-the-art coherence measures focus on the marginal probabilities of words and their co-occurrence. However, none of them takes into account the randomness of co-occurrences. In this work, we undertake a rigorous statistical approach based on hypotheses testing to develop a new

topic-coherence measure, $Coh_{SVN}$. To automatically evaluate how semantically close the top words of the topics are, we represent each topic as a weighted network of its most probable words. The presence of a link between two words indicates that their co-occurrence in sentences is statistically significant against the null hypothesis of random co-occurrence. The proposed global measure of coherence, $Coh_{SVN}$, is derived by considering the number of statistically validated links, the strength of the association between word pairs, and the relative relevance of each word in the topic. To prove the effectiveness of our method, we administered a survey on 222 PhD students from University of Palermo, Italy, and construct a benchmark dataset of human judgements. These judgments were taken as ground truth, and it was shown that the proposed measure reproduces human judgment more closely than the state-of-the-art (Table 2.5). As for future research, the results reported in this paper suggest to explore the possibility to develop a topic similarity index based on Statistically Validated Networks and including NLP tools, e.g., entity recognition and part-of-speech tagging. Finally, the development of a rigorous statistical method for validating the similarity between two topics could prove beneficial, following the theory of recommendation systems [234], to promote *diversity* in the final ranking of topics. Indeed, the ordered list of topics could be determined by considering both the point-wise quality score ($Coh_{SVN}$) and the correlations between topics.

# Chapter 3

# Networks and text mining approach to perform systematic literature reviews

**Abstract**

*Scholars conduct systematic literature reviews (SLR) to summarize what we know and discern what we should know about a specific theme. Machine learning (ML) can support researchers conduct systematic literature reviews. We present an ML approach based on network analysis and Natural Language Processing (NLP) that allows extracting textual features to categorize papers. The method consists of an algorithm that allows to: (a) select relevant studies on a specific theme; (b) discern the main topics around the theme. Additionally, we offer two applications of the toolkit. Specifically, we select relevant studies and discern the main topics around two themes: cobranding and coopetition. We juxtapose our results with previous systematic literature reviews on the abovementioned themes. We show how ML may boost the rigour of SLR by improving their transparency, completeness, saturation, and universalism.*

## 3.1 Introduction

Systematic literature reviews (SLRs) frame the current state of the art of a specific theme to gather insights and set what directions we should move forward [122, 123]. Given the proliferation of business journals, the mushrooming numbers of articles in journal is-

sues, the increasing fragmentation of studies on specific themes, and the rapid dissemination of empirical findings [78], the recourse to SLRs has turned more commonly than in the past. Correspondingly, scholars are committed to boosting methodological rigour in SLRs. Nonetheless, SLRs remain "bespoke, haphazard, and inconsistent across the universe of studies" [199]. One of the most critical aspects of conducting an SLRs is selecting articles [98]. There exist several and contradictory protocols. Some SLRs protocols consider the papers published in top journals [44, 117]; others SRLs also include papers published in 3, 4, or 4* ABS journals or FT50 [65, 117]; other SLRs enlarge the spectrum of analysis to the 2 stars ABS journals [181], and finally, some SLRs straightforward consider peer-reviewed journals [14]. Once the authors identify the original sample of papers, regarding the inclusion of the articles for the review, we see several protocols of categorizing the articles as "accepted", "possibly accepted", and "rejected". Some authors clearly stated to have reviewed the full text and remove those that, in their own opinion, are not pertinent to the theme [40]; other times, papers are selected based on reading the abstracts by at least two authors independently [46, 63, 230]. The protocols mentioned earlier allow authors to manage a reduced amount of articles at the expense - at least potentially - of the rigour of SLRs in terms of transparency, completeness, saturation, and universalism [199]. However, time constraints and limited human capacity to manage many papers justify the recourse to such protocols [188].

Recent advances in machine learning (ML), in general, and natural language processing (NLP), in particular, turn potentially valuable for making complex automated tasks and managing vast portions of text [98]. The NLP has been employed in many industries, such as the health industry [52]. ML overcomes the challenge of detecting a portion of texts and recognizing "the knowledge/wisdom in it, specifically within any given time limits" [55] without direct human intervention. Accordingly, ML supports researchers in decoding data, learning from and drawing from those learnings to reach specific purposes [97, 215]. These considerations lead to our research question:

*Can SRL authors boost the methodological rigour by adopting the ML toolkits in selecting of articles?*

This paper provides a method to select papers for SLRs on a given theme and discern the main topics around it. In addition, we offer two applications. Specifically, we select papers and discern the main topics related to the literature on cobranding and coopetition. A few

motivations justify our focus. First, recent SLRs [68, 171] show that both literatures on cobranding and coopetition are now at a mature stage and the copious amount of studies justify a literature review. Second, cobranding and coopetition are two themes explored from different disciplinary perspectives: marketing and strategic management. In doing so, we assess the validity of our methodology to select papers and discern topics within two different disciplines. Third, both literature on cobranding and coopetition encompass the agreements among firms, although they focus on different levels of analysis and types of shared resources [53]. Therefore, juxtaposing the results of our study, we are able to assess the convergence/divergence between the two streams of literature.

We structure this paper as follows. Section 3.2 reports the background of this paper; we focus on the importance of sample selection in SLRs. Section 3.3 illustrates our methodology. Sections 3.4 offers the applications of our proposal to cobrand and coopetition literature. Section 3.5 and and 3.6 provide a discussion of results and summarizes the contributions of our study.

## 3.2 Systematicity in the process of articles selection

Business studies are generally positioned within a "conceptual space" and offer incremental contributions to theorizing around a given theme [140, 180]. Thus, what we know about a specific theme is frequently associated with an "accumulation of knowledge" spread among several articles focusing on specific aspects. For instance, in exploring the core tenant underlying a theme, studies may focus on detecting the antecedents, the consequences, the drivers of success (or unsuccess), the role of the context, and so on [78]. Moreover, frequently, scholars adopt new labels that they have adopted in the past to depict other related constructs [78]. Researchers have progressively increased their abilities to publish in international journals, and the number of journals has gradually increased. These circumstances explain the exponential and fragmented development of the literature on a given theme and, consequently, the proliferation of literature reviews [201]. Literature reviews offer a picture of the accumulated state of knowledge and build a "foundation for advancing knowledge" [222]; they illustrate how research on a specific theme has established the literature into topics that can make a more holistic understanding of that theme available. From an academic perspective, scholars summarize the accumulated state of knowledge as a necessary step to discern current and emerging conceptual understandings [123]. From a

practical perspective, since managers should consider the broadest range of factors before deciding, the literature review benefits them by offering a "big picture" of a given theme [190]. While different approaches exist in conducting a literature review (e.g., systematic, semi-systematic, and integrative), only the SLRs employ rigorous protocols to guarantee a "comprehensive accumulation, transparent analysis, and reflective interpretation of all empirical studies" [190, 201]. SLRs provide a picture of a given theme by summarizing previous literature "according to explicit and reproducible methodology" [92]. Specifically, the authors report the definition of keywords, the choice of database, the selection of papers, and then the identification of topics occurring within a given literature, and the final interpretation.

In this paper, we focus our attention on the selection of papers and the identification of topics as preliminary steps for an effective SLR. A proper selection process should meet the following criteria. First, the *criterion of transparency*: authors may reveal the same processes and reproduce the same methods used to select articles [199]. So, other researchers may reproduce the same selection of papers [103, 177]. Second, the *criterion of completeness*: the selection of articles should include all relevant and essential studies related to the theme[42, 169]. Third, the *criterion of saturation*: the selection of articles to review does not leave out one or more relevant topics related to the theme; hence, all the relevant topics are reported in the map of literature that emerges from the SLR. Eventually, authors may discuss the over-expression of some topics in specific periods. Fourth, the *criterion of universalism*: authors conducting the paper's selection for SLR should assume an impartial perspective that precludes any forms of particularism [199]. Quite surprisingly, multiple contradictory protocols regarding the paper selection [103] make SLRs not consistent and haphazard across studies [199]. For instance, considering only papers published in top journals or the basis of ABS classification may lead to a lack of completeness and saturation. Additionally, we observe the lack of transparency in categorizing the articles in "accepted", "possibly accepted", and "rejected" and a biased perspective in selecting key topics around a given theme.

## 3.3 Method

Recent studies [144, 188] examine the application of ML and NLP in the literature review, providing the power and limits of such applications. For example, Watanabe et

al. (2020) [221] use NLP tools to retrieve concepts from academic papers, and Porciello et al. (2020) [174] implement ML methods to select pertinent papers related to a specific topic. The application of ML methods in literature reviews can be done in three ways [144]: *searching*, *screening*, and *data extraction*. We focus our analysis on *searching*, a tool that relies on applying NLP techniques to organize papers into different categories, and on *screening* which refers to selecting academic literature according to a training step made with human interaction.

### 3.3.1 Pre-processing

The pre-processing of textual data is a fundamental step in text analysis. It helps to reduce noise and remove meaningless parts of the text, such as punctuation, numbers and stop words since they don't provide information, particularly when the analysis aims to discover topics as clusters of words. Then we decided to remove the inflectional ending parts of the words with the stemming technique through the package NLTK implemented in Python. Moreover, we decide to include in the list of stop words also general words related to research such as:
*author, result, studi, research, effect, find, paper, provid, examin, develop, new, evalu, implic, base, investig, categori, context, suggest, purpos, intent, previou, indic, contribut, publish, articl, amongst, book, approach, method, analys, analysi, shed, light, abstract, science, summary, purpose, background, conclusion, chapter, proceed.*
The words reported are in the stemmed form.

### 3.3.2 Background: extraction of textual features

Clustering is one of the primary goals of text mining in many applications, such as document classification and organization. The task regards finding groups of similar texts in a collection of documents. In such applications, documents generally are the basic elements. Text documents can be represented as binary vectors, i.e., considering the presence or absence of a word in the document. So, we can represent a collection of documents in a matrix form, the so-called "document-term matrix", where usually rows are documents and columns are words. Words are the variables of the vector space of dimension $V$, the vocabulary length. This representation is referred to "Vector Space Model" (VSM), and it is suitable for many algorithms that aim to discover thematic information of a large collection of documents. Latent Semantic Analysis (LSA), introduced by Dumais et al. (1988)

[77], is one of the first unsupervised methods to extract a representation of text based on observed words. The model uses Singular Value Decomposition (SVD) to decompose the "document-term matrix" to retrieve a dimensionality reduction to achieve the best representation of documents. A direct extension is the probabilistic LSA (pLSA) proposed by Hofmann (1999) [106], which captures the possibility that a document may contain multiple topics. The clusters of words that occur together are referred to as topics. A topic model can group words with similar meanings and distinguish between uses of words with multiple meanings. Topic modelling is a widely used probabilistic text clustering algorithm that has gained increasing importance in recent years. Probabilistic topic models are statistical methods that aim to extract the hidden thematic structure in a collection of documents, how these themes are connected, and how they change over time. The Latent Dirichlet Allocation (LDA) model is the state-of-the-art unsupervised technique for extracting thematic information (topics) from a collection of texts. It is a bayesian probabilistic model [33], which assumes a fixed number of topics, and each document reflects a combination of these topics. It is closely related to classical principal component analysis [45]. However, the number of topics must be fixed in advance, which is a severe limitation. Choosing the number of topics in LDA is a well-known issue [23]. It is usually established by examining the fit to held-out documents [33] or by selecting based on the marginal probability of the whole collection [93]. Indeed, nonparametric bayesian methods [163] provide a solution leading to an "infinite" topic model. Still, the problem is only shifted to set the values of many hyper-parameters. However, in both cases, the implementations rely on stochastic initialization of values, potentially leading to different outputs using the same parameter values. This issue corresponds to the concept of "instability" [232, 23]. Another well-known limitation of the LDA model regards the application to short documents [228]. Yan et al. (2013) [228] proposed the "biterm topic model", which learns topics over short texts by modelling all the unordered word-pair within a text. However, the model needs hyperparameter fine-tuning, such as the number of topics. The proposed method tries to solve these issues, and it consists of the following steps:

1. Construction of a network of words in terms of semantic similarity

2. Clustering of words to define hidden themes (topics) as groups of strongly connected components of the network

3. Assigning how much a topic is present in a document

### 3.3.3 Statistically Validated Networks

The papers' abstracts under our analysis can be considered paragraphs or short texts. To overcome the abovementioned issues, we apply the Statistically Validated Networks method, introduced by Tumminello et al. (2011) [214]. It is an unsupervised method to statistically test the significance of links in a projected weighted network obtained from a bipartite network. A bipartite network consists of two separate sets of nodes where the links connect only nodes belonging to different sets. So, we represent the document-term matrix as a bipartite network, where the two sets of nodes consist of words and documents (abstracts). A link is set between a word and a document if a word occurs in a document. The proposed method aims to project the set of words and apply a statistical test to validate each link. We set a link between two words if their co-occurrence within the documents is statistically significant against a null hypothesis of random co-occurrence. So, we build a word co-occurrence network through the Statistically Validated Networks method. Here contrary to Chapter 2, we focus on similarities between documents (abstracts). Therefore, we analyse the use of words in the abstracts to retrieve the latent topics behind the collection of texts instead of measuring a certain degree of coherence between two words. So, with this approach, we study words' semantic similarities discovering the hidden thematic structure. Specifically, since we are analysing a homogeneous collection of texts, we find sub-topics of the main theme of cobranding and coopetition, respectively. Moreover, the hypergeometric test allows us to manage the marginal occurrences of words, both frequent and rare, to overcome issues arising from applying topic models in a more heterogenous corpus of documents.

The network construction involves multiple hypothesis testing, and we use the False Discovery Rate (FDR) criterion [28], a less stringent correction than the Bonferroni correction [155], to control Type I errors. The FDR correction is defined as follows. Specifically, we first arrange the p-values of different tests in increasing order ($p_1, < \cdots < p_N$), and we obtain the FDR threshold by finding the largest $j_{max}$ such that $p_{j_{max}} < j_{max} \cdot \alpha/N$. We set the value of $\alpha$ equal to 0.05. Although, the FDR criterion is valid when the tests are independent, as pointed out by Benjamini and Yekutieli (2001) [29], the controlling procedure holds also when the test statistics are positive dependent, which is the case under our analysis. Then, we transform the p-values into correlation coefficients $\rho$ through Eq.(2.13). Once we obtain the weighted network of word co-occurrences, we are interested in finding

groups of similar words. To this end, we apply a community detection algorithm based on modularity optimization. In network science, a community is defined as a group of nodes more likely to connect to each other than to nodes from other communities. Modularity measures the quality of each partition and allows us to infer if a particular community partition is better than some other one. The idea consists of comparing a community's link density with the link density of the same group of nodes obtained by a hypothetical random network structure. Then, we could decide if the original community corresponds to a dense subgraph or if its connectivity pattern emerged by chance. Modularity optimization relies on two central hypotheses. The first one states that a community is a locally dense connected subgraph in a network. The second one asserts that for a given network, the partition with maximum modularity corresponds to the optimal community structure. Therefore, we identify these groups of words as topics or sub-topics of the two datasets' main "general" topics related to coopetition and cobranding. Then, we study how much a topic is expressed in each document/abstract. So, we test the over-expression of topics in each document. To perform the test, we count the words shared by a document and a topic and their marginal counts, considering only the words in the resulting validated network. However, since some documents have no over-expression with any topics, we test if such documents are at least over-expressed on the whole set of validated words, as shown in Figure 3.1. So, we assign these documents to a topic called "General", i.e. documents related to the main topic of cobranding or coopetition.

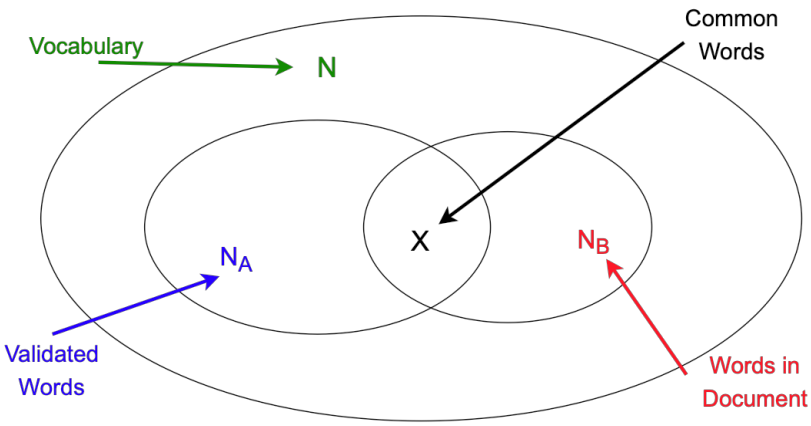



**Figure 3.1:** Venn Diagram showing the overlap

## 3.4 Papers selection and topics discovering

### 3.4.1 Illustrative case 1: cobranding

To provide an exemplification of using the proposed method, we employ it to select papers on cobranding and discern the main topics around this theme. Comparing the findings of our approach and Pinello et al. (2022) [171] supports our evaluation of the effectiveness of ML for SLR. Following Pinello et al. (2022) [171], we select articles that include the terms "cobrand*", "co-brand*", "brand alliance*", "joint branding", "dual branding", "co-marketing alliance", "ingredient branding", and "multiple branding" in their title, abstract, or list of keywords. We extract data from Scopus.

Once download the list of papers, we focus the analysis on the abstracts. Table 3.1 compares the papers on cobranding selected by our method and the selection provided by Pinello et al. (2022) [171]. As described in section 3.3, at this point, we can opt for two approaches: (a) a narrow approach that considers only the internal citation and (b) a broad approach that considers both the internal and external citations. As regards cobranding literature, the results of the two approaches substantially analogous. Several articles have been internally referred to as cobranding literature, resulting in cobranding scholars as community isolated within the wider marketing community. In the rest of this section, we consider the findings - proposed in Table 3.1 - resulting from the narrow approach that considers only internal citations. The number of overlapping papers between the automatic selection and Pinello et al. (2022) [171] is 154, namely about 75% of the selection performed by Pinello et al. (2022) [171]. On the contrary, 25% of our paper selection is inconsistent with Pinello et al. (2022) [171]. Our approach did not select 51 papers that Pinello et al. (2022) [171] considered relevant in cobranding literature. Additionally, Pinello et al. (2022) [171] ignored 52 papers that are included by the ML algorithm, as they are definitively important for cobranding literature.

At this point, we independently read the articles' and found that only 3 papers selected automatically are not considered strictly pertinent to cobranding . The remaining 49 papers out of the 52 selected by our method were disregarded by Pinello et al. (2022) [171]. A selection bias explains this finding in Pinello et al. (2022) [171]: authors focused on recent papers published in top journals and/or higher cited. This procedure led to a lack

of completeness in sample selection.

**Table 3.1:** Comparison the papers selection between ML and Pinello et al. (2022) [171]

| | | Authors | | |
|---|---|---|---|---|
| | | Seletcted | Not selected | Tot |
| Automatic tool | Seletcted | 154 | 52 | 206 |
| | Not selected | 51 | 300 | 351 |
| | Tot | 205 | 352 | 557 |

Another point is whether the lack of completeness in Pinello et al. (2022) [171]'s article selection jeopardizes the saturation criterion. Considering the automatic selection of papers, we extract topics (sub-topics) related to cobranding. Then, at least two authors analyzed the words aiming to discern the main topics related to cobranding. From the analysis, we recognize 22 key topics. Table 3.2 reports the descriptive statistics of topics.

Topic 1 echoes the *conceptualization of cobranding*; we reckon this label by considering terms such as review, framework, heuristics, and originality. Topic 13 is centred on quantitative research to investigate cobranding, as evocated by some words such as regression, logistic, data, collection, and questionnaires. Finally, interviews, semi-structured, qualitative, and in-depth words push us to label topic 17 as a qualitative method to investigate cobranding. The three topics mentioned above are consistent with a common idea in marketing that methodological approaches are complementary among them. For instance, case studies are frequently provided to shed light on the contexts where quantitative studies are performed or supply insights to formulate hypotheses that future studies quantitative test [24]. In addition, quantitative studies contribute to unveiling the micro-dynamics of cobranding configuration, development, and dissolution and detect the causal mechanisms interested [24].

We also recognize eight topics related to industry contexts where cobranding strategy emerges: *sports events* (topic 3), *oil companies* (topic 6), *tourism and destination* (topic 7), *hotels/restaurants* (topic 9), *supermarkets* (topic 15), *biochemical* (topic 14), *higher education* (topic 12), *banking credit system* (topic 22), and *car services* (topic 24). Furthermore, topic 8 regards a more general context: the *product value chain*, which stresses

the importance of cobranding to a practical experience delivery. The other two topics are related to the geographic context with a particular focus on Asian South East and Europe (topics 16 and 22, respectively). All topics mentioned above are related to the importance of factors that compose the context of cobranding in Pinello and colleagues' framework. Specifically, Pinello et al. (2022) [171] consider the country of origin, the product industry, and the cobranding contract specificities as they are critical factors shaping the perception of the brand fit and, in turn, the purchase intentions [66]. While Pinello et al. (2022) [171] recognize the importance of context in cobranding , they fall short of reporting (and hence analyze) the peculiarity of each contexts. By citing Cheah et al. (2016) [50], Pinello et al. (2022) [171] argue about the importance of country of origin as a driver of consumer perception. However, Pinello et al. (2022) [171] did not consider whether and how may apply cross-cultural theories to specific types of brands (luxury vs grocery) [220]. Remarkably, the fact that the eight topics over twenty-two that ML provide are related to context suggests a paramount emphasis in the literature.

Four topics are related to cobrand motivations: the drivers that lead firms to draft and implement a cobranding strategy. specifically, Topic 4 regards the *ingredient brand*; precisely, the words that our toolkit detects around topic 4 are product, supplier, origin, innovation, ingredient, and component. Topic 5 embraces words such as cause, nonprofit and for-profit, donation, and social; thus, it regards *cause-related cobranding*. Topic 2 regards the *consumer perception*, it encompasses the following words: perception, experiment, partnership, design, and co-market. Topic 10 frames the *penetration of foreign market*; Indeed, the words include cross border, country of origin, and countries. All the topics identified by the ML algorithm are traceable in Pinello et al. (2022) [171]. However, different from Pinello et al. (2022) [171], the method did not call specific attention to brand development and equity.

Finally, the last four topics identified by our method are related to cobranding outputs. Specifically, reading the words for each topic,we recognize the importance of the *spillover effect* (topic 11) between allied brands with a particular focus on the effect of the brand name, the *competitive advantage* (topic 19), *dimensions of trust* (topic 20), and finally, the *customer fidelisation* (topic 21). Generally, all outputs we identified using ML toolkit are consistent with Pinello et al. (2022) [171], even if they use a different level of detail.

At this point, we can conclude that despite the lack of completeness in article selection may jeopardize the criterion of saturation, in Pinello et al. (2022) [171], it did not happen. In addition to the personal perspective of author(s) that can jeopardize the criterion of universalism, we argue that bias in the selection of articles may be related to the over/under-representation of topics when authors select specific journals. For instance, Table 3.2 shows the topic "consumer perception" overrepresentation of 202 from our database but only 45 from top journals. It means that "consumer perception" is a core topic for cobranding literature, but it has limited space in top-tier journals. Similarly, Table 3.2 shows "Sports events", "Cause-related", and "ingredient" are central topics in cobrand literature but are practically ignored in top journals. This finding should alert scholars that SLR focuses only on articles published in top-tier journals.

In summary, our study extend the results of previous SLRs by stressing the importance of context, which is de facto more emphasized in the literature than indicated in Pinello et al. (2022) [171]. Similarly, our study enlarges the perspective provided in the meta-analysis proposed by Paydas Turan (2021) [165] that considers the cobranding (vertical vs horizontal) as a theoretical moderator and the type of business (B2B vs B2C) and type of industry (service vs no-service). Furthermore, we echo Chiambaretto and Gu$\check{a}$ru (2017) [53] to call attention to the cobranding types by focusing on the ingredient cobrand that appears to play a pivotal role.

**Table 3.2:** Cobrand: descriptive Statistics of topics as revealed from the abstracts

| Topic | Topic description | Modularity contribution | Number of stemmed words | Number of papers over-expressed | Number of papers from top journals over-represented | Average number of citations | Average number of citations within dataset | Average number of citations within topic |
|---|---|---|---|---|---|---|---|---|
| 0 | General | – | 446 | 83 | 11 | 13.9 | 2.6 | 0.3 |
| 1 | Conceptualization | 0.067 | 34 | 130 | 23 | 18.4 | 3.0 | 1.3 |
| 2 | Consumer perception | 0.065 | 34 | 202 | 45 | 25.7 | 5.7 | 3.9 |
| 3 | Sport events | 0.035 | 17 | 13 | 1 | 11.8 | 0.3 | 0.2 |
| 4 | Ingredient brand | 0.045 | 16 | 42 | 6 | 19.7 | 6.1 | 1.3 |
| 5 | Cause related cobranding | 0.033 | 12 | 19 | 2 | 37.1 | 2.7 | 0.5 |
| 6 | Oil companies | 0.014 | 9 | 3 | 0 | 0.0 | 0.0 | 0.0 |
| 7 | Tourism and destination | 0.015 | 9 | 7 | 0 | 12.9 | 1.3 | 0.0 |
| 8 | Retail/Product value chain | 0.028 | 8 | 10 | 0 | 0.4 | 0.0 | 0.0 |
| 9 | Hotels / restaurants | 0.014 | 8 | 7 | 0 | 6.1 | 1.7 | 0.1 |
| 10 | Market penetration / Heritage | 0.008 | 7 | 5 | 0 | 11.0 | 1.0 | 0.2 |
| 11 | Spillover effect | 0.020 | 7 | 10 | 5 | 29.0 | 10.2 | 1.0 |
| 12 | Higher Education | 0.011 | 7 | 5 | 0 | 4.8 | 0.2 | 0.0 |
| 13 | Quantitative research | 0.009 | 7 | 6 | 0 | 13.3 | 1.5 | 0.2 |
| 14 | Biochemical | 0.018 | 6 | 3 | 0 | 2.3 | 0.3 | 0.3 |
| 15 | Supermarket | 0.007 | 4 | 5 | 1 | 11.8 | 2.8 | 0.0 |
| 16 | Asian South East | 0.005 | 4 | 3 | 0 | 7.7 | 0.0 | 0.0 |
| 17 | Banking credit system | 0.006 | 4 | 6 | 1 | 3.2 | 0.2 | 0.0 |
| 18 | Oil companies Car services | 0.006 | 3 | 3 | 0 | 0.0 | 0.0 | 0.0 |
| 19 | Competitive adbantage | 0.007 | 3 | 8 | 0 | 5.8 | 1.9 | 0.0 |
| 20 | Trust / Reputation | 0.011 | 3 | 3 | 0 | 55.3 | 2.3 | 0.3 |
| 21 | Customer fidelization | 0.005 | 3 | 3 | 0 | 4.7 | 1.3 | 0.0 |
| 22 | Events Germany | 0.004 | 3 | 3 | 0 | 18.0 | 7.3 | 0.0 |

### 3.4.2 Illustrative case 2: coopetition

In this section, we offer an additional application: selecting articles and discerning the main topics around the coopetition strategy. Coopetition regards the simultaneous competition and cooperation between two or more firms [26, 62, 68]. To provide our application of how our toolkit works, we juxtapose the results of our study with Devece, et al. (2019) [68]. Devece, et al. (2019) [68] select papers that include the keywords: "coopet*" or "co-opet*" in the titles; we use the exact keywords, but we extend our research to the title, abstract, or list of keywords. We argue that searching the keywords by focusing on the title leads to a bias, and the search does not include all relevant and essential studies related to the research [41, 168]. Initially, we apply our method by using a narrow selection that considers internal citations. We find 44 articles. Table 3.3 compares the papers on coopetition that we find through the ML algorithm and those included in Devece, et al. (2019) [68]. We find that the number of overlapping papers is 28, about 16% of the selection performed by Devece, et al. (2019) [68]. Then, we launch the ML algorithm using a broad selection that considers SCOPUS citations. We find 66 articles. Table 3.4 show a comparison between the selection of the papers through the broad approach of our ML algorithm and the one provided by Devece, et al. (2019) [68]. We found that the number of overlapping papers is 66, about 86% of the selection performed by Devece, et al. (2019) [68].

**Table 3.3:** Comparison of papers selection (Conservative: with internal citations) between and Devece, et al. (2019) [68]

| | | Authors | | |
|---|---|---|---|---|
| | | Seletcted | Not selected | Tot |
| Automatic tool | Seletcted | 28 | 16 | 44 |
| | Not selected | 49 | 78 | 127 |
| | Tot | 77 | 94 | 171 |

**Table 3.4:** Comparison of papers selection (selection Large) and Devece, et al. (2019) [68]

| | | Authors | | |
|---|---|---|---|---|
| | | Seletcted | Not selected | Tot |
| Automatic tool | Seletcted | 66 | 78 | 144 |
| | Not selected | 11 | 16 | 127 |
| | Tot | 77 | 94 | 171 |

Tables 3.3 and 3.4 lead an additional consideration about the importance of internal citations. The results show that the nature of the theme plays a crucial role in shaping the citations for publications. Coopetition represents an interpretative lens of relationships between competitors and thus helpful to investigate multiple levels of analysis, e.g., business ecosystems [17, 186], divisions [12], and functions within firms [206], and so on. Therefore, the nature of the theme helps researchers develop a high number of citations in papers that are not directly linked to the coopetition. This circumstance represents an essential difference between coopetition and cobranding research.

Table 3.5 reports the descriptive statistics of topics related to coopetition as revealed in the abstracts. From a methodological perspective, we recognize two specific topics in literature: topics 6 and 11. Topic 6 regards specific terms - such as applicability, running, modelling, and test - that echo the fact that studies on coopetition leverage *quantitative analysis and empirical investigations*. Recently, coopetition studies are advancing quantitative investigation to test the bright [224] and dark sides of coopetition [60] and its impact on performance. The growing number of empirical studies on coopetition mirrors the evolution of the field and the fact that coopetition research is progressively shifting from childhood to the young-adulthood stage of evolution [36]. Also, topic 11 encompasses quantitative methods in coopetition, but it focuses on *surveys and moderation effects*. This topic pictures the upsurge ion interest in exploring whether or not coopetition is positively or negatively related to performance and what moderates these relationships [59, 125].

From a thematic perspective, we acknowledge topics regarding the antecedents, management, and consequences of coopetition. We shall depart from considering the antecedents

**Table 3.5:** Coopetition: descriptive Statistics of topics as revealed from the abstracts

| Topic | Topic description | Modularity contribution | Number of stemmed words | Number of papers over-expressed | Number of papers from top journals over-represented | Average number of citations | Average number of citations within dataset | Average number of citations within topic |
|---|---|---|---|---|---|---|---|---|
| 0 | General | – | 601 | 3 | 1 | 125.3 | 0.7 | 0.0 |
| 1 | Coopetition-setting | 0.057 | 42 | 17 | 9 | 29.6 | 0.5 | 0.3 |
| 2 | Positioning coopetition strategic managment | 0.056 | 38 | 14 | 2 | 30.0 | 0.0 | 0.0 |
| 3 | Resources | 0.038 | 25 | 23 | 16 | 116.2 | 1.9 | 0.7 |
| 4 | Co-creation performance | 0.028 | 21 | 7 | 5 | 177.0 | 6.1 | 0.7 |
| 5 | International coopetition | 0.034 | 21 | 10 | 5 | 146.9 | 2.1 | 0.2 |
| 6 | Coopetition quantitative model | 0.028 | 20 | 6 | 3 | 62.7 | 0.8 | 0.2 |
| 7 | Policy | 0.03 | 19 | 4 | 2 | 38.3 | 1.0 | 0.0 |
| 8 | Market failure and third-party role | 0.027 | 14 | 9 | 8 | 33.8 | 0.7 | 0.0 |
| 9 | Coordination | 0.024 | 13 | 4 | 2 | 173.0 | 0.0 | 0.0 |

of coopetition. Topic 1 encompasses aspects related to the *setting in which coopetition occurs*, i.e., the aspects that matter for developing coopetition. Among them, we find some specific words that are particularly evocative: intensity, profit, outsourcing, channel, and competition. Topic 9 considers the *contingencies* to understand how and under which conditions coopetition may occur. We find some particularly evocative words within this topic, such as drivers, insights, and coordination. Such words are consistent with several studies that have widely acknowledged the relevance of unpacking drivers of coopetition [60, 2] and identifying which factors may push firms to collaborate with competitors. Topic 3 sheds light on the role of *resource deployments in developing a coopetition strategy.* Within this topic, we find specific words such as capabilities, innovation, functions, technology, synergies, strengths, sources, and culture. They evocate that firms may cooperate with competitors to access resources they could not otherwise and share risks and costs related to investing in innovation in highly technological and turbulent environments. This topic roots in the studies that draw on the resource-based view and conceive coopetition as instrumental for resource access [187, 137]. Topic 7 depicts the *market exploration and development* on the basis of coopetition. We recognize this topic based on the following words: barriers, intervention, regulation, regulators, and policy. This topic relates to one of the seminal pieces on coopetition that draws from the pie metaphor to identify the relevance of shifting from Porter's five forces schema to a model that includes the role of complementors in a firm's competitive advantage [39]. This approach has emerged in several studies in the smart card industry [150], the tourism industry [85], and the high-tech industry [90]. Topic 8 focuses on *third-party organizations* - such as governments or private clients - in coopetition; they may facilitate cooperation and/or competition between competitors. Specifically, this topic encompasses the following words: formalization, forces, governments, and driving. Usually, the third parties may initiate coopetition by imposing cooperation among competitors [143] or, differently, enhancing competition among co-opetitors [223]. Finally, topic 10 summarizes the *coopetition orientation and experience.* The words we find within this topic are perspectives, requirements, attitudes, alternatives, roles, and orientation. Coopetition requires specific attitudes, such as dealing with tensions underlying the interplay of competitive and cooperative actions.

As regards the management of coopetition, we recognize the following topics. Topic 2 emphasizes the relevance of considering *collaboration and governance models* based on

the interplay of competition and cooperation. Specifically, we find specific words such as strategy, strategically, management, and practice that echo the relevance of developing a practice to deal with coopetition. Topic 12 is related to the *paradoxical view of coopetition*. Words - such as paradox and dual that we found with our method - frequently represent the coexistence of apparently opposites forces that co-occur [126, 200]. Coopetition has widely drawn on paradoxes to explain how to manage the inner tensions between competition and cooperation [183, **?**]). Finally, as regards the consequences of coopetition, we find the following topics. Topic 4 focuses on the coevolution and value of coopetition. Some words such as creation, appropriation, protection, imitation, partner, and firm-specific underscore the need to protect internal resources and innovation from collaborating with a rival. Private and shared benefits coexist in coopetition [115]. Therefore, threats of imitation and the need to protect innovations emerge. Arguably, this topic is also retraceable in SLR proposed by [68] within the discussion on "alliance dynamics". Finally, Topic 5 focuses on multi-market coopetition and global strategies [138]. Within this topic, we find specific terms that are particularly evocative: emergent, regional, multinational, multi-market, global, and economies. The interplay of competition and cooperation affects entry into markets with rival incumbents [118] and the firm's competitive position within a coopetition network [54].

At this point, we can compare ML findings with Devece et al. (2019) [68]. First, we observe that Devece et al. (2019) [68] classify studies using criteria developed in literature - analysis level (as in [27]), method (as in [36]) - or however pre-established (objectives, focus, and firm size). This choice inevitability leads to a rigid classification, and new and emerging topics may be overlooked. Second, we note that the topic of third-party organizations emerged in our analysis generated by the ML algorithm; it is neglected in [68]. While pioneering studies on coopetition investigate the role of third parties in stimulating competition [143] , this idea is less dominant in recent literature. Correspondingly, Devece et al. (2019) [68] did not emphasize this topic because it is linked to the oldest studies on coopetition.

Finally, applying our methodology contributes to coopetition by shedding light on aspects that the recent review of Gernsheimer et al. (2021) [89] overlooked. Our approach stressed the importance of the setting in which coopetition occurs. Similarly, Gernsheimer

et al. (2021) [89] focus on partner interdependence. However, while ML algorithm calls attention to resource deployments and market power, Gernsheimer et al. (2021) [89] refer to mutual benefits. Furthermore, both studies called attention to contingencies. Interestingly, Gernsheimer et al. (2021) [89], like Devece et al. (2019) [68] neglect to consider the importance of third-party organizations. Some of the topics related to the management of coopetition are reported in our study and Gernsheimer et al. (2021) [89]: the *collaboration and governance model and the paradoxical view of coopetition* (that Gernsheimer et al. (2021) [89] consider as quest for separation and integration and the emergence of tensions). Other variables that Gernsheimer et al. (2021) [89] discern are, however, traceable in the above mentioned topics (value creation and appropriation, trust and opportunisms etc.). Gernsheimer et al. (2021) [89] call attention to the impact of coopetition on organizational learning and innovation and, more generally, firms performance, our approach evocates the importance of coevolution and value. Surprisingly, this aspect is not reported in the study. Instead, our results stress the impact of coopetition from a multimarket and global perspective. On the contrary, we did not find a crucial role in the interplay of coopetition and sustainability. Above mentioned comparison among Devece et al. (2019) [68], Gernsheimer et al. (2021) [89] and the results of our method underscore the importance of an impartial perspective in reviewing literature and the biased role of authors in selecting the topics to analyze in dept.

## 3.5 Discussion

This paper offers a method to select papers for SLRs and discern the main topic around the main theme. Additionally, we test the effectiveness of our technique to study cobranding and coopetition. In addition to the contents of the literature on cobranding and coopetition, our analysis supports also the comparison between them. The convergence of the two literatures is expected because coopetition represents a form of relation that encompasses cooperation and competition, and frequently cobranding strategies involve brand of rival firms. Surprisingly, we note that while cobranding topics represent motivations and consequences well, coopetition literature is quite unbalanced toward the studies on coopetition antecedents. Likely, it happens because the paradoxical nature of coopetition is self-evident. Conversely, although cobranding leads to spillover effects between the two brands, this aspect is more opaque. Assuming the paradoxical nature of coopetition, scholars wanted to explore why firms decided to compete with a rival.

Furthermore, our analysis shows that cobranding studies neglect the importance of collaboration and governance models. Indeed, cobranding studies simplicity assumes that ex-ante negotiation is an excellent solution to any problems that may emerge in the partnership [171]. Shifting our attention on the effectiveness of our method in conducting an SLR, we underscore that ML is helpful in processing extensive paper databases as input for isolating and categorizing literature patterns [146] without being explicitly managed by researchers [96]. Accordingly, the proposed ML tools for SLRs support the criterion of *transparency*. The procedure proposed is easily applicable, and other researchers may replicate the selection process of papers arriving at the same findings. Additionally, since our approach is based on a precise process described in the mathematical formula, it supports the criterion of *coherence* in selecting the articles to review. While the collaborative evaluation of abstracts or papers involving at least two authors improves the coherence of paper selection, unfortunately, it is enormously timing expensive. Automatic selection reduces the authors' bias and has terrific timing advantages.

Furthermore, the proposed methodology improves the *completeness* of the SLRs. As in the case of cobranding, to have a manageable number of articles, Pinello et al. (2022) [171] consider papers published in "high" quality journals and articles with the highest number of citations. Such protocols employ a subjective definition of thresholds and may not capture emerging research trends in "secondary" journals. Moreover, machine learning applications support the criterion of saturation in SLRs. As we show, the selection of papers proposed by Pinello et al. (2022) [171] led to a bias in stressing pertinent topics related to the research as regards the meso-context in which a cobranding strategy works. Conversely, the proposed approach does not leave out one or more relevant topics related to the theme.

Finally, the ML algorithm supports the principle of universalism by assuming an impartial perspective in selecting papers for SLR [199]. For instance, we recognize that a specific topic in coopetition has not been emphasized by Devece et al. (2019) [68]. In this case, one might suppose the authors did not adopt an impartial perspective but assumed a form of particularism based on timing. Likely, they preferred to overlook reporting one of the oldest ideas around coopetition.

## 3.6 Conclusions

This paper offers a threefold contribution. First, this paper leverages ML, based on NLP and network analysis, to conduct an SLR in business. It improves the quality of the selection of papers and the preliminary analysis of extant literature. Accordingly, we believe that the progressive application of ML for reviewing literature may reduce the general scepticism toward the validity of the literature review as a research methodology [201].

Second, from a methodological perspective, our method provides a network approach to discovering topics. Notably, our contributions overcome some frequent issues of topic modelling, such as hyperparameters fine-tuning and setting the number of topics a priori. Indeed, the implementation of topic models involves stochastic elements in their initialization phase, leading to different results that affect the composition of the topics and the rankings of the terms that describe those topics. We propose an unsupervised approach, suitable for applications involving short texts (abstracts), that does not suffer of random initialization providing reproducibility of the outputs. Moreover, topic models generally have poor performance in the homogeneous dataset: words representing a given topic may be ranked high because they are globally frequent across a corpus. Our method solves the issue since it considers the marginal occurrences of words.

Third, and finally, by examining research on coopetition and cobranding, this article juxtaposes the network analysis and recent reviews to select papers on the same literature[68, 171]. The approach provides valuable insights into the "high degree of terminological, conceptual, and explanatory heterogeneity" [74] around such studies. At the practical level, our toolkit is also helpful for researchers in selecting journals where their papers have higher chances to be published. Information that our toolkit provides about the over/under representation of the topics is a good proxy of editors interested in a specific topic.

We conclude our account by sketching further domains of ML developments where the application of our method represents only the first step. Additional, we recognize the risk of considering the ML as a substitute for scholars' interpretive critical reflection. Although the automatic selection of papers in SLR boosts transparency, completeness, saturation, universalism, and coherence [199], it cannot abolish the role of researchers' interpreta-

tion in the review process entirely. Indeed, some of the findings complementary to the ML tools are unavoidably left to the authors' understanding. Additionally, we stress that while ML supports selecting and analyzing relevant literature on a specific theme, ML cannot support look-ahead reasoning [173] and recognize new research gaps and managerial implications alone.

# Chapter 4

# Entrainment model

**Abstract**

*We introduce exact statistics that can be used to test the presence of an excess intra-group similarity against a null hypothesis in which similarity among attributes of group members occurs randomly. We present an application in natural language processing, focusing on attributes such as Part-of-Speech tags and sentiment categories to evidence the linguistic style of bankruptcy language. Moreover, the introduced statistics allows us to demonstrate the presence of a similarity excess among twins affected by neuronal disorders and children in the household, with respect to their status as NEET, student, or worker, and with respect to gender. The model depends on a single parameter and adequately describes the excess of similarity revealed in all of the empirical cases analyzed.*

## 4.1 Introduction

Studying entrainment, homophily, and social herding involves the natural and social sciences. Entrainment has been used first in physics and biology and later in other disciplines [116] to indicate a form of simultaneous behaviour. Over the last decades, the concept of entrainment has grown as a tool for investigating the transmission mechanism of attitudes and beliefs among cohorts or small groups. Indeed, the introduction within the scientific debate of the term "social entrainment", operated by McGrath et al. (1986) [148], has exacerbated the idea that similarities in human interactions, including those between two or more groups of individuals, derive from the socio-cultural background. However, although it is almost clear that entrainment, homophily and social herding increase the degree of intra-group similarity, little research has confirmed the presence of

such an excess of similarity among subjects belonging to small-size groups, like families or close friends.

## 4.2 Model: Testing excess of intra-group similarity

In this section, we introduce exact statistics that can be used to test the presence of an excess of intra-group similarity against a null hypothesis in which similarity among attributes of group members occurs randomly and depends on the number of members, $m$, in each one of the $f_m$ groups in the dataset, and the overall number, $K_m$, of individuals who display attribute $A$ in the dataset. The following subsection reports a description of the distribution associated with null hypothesis ($H_0$), while the test statistics is provided immediately afterwards, for groups of size, $m$, equal to 2 and 3. However, as it is better clarified in the remainder of this section, excess of similarity and its implications might vary depending on the specific attribute under investigation, as well as group size. Moreover, we provide the analytical expression of the probability distribution relaxing the assumption of fixed size of groups and considering more than 2 attributes.

### 4.2.1 Null hypothesis $H_0$

**Proposition 1.** *Let $S_m$ be a population consisting of $T_m = m \cdot f_m$ individuals divided into $f_m$ families with exactly $m$ family members each. Suppose that $K_m$ individuals are randomly selected from the overall $T_m = m \cdot f_m$ individuals. We denote by $\mathbf{d} = (d_0, \ldots, d_m)$ a vector of random variables $d_i$'s ($i = 0, \ldots, m$), where $d_i$ is the number of families from which* exactly *$i$ individuals have been selected*[1]*. We note that*

$$\mathbf{d} \in \left\{ \boldsymbol{\delta} = (\delta_0, \ldots, \delta_m) \in \mathbb{N}_0^{m+1} : \sum_{i=0}^{m} \delta_i = f_m, \ \sum_{i=0}^{m} i\delta_i = K_m \right\}.$$

*Then, we have*

$$P(\mathbf{d} = \boldsymbol{\delta}) = P(d_0 = \delta_0, d_1 = \delta_1, \ldots, d_m = \delta_m \,|\, K_m, f_m) = \frac{\binom{f_m}{\delta_0, \delta_1, \ldots, \delta_m} \prod_{i=0}^{m} \binom{m}{i}^{\delta_i}}{\binom{T_m}{K_m}} \quad (4.1)$$

Eq. 4.1 is obtained by dividing the (total) number of equally likely ways in which the outcome $(\delta_0, \delta_1, ..., \delta_m)$ can occur by the (total) number of arrangements of the total $T_m$ individuals in groups of size $K_m$ and $T_m - K_m$, i.e., $\binom{T_m}{K_m}$. The total number of ways in

[1] We used the notation $d_0$ to denote the number of families from which no one has been selected.

which the outcome $(\delta_0, \delta_1, ..., \delta_m)$ can occur is obtained by multiplying the number of ways in which the set of $f_m$ families can be partitioned in $m+1$ groups of size $\delta_0$,..., $\delta_m$, i.e., $\binom{f_m}{\delta_0, \delta_1, ..., \delta_m}$, by the product, over the $m+1$ distinct groups of families, of the number of ways in which $i\,(\leq m)$ individuals can be selected within each one of the $\delta_i$ families, $\prod_{i=0}^{m} \binom{m}{i}^{\delta_i}$.

**Comment**

The probability mass function (pmf) reported in Eq. 4.1 looks pretty similar to a multinomial distribution with probabilities $p_i \propto \binom{m}{i}$ (i=1,. . . ,m). However, this is not the case, since the support of the distribution not only requires that $\sum_{i=0}^{m} \delta_i = f_m$ as it is for the multinomial distribution, but also that $\sum_{i=0}^{m} i\delta_i = K_m$ which prevents Eq. 4.1 from being the pmf of a multinomial distribution, and explains the presence of the normalization constant $\binom{T_m}{K_m}$.

### 4.2.2 Illustrative example

Figure 4.1 shows an illustrative example of the distribution described in Eq. 4.1. Specifically, all of the five allowed configurations of $\boldsymbol{\delta} = (\delta_0, \delta_1, \delta_2)$ are reported for a system made of $f_2 = 12$ groups (the urns) with $m = 2$ members each (marbles), and including $K_2 = 9$ subjects with attribute $A$ (red marbles). Distribution (4.1) describes the number of groups $\delta_i$ with $i$ members who present attribute $A$, and completely disregards the role played by specific groups in the count, i.e., neither the order of groups, nor their labels are relevant. Therefore, no label is associated with urns in the figure, and shuffling the urns within a specific configuration does not affect the variables' values and the configuration's probability. The configurations reported in the figure can easily be listed by noticing that, according to the constraints, $\delta_0 + \delta_1 + \delta_2 = f_2 = 12$ and $\delta_1 + 2\delta_2 = K_2 = 9$, and therefore: $\delta_1 = 9 - 2\delta_2$ and $\delta_0 = 3 + \delta_2$, where $\delta_2$ can only take values in $\{0, 1, 2, 3, 4\}$, since both $\delta_0$ and $\delta_1$ must be non negative numbers. The first configuration from the top of the figure only includes groups with 0 or 1 member with attribute $A$. Accordingly, the probability to observe a number of groups with two members with attribute $A$ larger or equal to 0, i.e., $P(\delta_2 \geq 0)$, is equal to 1. Moving downwards in the figure, the number of groups

with both members displaying attribute $A$ increases, until the maximum value of $\delta_2 = 4$ is reached in the configuration at the bottom of the figure. In that case, $P(\delta_2 \geq 4) = 0.006$, according to the distribution (4.1), indicating that a configuration equally or more extreme than that one, in terms of the number of groups with both members displaying attribute $A$, is unlikely to occur under the null hypothesis $H_0$. Such a deviation from $H_0$ can be considered as a mark of excess of intra-group similarity as regards attribute $A$. Such a consideration is at the basis of the test statistics presented in the next subsection.

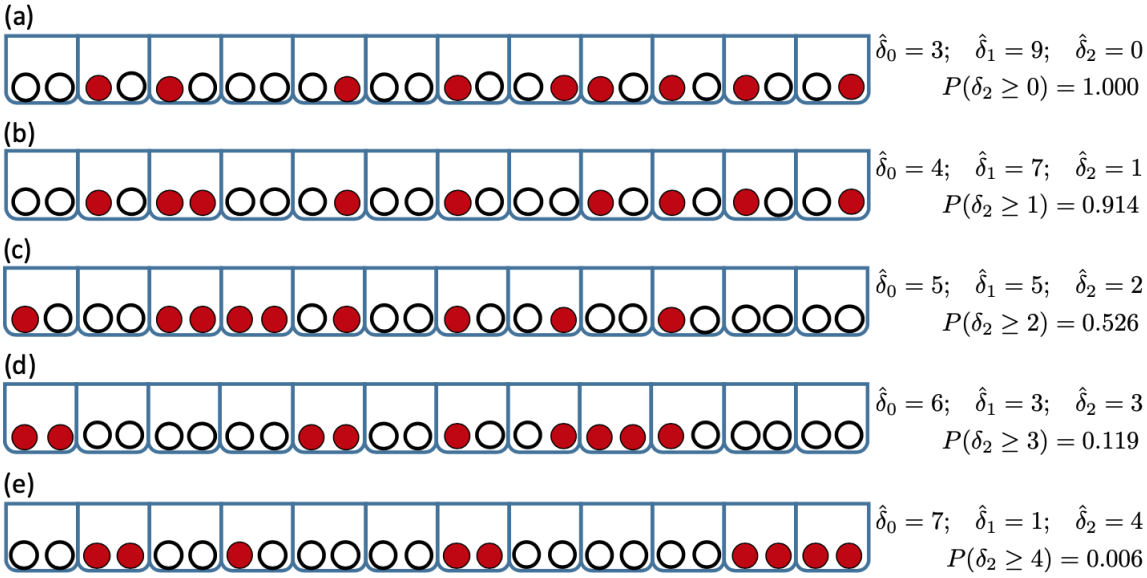



**Figure 4.1:** Illustrative example of the distribution (4.1) of a system made of $f_2 = 12$ family groups (the urns) with $m = 2$ members each (marbles), and including $K_2 = 9$ subjects with attribute $A$ (red marbles)

### 4.2.3 Test statistics

Let's consider a general system with groups of different size, $m(m = 1, \ldots, N)$ and $Q$ mutually incompatible attributes, $A_1, A_2, \ldots, A_Q$. Should one investigate intra-groups similarity with respect to attribute $A_p$, all of the other attributes are relevant just because they are other than $A_p$. Therefore they can be considered as a single attribute that we name $\bar{A}_p$. In other words, it is possible to group together all of the attributes $A_i$ with $i \neq p$ in the single attribute $\bar{A}_p$. The presence of an intra-group excess of similarity for attribute $A_P$ implies that elements with that attribute should appear more likely in the same group than it could be anticipated according to the distribution under $H_0$, which only depends on the composition of the system. Furthermore, such an excess of similarity may occur at a different extent depending on group size, $m$. Therefore, a suitable statistics should

automatically gauge to be independently applied to groups of different size, $m$, and only focus on attribute $A_p$ (equivalently $\bar{A_p}$). Accordingly, we propose the statistics

$$pair(A_p, m) = \sum_{i=0}^{m} \frac{i \cdot (i-1)}{2} d_i = \sum_{i=2}^{m} \frac{i \cdot (i-1)}{2} d_i \tag{4.2}$$

where the distribution of $d_i$, under null hypothesis $H_0$, is reported in Eq.(4.1). Statistics $pair(A_p, m)$ represents the total number of different intra-group pairs of elements with attribute $A_p$. Prominent qualities of $pair(A_p, m)$ are (i) the linearity of the statistics with respect to variables $d_i$, (ii) the fact that it automatically disregards groups with only one member ($pair(A_p, 1) = 0$), that is, groups where discussing intra-group similarity doesn't make sense, and, (iii) that the weight of $d_i$ depends quadratically on $i$, i.e., it depends quadratically on the number of subjects that share attribute $A_p$ within a group, in such a way to magnify the impact on the statistics of the presence of large homogenous (according to $A_p$) subgroups of elements. As a special case, which might help the reader to better interpret the properties of the test statistics, let's consider the case of group size $m = 2$, that is, $pair(A_p, 2) = \sum_{i=2}^{2} \frac{i \cdot (i-1)}{2} d_i = d_2$, where $\boldsymbol{d} = (d_0, d_1, d_2)$ takes values $(\delta_0, \delta_1, \delta_2)$,such that $\delta_0 + \delta 1 + \delta_2 = f_2$ and $\delta_1 + 2\delta_2 = K_2$, that is, $\delta_1 = K_2 - 2\delta_2$ and $\delta_0 = f_2 - K_2 + \delta_2$, where $K_2$ is the total number of subjects with attribute $A_p$ among the total $T_2 = 2f_2$ subjects in the set of $f_2$ families with two members each. Therefore, if $\widehat{pair(A_p, 2)} = \hat{\delta_2}$ is the observed value of the statistics, then

$$\begin{aligned} P(pair(A_p, 2) \geq \widehat{pair(A_p, 2)}) &= P(d_2 \geq \hat{\delta}_2) = \\ &= \frac{1}{\binom{T_2}{K_2}} \cdot \sum_{\delta_2 = \hat{\delta}_2}^{\lfloor \frac{K_2}{2} \rfloor} \left[ \binom{f_2}{\delta_0, \delta_1, \delta_2} \prod_{i=0}^{2} \binom{2}{i}^{\delta_i} \right] = \frac{1}{\binom{T_2}{K_2}} \cdot \sum_{\delta_2 = \hat{\delta}_2}^{\lfloor \frac{K_2}{2} \rfloor} \left[ \binom{f_2}{\delta_0, \delta_1, \delta_2} 2^{\delta_1} \right] \end{aligned} \tag{4.3}$$

After setting $f_2 = 12$ and $K_2 = 9$, the previous equation has been used to calculate the probabilities reported in Figure 4.1, where $\hat{\delta}_2 = 0$ for the configuration at the top of the figure, and $\hat{\delta}_2 = \lfloor \frac{K_2}{2} \rfloor = \lfloor \frac{9}{2} \rfloor = 4$ for the configuration at the bottom (the most extreme, according to statistics $pair(A_p, 2)$). Similarly, if we focus on groups with only three members, that is, we set $m = 3$, we have that $pair(A_p, 3) = \sum i = 23 \frac{i(i-1)}{2} d_i = d_2 + 3d_3$, and, if $\widehat{pair(A_p, 2)} = \hat{\delta}_2 + 3\hat{\delta}_3$ is the observed values of the statistics, then the associated p-value is obtained as:

$$P(d_2 + 3d_3 \geq \hat{\delta}_2 + 3\hat{\delta}_3) = \frac{1}{\binom{T_3}{K_3}} \cdot \sum_{\boldsymbol{\delta}: \delta_2 + 3\delta_3 \geq \hat{\delta}_2 + \hat{\delta}_3} \left[ \binom{f_3}{\delta_0, \delta_1, \delta_2, \delta_3} 3^{\delta_1 + \delta_2} \right] \tag{4.4}$$

where the sum is conditioned to the following constraints: $\delta_0+\delta_1+\delta_2+\delta_3 = f_3$, $\delta_1+2\delta_2+3\delta_3 = K_3$, and $K_3$ is the actual number of subjects with attribute $A_p$ among the total $T_3 = 3f_3$ subjects in the $f_3$ groups. Therefore, $\delta_1 = K_3-2\delta_2-3\delta_3$, and $\delta_0 = f_3+\delta_2+2\delta_3-K_3$, in Eq. 4.4.

## 4.3 Model extensions

### 4.3.1 Groups with different number of members and two attributes

**Theorem 1.** *Suppose that a system consists of $f$ families, $f_m$ of them with $m$ family members each ($f_1+f_2+...+f_N = f$), where $N$ is the maximum number family members present in the system.*
*For any $m = 1,\ldots,N$, let $S_m$ be a population consisting of $T_m = m\cdot f_m$ individuals divided into $f_m$ families with exactly $m$ family members each. We denote by $S = S_1\cup S_2\cup\cdots\cup S_N$ the whole population, by $T = \sum_{m=1}^{N} T_m$ the total number of individuals in $S$ and by $f = f_1+f_2+...+f_N$ the total number of families in $S$.*
*Suppose that $K$ individuals are randomly selected from $S$. We denote by $\mathbf{K} = (K_1,\ldots,K_N)$ a vector of random variables $K_m$'s ($m = 1,\ldots,N$), where $K_m$ is the number of individuals among the total $K$ that are selected from the sub-population $S_m$. It holds true that $K_1+\ldots+K_N = K$ and $K_m \le T_m$, $m = 1,\ldots,N$. Moreover, for each $m = 1,\ldots,N$, we denote by $\mathbf{d_m} = (d_{m,0},\ldots,d_{m,m})$ the vector of random variables $d_{m,i}$'s ($i = 0,\ldots,m$), where $d_{m,i}$ is the number of families, among the $f_m$'s families with $m$ family members each, from which* exactly *$i$ individuals are selected. For each $m = 1,\ldots,N$, we have that*

$$\sum_{i=0}^{m} d_{m,i} = f_m;\ \ \sum_{i=0}^{m} i\cdot d_{m,i} = K_m \tag{4.5}$$

*Moreover, since $f_1+\ldots+f_N = f$ and $K_1+\ldots+K_N = K$, it holds true that*

$$\sum_{m=1}^{N}\sum_{i=0}^{m} d_{m,i} = \sum_{m=1}^{N} f_m = f,\quad \sum_{m=1}^{N}\sum_{i=0}^{m} i d_{m,i} = K \tag{4.6}$$

*By setting $\mathbf{D} = (\mathbf{d}_1,\ldots,\mathbf{d}_N)$ and $\mathbf{\Delta} = (\boldsymbol{\delta}_1,\ldots,\boldsymbol{\delta}_N) \in \mathbf{N}_0^{N+1}$, where $\boldsymbol{\delta}_m = (\delta_{m,0},\ldots,\delta_{m,m})$ is a possible value vector that can be assumed by $\mathbf{D}$, it holds true that*

$$P(\mathbf{D} = \mathbf{\Delta}|K) = \frac{1}{\binom{T}{K}}\prod_{m=1}^{N}\left[\binom{f_m}{\delta_{m,0},\ \ldots,\ \delta_{m,m}}\prod_{i=0}^{m}\binom{m}{i}^{\delta_{m,i}}\right]. \tag{4.7}$$

*Proof.* We decompose probability $P(\mathbf{D} = \boldsymbol{\Delta} \mid K)$ as

$$P(\mathbf{D} = \boldsymbol{\Delta}|K) = \sum_{\mathbf{k}:\, k_1+\ldots+k_N=K} P[\mathbf{D} = \boldsymbol{\Delta}|\mathbf{K} = \mathbf{k}, K] \cdot P(\mathbf{K} = \mathbf{k}|K)$$

where

$$\begin{aligned} k_1 &= \delta_{1,1} \\ k_2 &= \delta_{2,1} + 2\,\delta_{2,2} \\ k_3 &= \delta_{3,1} + 2\,\delta_{3,2} + 3\,\delta_{3,3} \\ &\ldots \\ k_N &= \sum_{i=0}^{N} i\,\delta_{N,i} \end{aligned}$$

The system above admits unique solution $\tilde{\mathbf{k}} = (\tilde{k}_1, \ldots, \tilde{k}_N)$ given $\delta$'s, then $P(\mathbf{K} = \tilde{\mathbf{k}}|\mathbf{D} = \boldsymbol{\Delta}, K) = 1$. Then by construction the complementary event $P(\mathbf{K} \neq \tilde{\mathbf{k}}|\mathbf{D} = \boldsymbol{\Delta}, K) = 0$. Let $E = \{D = \boldsymbol{\Delta}|K\}$, $A = \{\mathbf{K} = \tilde{\mathbf{k}}|K\}$ and $A^c = \{\mathbf{K} \neq \tilde{\mathbf{k}}|K\}$ then

$$P(A^c|E) = \frac{P(E|A^c) \cdot P(A^c)}{P(E)} = 0$$

. Since $P(E) \neq 0$ and $P(A^c) \neq 0$, $P(E|A^c) = 0$. By the law of total probability

$$P(E) = P(E|A)\,P(A) + \underbrace{P(E|A^c)}_{=0}\,P(A^c)$$

Then,

$$\begin{aligned} P(\mathbf{D} = \boldsymbol{\Delta}|K) = &\sum_{\substack{\mathbf{k}:\, k_1+\ldots+k_N=K \\ \mathbf{k}\neq\tilde{\mathbf{k}}}} P[\mathbf{D} = \boldsymbol{\Delta}|\mathbf{K} = \mathbf{k}, K] \cdot P(\mathbf{K} = \mathbf{k}|K) \\ &+ P[\mathbf{D} = \boldsymbol{\Delta}|\mathbf{K} = \tilde{\mathbf{k}}, K] \cdot P(\mathbf{K} = \tilde{\mathbf{k}}|K) \end{aligned} \tag{4.8}$$

where the summation (the first part of the sum) is equal to zero.
For any possible value $\mathbf{k}$ of $\mathbf{K}$, the probability $P(\mathbf{K} = \mathbf{k}|K)$ is the probability that, randomly selecting $K = K_1 + \ldots + K_N$ individuals from the overall $T = \sum_{m=1}^{N} m\,f_m$ individuals in the system $S$, $K_1 = k_1$ individuals are selected among the $T_1$'s individuals of the sub-population $S_1$, $K_2 = k_2$ individuals are selected among the $T_2$'s individuals of the sub-population $S_2$, and so on. Therefore, $P(\mathbf{K} = \mathbf{k}|K)$ is just given by the pmf

of a multivariate hypergeometric distribution. Specifically, for any possible value $\mathbf{k} = (k_1, ..., k_m)$ of $\mathbf{K}$ we have that

$$P(\mathbf{K} = \mathbf{k}|K) = \frac{\prod_{m=1}^{N} \binom{T_m}{k_m}}{\binom{T}{K}}.$$

Then,

$$\begin{aligned} P(\mathbf{D} = \mathbf{\Delta} \,|\, K) &= P[\mathbf{D} = \mathbf{\Delta}|\mathbf{K} = \tilde{\mathbf{k}}, K] \, \frac{\prod_{m=1}^{N} \binom{T_m}{\tilde{k}_m}}{\binom{T}{K}} = \\ &= P[\mathbf{D} = \mathbf{\Delta}|\mathbf{K} = (\tilde{k}_1, ..., \tilde{k}_N), K] \, \frac{\prod_{m=1}^{N} \binom{T_m}{\tilde{k}_m}}{\binom{T}{K}}, \end{aligned} \tag{4.9}$$

On the other hand, as soon as $\tilde{k}_1, \tilde{k}_2, ..., \tilde{k}_N$ are given, then probability $P(\mathbf{D} = \mathbf{\Delta} \,|\, \tilde{k}_1, ..., \tilde{k}_N)$ can be factorized with respect to the groups of families with the same number of family members, since events within each group are mutually independent. Therefore, we have that

$$P(\mathbf{D} = \mathbf{\Delta} \,|\, K) = \left[ \prod_{m=1}^{N} P(d_{m,1} = \delta_{m,1}, d_{m,2} = \delta_{m,2}, ..., d_{m,m} = \delta_{m,m} \,|\, \tilde{k}_m) \right] \frac{\prod_{m=1}^{N} \binom{T_m}{\tilde{k}_m}}{\binom{T}{K}}. \tag{4.10}$$

Probability $P(d_{m,1} = \delta_{m,1}, d_{m,2} = \delta_{m,2}, ..., d_{m,m} = \delta_{m,m} \,|\, k_m)$ is provided by Eq. (**??**). So,

$$P(\mathbf{D} = \mathbf{\Delta} \,|\, K) = \left[ \prod_{m=1}^{N} \frac{\binom{f_m}{\delta_{m,0},\, \delta_{m,1},\, \delta_{m,2},\, ...,\, \delta_{m,m}}}{\binom{T_m}{\tilde{k}_m}} \prod_{i=0}^{m} \binom{m}{i}^{\delta_{m,i}} \right] \frac{\prod_{m=1}^{N} \binom{T_m}{\tilde{k}_m}}{\binom{T}{K}} = \tag{4.11}$$

$$= \frac{1}{\binom{T}{K}} \prod_{m=1}^{N} \left[ \binom{f_m}{\delta_{m,0},\, \ldots,\, \delta_{m,m}} \prod_{i=0}^{m} \binom{m}{i}^{\delta_{m,i}} \right] \tag{4.12}$$

which is equal to Eq. (4.2). □

### 4.3.2 Groups with different number of members and three attributes

There are systems in which one may need to consider the presence of more than two attributes, assuming that attribute $A$, $B$, and $C$ cannot occur simultaneously in the same person, but they can occur within the same family. The null hypothesis introduced in the previous section can be easily generalized to deal with three attributes. First of all, a straightforward generalization of the notation should be introduced as follows. The variable $d_{m,i,j}$ describes the number of families with $m$ members each, where $i$ members present attribute $A$, $j$ members attribute $B$, and the remainder $m - i - j$ attribute $C$.

Therefore, by construction, variables $d_{m,i,j}$ are not independent:

$$\sum_{i=0}^{m}\sum_{j=0}^{m-i} d_{m,i,j} = f_m \quad \forall m \in 1,\ldots,N; \quad \sum_{m=1}^{N}\sum_{i=0}^{m}\sum_{j=0}^{m-i} i\, d_{m,i,j} = K_A; \quad \sum_{m=1}^{N}\sum_{i=0}^{m}\sum_{j=0}^{m-i} j\, d_{m,i,j} = K_B;$$

$$P(\mathbf{D}=\boldsymbol{\Delta}) = \frac{1}{\binom{T}{K_A, K_B, T-K_A-K_B}} \prod_{m=1}^{N}\left[\binom{f_m}{\vec{\delta}_m}\prod_{i=0}^{m}\prod_{j=0}^{m-i}\binom{m}{i,j,m-i-j}^{\delta_{m,i,j}}\right]. \quad (4.13)$$

where

$$\vec{\delta}_m = (\delta_{m,0,0},\ldots,\delta_{m,m,0},\ldots,\delta_{m,0,1},\ldots,\delta_{m,0,m})$$

### 4.3.3 Groups with different number of members and Q different attributes

In the case with $Q$ different attributes, the variable $d_{m,\vec{q}}$ ,with $\vec{q} = (q_1,\ldots,q_Q)$, describes the number of families with $m$ members each, where $q_i$ indicates the number of members with the $q_i$ attribute. Therefore, the following conditions hold:

$$\sum_{j=1}^{Q}\sum_{q_j=1}^{m-q_{j-1}} d_{m,\vec{q}} = f_m \quad , \quad \sum_{m=1}^{N} f_m = f$$

$$\sum_{j=1}^{Q}\sum_{q_j=0}^{m-q_{j-1}} q_j\, d_{m,\vec{q}} = k_m \quad , \quad \sum_{m=1}^{N} k_m = K$$

Let $q_{j^*}$ be a specific attribute, with $q_{j^*} \in \{q_1,\ldots,q_Q\}$, then the following holds:

$$\sum_{m=1}^{N}\sum_{q_{j^*}=0}^{m} q_{j^*}\, d_{m,\vec{q}} = K_{q_{j^*}} \qquad \sum_{q_j^*}^{m} q_{j^*}\, d_{m,\vec{q}} = K_{m,q_{j^*}}$$

where $K_{q_{j^*}}$ is the the number of individuals with attribute $q_{j^*}$ and $K_{m,q_{j^*}}$ is the number of individuals with attribute $q_{j^*}$ within the families with $m$ members Then

$$P(\mathbf{D}=\boldsymbol{\Delta}|K) = \frac{1}{\binom{T}{K_1,\ldots,K_Q}} \prod_{m=1}^{N}\left[\binom{f_m}{\delta_{m,\vec{q}}}\prod_{j=1}^{Q-1}\prod_{q_j=0}^{m-q_{j-1}}\binom{m}{q_1,\ldots,q_Q}^{\delta_{m,\vec{q}}}\right] \quad (4.14)$$

where $\delta_{m,\vec{q}} = \delta_{m,q_1,\ldots,q_Q}$, constrained to $\sum_{j=}^{Q} q_j = m$.

## 4.4 Applications

### 4.4.1 An application to word attributes on sentences

Across companies and conditions, the linguistic features and grammar styles are not necessarily homogenous. Differences in how individuals make grammatical choices to connect heterogeneous types of content tend to be a sub-conscious activity, which may effectively signal how they relate to their social and environmental policies [58]. Following Chapter 1, we explore the concentration of linguistic attributes in windows of length 2 (2-grams), as the order of the Language Model used and the window's length for the context of words. As in studying the language of bankruptcy, we focus the analysis on the importance of the different contexts of words splitting the corpus into two sub-datasets, bankruptcy and healthy. Here, we retrieve the POS-tags and sentiment of words during the pre-processing step. So, we collect sentences of documents forgetting the meaning of words and focusing only on their POS-tags and sentiment tags. We split all sentences in a window of length 2 ( 2-grams), replacing words with Pos-tags and sentiments. We get the POS-tags by the parsing tool of Spacy, implemented in Python and the sentiment tags from the popular sentiment word list of Loughran and McDonald [135]. From the sentiment word list, we consider only *positive* and *negative* sentiments, considering the tags *uncertainty*, *litigious*, *constraining* as *negative*. Moreover we add the tag *neutral* to assign at least a tag to all words. From the POS-tags we consider *nouns*, *verbs*, *adjectives*, and *adverbs*. Finally, we study 3 different attributes: POS-tags, sentiment tags and pair of POS-tag and sentiment tag (as unique attribute). Tables 4.1, 4.2, and 4.3 report the results. The first column indicates the attribute used for the test. In the second, we have the number of documents for which we accepted to reject the null hypothesis. *Tail* and *Correction* columns refer to the left or right tail, meaning anti-entrainment or entrainment, and which correction was used (at significant level $\alpha = 0.01$). In Table 4.1, we observe that only some healthy documents have an effect of entrainment for the sentiment tags, also if we use a less conservative correction (FDR). It shows an informative disclosure on reports of healthy companies. In Table 4.3, there is more anti-entrainment, in terms of *nouns* and *verbs*, and more entrainment for the *adverbs* in the bankruptcy corpora than in the healthy one. These results suggest how the two corpora styles show the opposite effect of *adverbs* with respect to *nouns* and *verbs*. However, the difference becomes less evident with the FDR correction. Finally, combining the POS-tags and sentiment tags

we observe more entrainment on bankruptcy documents relative to *nouns-negative*. Although, from Tables 4.2 and 4.1 we have documents that have anti-entrainment in terms of POS-Tags and no ones of entrainment for the negative sentiment. In contrast with the healthy language style, the bankruptcy language shows strong associations of negative sentiment words only referring to *nouns*, so it is informative only in some parts of the report describing adverse facts. Moreover, the entrainment of *adverbs-neutral* and *adverbs* suggests that the bankruptcy language is characterized by a more convoluted and fragmented speech, with many subordinates in sentences. The literature on narrative accounting confirms some of these results. Some studies show that when the language is mainly associated with reporting stories (i.e., more narrative), it tends to include a more widespread use of adverbs, conjunctions, impersonal pronouns, negations, and personal pronouns [49]. Moreover, after a negative event (e.g., an irresponsible act), we expect that a company could develop a more narrative language style. Reports reveal that the more a firm is involved in irresponsible business conduct, the more likely it is to use narrative (instead of analytical) language.

**Table 4.1:** Summary statistics of excess of similarity with respect to sentiment attributes

| Sentiment | Healthy documents | Bankruptcy documents | Tail | Correction |
|---|---|---|---|---|
| Negative | 12 | - | right | Bonferroni |
| Neutral | 25 | - | right | Bonferroni |
| Negative | 21 | - | right | FDR |
| Neutral | 45 | - | right | FDR |
| Positive | 1 | - | right | FDR |

**Table 4.2:** Summary statistics of excess of similarity with respect to POS-tag attributes

| POS-tag | Healthy documents | Bankruptcy documents | Tail | Correction |
|---|---|---|---|---|
| Adjectives | 2 | - | right | Bonferroni |
| Adverbs | 15 | 20 | right | Bonferroni |
| Adjectives | 4 | 3 | right | FDR |
| Adverbs | 30 | 33 | right | FDR |
| Adjectives | 4 | 8 | left | Bonferroni |
| Nouns | 25 | 34 | left | Bonferroni |
| Verbs | 13 | 27 | left | Bonferroni |
| Adjectives | 12 | 16 | left | FDR |
| Nouns | 45 | 58 | left | FDR |
| Verbs | 30 | 39 | left | FDR |

**Table 4.3:** Summary statistics of excess of similarity with respect to POS-tag & sentiment attributes

| POS-tag - Sentiment | Healthy documents | Bankruptcy documents | Tail | Correction |
|---|---|---|---|---|
| Adverbs - Neutral | 9 | 16 | right | Bonferroni |
| Nouns - Negative | 5 | 9 | right | Bonferroni |
| Verbs - Negative | 3 | 2 | right | Bonferroni |
| Adverbs - Neutral | 27 | 33 | right | FDR |
| Nouns - Negative | 12 | 17 | right | FDR |
| Verbs - Negative | 3 | 3 | right | FDR |
| Adjectives - Neutral | 4 | 4 | right | FDR |
| Adjectives - Negative | - | 3 | right | FDR |
| Nouns - Neutral | - | 2 | right | FDR |
| Adjectives - Neutral | - | 11 | left | Bonferroni |
| Nouns - Neutral | 11 | 16 | left | Bonferroni |
| Verbs - Neutral | 29 | 51 | left | Bonferroni |
| Adjectives - Neutral | 8 | 11 | left | FDR |
| Verbs - Neutral | 51 | 69 | left | FDR |
| Nouns - Neutral | 24 | 33 | left | FDR |

### 4.4.2 An application to real data of patients affected by neuronal disorders.

In a recent empirical study of twins affected by neurodevelopmental disorder (PMA-REF), some authors faced the problem of testing the disorder's familiarity to limit its impact on the investigation of other risk factors, such as the conception method. The dataset includes information about 41 pregnancies—37 twin-pregnancies and 4 three-baby pregnancies. All of the latter pregnancies resulted from assisted reproductive technology (ART)—two from FIVET and two from ICSI technology. One of the three-baby pregnancies from ICSI technology led to the death of one offspring. One death also occurred in a twin-pregnancy stemming from FIVET technology. Such a feature allows one to untangle the effects of the association between ART and twin births and the association between twin births and neurological disorders. However, the database population might bring, as a side effect, to observe that the event that a child is affected by a neurological disorder is not independent of the event that her twin sibling is also affected. In other words, that database might suffer from a bias due to the familiarity with neurodevelopmental disorders. To check for the presence and statistical significance of that bias, we tested whether children affected by neurological disorders tended to group in an overall number of pregnancies, that is, families, which is smaller than what we could anticipate by assuming no familiarity effect. By excluding the dead from the analysis and using the notation introduced in the previous sections, the database consists of $f = 41$ families that can be divided into three homogeneous groups, according to the number of offspring. Specifically, it is $f_1 = 1$, $f_2 = 37$, and $f_3 = 3$. The total number of children with a neurodevelopmental disorder is $K = 37$ over a total of 84. The observed value of the statistics $q = \sum_{m=1}^{N} d_{m,0}$ is $\hat{q} = 16$. Before calculating the p-value associated with $\hat{q}$, we checked if the null hypothesis well describes the actual experiment, when the familiarity effect, if any, is removed. We consider a matrix, $\mathbf{M}$, with two columns and a number of rows equal to the total number of children in the dataset (after removing the dead), the first column reporting progressive numbers that identify the mother of children—so that twins are identified by the same number in the matrix—and the second column reporting whether a child shows a neurological disorder or not. Then we perform a random shuffling of the second column of matrix $\mathbf{M}$, in such a way to destroy any familiarity effect, and calculate the value of statistics $q$, $\hat{q}^*$ as the number of mothers with no offspring affected by a neurodevelopmental disorder in the shuffled matrix. We construct a total of $10^7$ (independently) shuffled

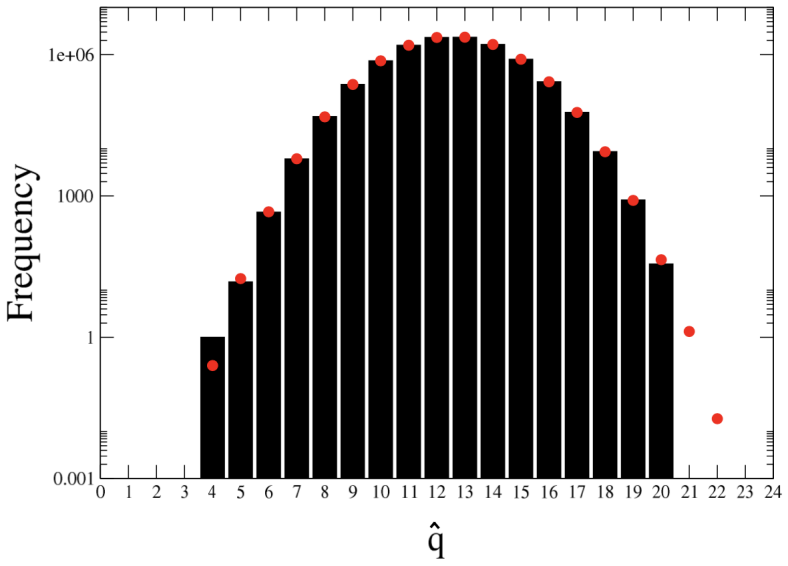



**Figure 4.2:** Comparison between the expected frequency of outcomes of $q$ (red dots), according to probability mass function of the presented null model, and the frequency of outcomes in the shuffling experiment (black bars), as calculated over $10^7$ independent realizations.

replicas of matrix $\mathbf{M}$, and calculate the frequency of each possible outcome of $q$. Figure 4.2 shows a comparison between the expected frequency of outcomes of $q$, according to probability mass function of the presented null model, and the frequency of outcomes in the shuffling experiment. The perfect agreement observed in the figure is also supported by the result of a Kolmogorov-Smirnov test for discrete distribution, which gives a p-value of 0.866. The analysis performed so far supports the effectiveness of our model as an appropriate null hypothesis to test for the presence of a familiarity effect in the considered data set. We obtain that $P(q \geq \hat{q} = 16) = 0.0332752$, which indicates that the familiarity effect cannot be excluded at a 5% confidence level.

### 4.4.3 An application of the test to children status similarity

Over the last few years, the percentage of young people who are not in Education, Employment, or Training, the so-called NEET [185], has increased dramatically in many European countries, especially in the Mediterranean ones. We investigate intra-group similarity by using data collected by the Italian Institute of Statistics (ISTAT) through the Survey on Household Consumption, in the years from 2001 to 2013. The data consists of a stratified sample of households that is representative of the Italian population. The analysis of intra-group similarity has been done on the subset of households with the

following family composition: two parents, with either one or two parents working, and two children (m=2), or three children (m=3), all of age between 21 and 30. The analysis of similarity concerns children, and, in particular, their status as student (attribute A1), worker (attribute A2), and NEET (attribute A3). Moreover, unless theoretical hypotheses about the entanglement of two or more attributes shall be tested, we believe that the best way to analyze intra-group similarity in a dataset is to split the data in subsets, which are homogeneous by group size, and focus on a single attribute at a time. We consider statistics $pair(A_{p2})$, i.e., the total number of pairs of children who share attribute $A_p$ and belong to the same family. Table 4.4 reports the summary statistics of the excess of similarity among children, according to the statistics $x_1 = pair(A_1, 2) = d_2$ with respect to attribute $A_1$=student, statistics $x_2 = pair(A_2, 2)$ with respect to attribute $A_2$=worker, and statistics $x_3 = pair(A_3, 2)$ with respect to attribute $A_3$=NEET for families with 2 children. Looking at Table 4.4, small p-values under $H_0$ indicate the significance level of the statistics for each year. Notably, the $Z$ score $(Z|H_0)$ values range from 5.5 (lower case) to 9.9 (upper case), suggesting similar differences for the attribute $A_p$. Our results suggest that Italian households' children entrain with the social habitus of their parents (in terms of educational level and working status), making some of their attributes (their status of students, workers or NEETs) predictable. We conceptualize this phenomenon as "family entrainment". Recent findings [132] indicate that being NEET among Italian children is the consequence of a sort of imitation effect in replicating the family background model. However, the mechanism underlying this process is far from being completely clear. Indeed, the NEET condition seems to be a combination of both economic and social deprivations and the result of habitus transmitted to the family of origin. Table 4.5 reports the summary statistics of the excess of similarity among children, according to the statistics $x_1 = pair(A_1, 3) = d_2 + 3d_3$ with respect to attribute $A_1$=student, statistics $x_2 = pair(A_2, 3)$ with respect to attribute $A_2$=worker, and statistics $x_3 = pair(A_3, 3)$ with respect to attribute $A_3$=NEET for families with 3 children. As the number of children increases, the entrainment effect (i.e., excess of similarity) slightly reduces, but it appears still evident. These results suggest that in households with three children, the children imitative behavior becomes more complex to replicate, and they tend to be more responsible for their status. Table 4.6 shows the average percentage of NEETs, workers, and students among children, as a function of the maximum level of education of parents (Low, Average, and High), and number of working parents (NWP), in families with

two children. Percentages are calculated over families homogeneous by NWP and level of education of the parents. The results show a higher presence of NEETs children, while the percentage of female NEETs declines rapidly as parental education increases, but not for male NEETs. Indeed, we note that the distribution of NEETs is higher in households composed of one working parent (often the father) than in households with two working parents. This finding is also in line with those of Lo Verde et al. (2022) [132] who have found as the family composition and in particular, the mother's working status influence the presence of NEEETs in Italian families.

**Table 4.4:** Summary statistics of the excess of similarity w.r.t. student, worker and NEET attributes for families with 2 children

| Year | Attribute 1: student | | | | | | Attribute 2: worker | | | | | Attribute 3: NEET | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $f_2$ | $\hat{x}_1$ | $E[x_1\|H_0]$ | $\sigma[x_1\|H_0]$ | $Z\|H_0$ | P-v.$\|H_0$ | $\hat{x}_2$ | $E[x_2\|H_0]$ | $\sigma[x_2\|H_0]$ | $Z\|H_0$ | P-v.$\|H_0$ | $\hat{x}_3$ | $E[x_3\|H_0]$ | $\sigma[x_3\|H_0]$ | $Z\|H_0$ | P-v.$\|H_0$ |
| 2001 | 542 | 114 | 65.4 | 5.3 | 9.2 | $6.6e-20$ | 142 | 92.9 | 5.7 | 8.7 | $1.2e-18$ | 70 | 30.6 | 4.2 | 9.3 | $9.1e-19$ |
| 2002 | 610 | 111 | 68.8 | 5.3 | 9.9 | $3.0e-22$ | 188 | 133.0 | 6.2 | 8.9 | $1.7e-19$ | 70 | 30.0 | 4.3 | 9.4 | $1.3e-18$ |
| 2003 | 581 | 96 | 58.8 | 5.2 | 7.1 | $2.3e-12$ | 168 | 130.0 | 6.0 | 6.3 | $1.8e-10$ | 49 | 25.1 | 4.0 | 6.0 | $1.2e-08$ |
| 2004 | 540 | 86 | 55.0 | 5.1 | 6.1 | $1.3e-09$ | 151 | 109.2 | 5.8 | 7.3 | $2.5e-23$ | 62 | 28.6 | 4.1 | 8.1 | $1.7e-14$ |
| 2005 | 490 | 112 | 72.0 | 5.2 | 7.6 | $2.1e-14$ | 124 | 79.5 | 5.3 | 8.4 | $4.7e-17$ | 45 | 22.2 | 3.7 | 6.1 | $6.3e-09$ |
| 2006 | 430 | 91 | 56.5 | 4.8 | 7.2 | $7.3e-13$ | 114 | 82.9 | 5.1 | 6.1 | $9.1e-10$ | 40 | 16.7 | 3.3 | 7.1 | $5.5e-11$ |
| 2007 | 445 | 105 | 66.7 | 5.0 | 7.6 | $2.0e-14$ | 109 | 71.9 | 5.1 | 7.3 | $2.1e-13$ | 50 | 19.6 | 3.5 | 8.7 | $6.9e-16$ |
| 2008 | 408 | 86 | 53.9 | 4.7 | 6.8 | $8.3e-12$ | 106 | 78.0 | 5.1 | 5.6 | $1.3e-08$ | 39 | 16.0 | 3.2 | 7.2 | $3.6e-11$ |
| 2009 | 400 | 92 | 58.4 | 4.7 | 7.1 | $1.2e-12$ | 110 | 62.2 | 5.0 | 8.4 | $2.6e-17$ | 42 | 16.1 | 3.2 | 8.1 | $1.1e-13$ |
| 2010 | 499 | 84 | 54.0 | 4.6 | 6.4 | $1.3e-10$ | 82 | 46.2 | 4.5 | 8.0 | $2.9e-15$ | 70 | 33.6 | 4.1 | 8.8 | $8.0e-18$ |
| 2011 | 379 | 82 | 55.4 | 4.6 | 5.8 | $6.7e-09$ | 80 | 51.2 | 4.5 | 6.4 | $2.3e-10$ | 46 | 23.5 | 3.6 | 6.2 | $2.9e-9$ |
| 2012 | 397 | 93 | 56.9 | 4.7 | 7.7 | $1.5e-14$ | 86 | 51.8 | 4.6 | 7.4 | $1.2e-13$ | 52 | 26.6 | 3.8 | 6.6 | $1.7e-10$ |
| 2013 | 303 | 81 | 51.9 | 4.2 | 6.9 | $4.0e-12$ | 66 | 35.6 | 3.9 | 7.7 | $1.5e-14$ | 34 | 17.7 | 3.2 | 5.1 | $1.1e-06$ |

**Table 4.5:** Summary statistics of the excess of similarity w.r.t. student, worker and NEET attributes for families with 3 children

| Year | Attribute 1: student | | | | | | Attribute 2: worker | | | | | Attribute 3: NEET | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $f_3$ | $\hat{x}_1$ | $E[x_1\|H_0]$ | $\sigma[x_1\|H_0]$ | $Z\|H_0$ | P-v.$\|H_0$ | $\hat{x}_2$ | $E[x_2\|H_0]$ | $\sigma[x_2\|H_0]$ | $Z\|H_0$ | P-v.$\|H_0$ | $\hat{x}_3$ | $E[x_3\|H_0]$ | $\sigma[x_3\|H_0]$ | $Z\|H_0$ | P-v.$\|H_0$ |
| 2001 | 42 | 34 | 19.6 | 2.7 | 5.4 | $1.3e-06$ | 31 | 18.0 | 2.6 | 4.9 | $8.9e-06$ | 13 | 6.0 | 1.9 | 3.6 | $1.5e-03$ |
| 2002 | 70 | 26 | 9.9 | 2.5 | 6.5 | $4.7e-08$ | 56 | 37.5 | 3.5 | 5.2 | $1.2e-06$ | 47 | 26.6 | 3.3 | 6.2 | $2.8e-08$ |
| 2003 | 51 | 24 | 15.5 | 2.7 | 3.2 | $2.7e-03$ | 47 | 36.5 | 3.1 | 3.4 | $1.1e-03$ | 16 | 5.3 | 1.9 | 5.7 | $4.4e-06$ |
| 2004 | 52 | 41 | 26.8 | 3.0 | 4.7 | $1.6e-05$ | 46 | 30.3 | 3.1 | 5.1 | $2.6e-06$ | 12 | 3.0 | 1.5 | 6.0 | $5.3e-06$ |
| 2005 | 38 | 31 | 19.1 | 2.6 | 4.6 | $2.8e-05$ | 20 | 10.5 | 2.3 | 4.2 | $1.8e-04$ | 18 | 8.8 | 2.1 | 4.3 | $1.5e-04$ |
| 2006 | 36 | 14 | 5.6 | 1.8 | 4.6 | $1.1e-04$ | 43 | 35.3 | 2.5 | 3.0 | $4.0e-03$ | 9 | 3.9 | 1.6 | 3.2 | $6.0e-03$ |
| 2007 | 40 | 34 | 21.4 | 2.7 | 4.7 | $1.7e-05$ | 20 | 13.1 | 2.4 | 2.8 | $6.8e-03$ | 14 | 6.8 | 2.0 | 3.6 | $1.4e-03$ |
| 2008 | 41 | 28 | 14.8 | 2.5 | 5.2 | $2.9e-06$ | 32 | 17.0 | 2.6 | 5.8 | $2.3e-07$ | 21 | 9.2 | 2.2 | 5.3 | $4.0e-06$ |
| 2009 | 38 | 20 | 11.8 | 2.3 | 3.5 | $1.2e-03$ | 20 | 11.8 | 2.3 | 3.5 | $1.2e-03$ | 21 | 13.8 | 2.4 | 3.0 | $4.9e-03$ |
| 2010 | 42 | 21 | 14.8 | 2.3 | 2.6 | $1.1e-02$ | 19 | 13.3 | 2.3 | 2.5 | $1.5e-02$ | 8 | 4.9 | 1.7 | 1.8 | $6.9e-02$ |
| 2011 | 41 | 29 | 17.7 | 2.6 | 4.3 | $7.7e-05$ | 22 | 10.9 | 2.3 | 4.8 | $2.5e-05$ | 20 | 12.1 | 2.4 | 3.3 | $2.3e-03$ |
| 2012 | 34 | 20 | 13.2 | 2.3 | 2.9 | $5.6e-03$ | 17 | 11.1 | 2.2 | 2.6 | $1.2e-02$ | 15 | 9.2 | 2.1 | 2.7 | $1.0e-02$ |
| 2013 | 23 | 15 | 8.8 | 1.9 | 3.2 | $3.2e-03$ | 21 | 11.1 | 2.0 | 4.9 | $1.2e-05$ | 36 | 3.5 | 1.5 | 1.7 | $9.5e-02$ |

**Table 4.6:** Average number of attributes as a function of parents' level of education in families with 2 children

| Year | Any | | | | | | Female | | | | | | Male | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Status | Neets | | Workers | | Students | | Neets | | Workers | | Students | | Neets | | Workers | | Students | |
| NWP | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| Low | 33% | 18% | 44% | 56% | 23% | 26% | 36% | 18% | 36% | 49% | 28% | 33% | 30% | 17% | 51% | 64% | 19% | 19% |
| Average | 26% | 19% | 33% | 38% | 41% | 43% | 26% | 20% | 28% | 33% | 46% | 47% | 25% | 19% | 39% | 42% | 36% | 39% |
| High | 18% | 15% | 24% | 23% | 58% | 62% | 16% | 14% | 23% | 19% | 61% | 67% | 19% | 15% | 25% | 27% | 56% | 58% |

### 4.4.4 An application of the test to children gender similarity

This application relies on the concentration of children's gender in households with two or three children. We consider statistics $pair(Ap2)$, i.e., the total number of pairs of children with the same gender (female or male) belonging to the same family. The results are shown in the following Tables. Table 4.7 reports the summary of the statistical test considering the families with only two children and the first two children of families with three children. In Table 4.7 we do not observe entrainment. According to the results, we assume that the concentration of gender (male or female) follows the distribution under the Null Hypothesis. The observed statistics are very close to the expected values under the null hypothesis $H_0$, and we cannot reject $H_0$ at any significant level $\alpha$, as shown by the right tail p-values. Table 4.8 summaries the results concerning only the families with three children, according to the statistic $pair(A, 3) = d_2 + 3d_3$. In this case, we observe a slight entrainment, as we can see from the column of right tail p-values. Then, comparing this result with the ones in the Table 4.7, the families with three children seem to have different behaviours in expressing gender similarity of their children. So, we analyse the two datasets separately.

Table 4.9 presents the summary of the test statistic considering only the families with two children. From these results, we observe an anti-entrainment phenomenon. It means that most families with only two children have no gender concentration, so one male and one female. Also, this result could be expected and reasonable since the parents have no control over the genders of their children.

Table 4.10 shows the test statistic results considering only the first two children of families with three children. Here, we observe the unexpected result: we reject the null hypothesis. Surprisingly, we observe gender entrainment in the first two children, which does not happen in the case of families with only two children. From a sociological perspective, this result, combined with the previous ones, could be explained by the Monty-Hall paradox. According to Table 4.7, we do not observe entrainment if we consider the families with only two children and the first two children of families with three children due to the mitigation effect of the anti-entrainment effect of families with only two children. However, only families with three children show entrainment if we consider the first two children. Since the parents have no control over the gender of their children, it could mean that

when a family has the first two children of the same gender, they are motivated to have a third child. This motivation could be explained in terms of willingness to have a more equal gender proportion in their children. So, for the Monty-Hall paradox, the choice to have or not the third child relies on the observed genders of the first two children. Finally, we can say that parents are not indifferent to the past (genders of their first two children) in choosing to have a third child.

**Table 4.7:** Summary statistics of the excess of similarity w.r.t gender: families with 2 children and first two children of families with 3 children

| Years | $f_2$ | $\hat{x}_1$ | $E[x_1\|H_0]$ | $\sigma[x_1\|H_0]$ | $Z\|H_0$ | Right P-v.$\|H_0$ |
|---|---|---|---|---|---|---|
| $2002-2007$ | 19510 | 4278 | 4312.737 | 34.7957 | $-0.9983$ | 0.8444 |
| $2008-2013$ | 14145 | 3094 | 3084.5436 | 29.6041 | 0.3194 | 0.3811 |

**Table 4.8:** Summary statistics of the excess of similarity w.r.t gender: families with 3 children

| Years | $f_3$ | $\hat{x}_1$ | $E[x_1\|H_0]$ | $\sigma[x_1\|H_0]$ | $Z\|H_0$ | Right P-v.$\|H_0$ |
|---|---|---|---|---|---|---|
| $2002-2007$ | 2569 | 1795 | 1721.8 | 21.9 | 3.3 | 0.00050 |
| $2008-2013$ | 1723 | 1215 | 1156.3 | 17.9 | 3.3 | 0.00067 |

**Table 4.9:** Summary statistics of the excess of similarity w.r.t gender: families with 2 children

| Years | $f_2$ | $\hat{x}_1$ | $E[x_1\|H_0]$ | $\sigma[x_1\|H_0]$ | $Z\|H_0$ | Left P-v.$\|H_0$ | Right P-v.$\|H_0$ |
|---|---|---|---|---|---|---|---|
| $2002-2007$ | 16941 | 3626 | 3732.0254 | 32.4177 | $-3.2706$ | 0.00057 | 0.99949 |
| $2008-2013$ | 12422 | 2651 | 2693.5191 | 27.7327 | $-1.5332$ | 0.06485 | 0.93959 |

**Table 4.10:** Summary statistics of the excess of similarity w.r.t gender: first two children of families with 3 children

| Years | $f_2$ | $\hat{x}_1$ | $E[x_1\|H_0]$ | $\sigma[x_1\|H_0]$ | $Z\|H_0$ | Right P-v.$\|H_0$ |
|---|---|---|---|---|---|---|
| $2002-2007$ | 2569 | 652 | 580.6702 | 12.6420 | 5.6423 | $1.0206e-08$ |
| $2008-2013$ | 1723 | 443 | $391,0772$ | 10.3558 | 5.0139 | $3.3133e-07$ |

## 4.5 Conclusions

In this Chapter we propose exact statistical to test the presence of an excess of attribute similarity among the elements of very small groups. We introduce a probability distribution that describes the distribution of attributes in a sample of small-size groups. The application shows how it is possible to automatically distinguish among different narrative styles by analysing the concentration of word attributes, such as, grammatical roles and sentiments, in small sequences of words. In Chapter 1, the results show the key role of 2-grams, as a unit in Language Model and as a context of a word. We extended such results studying the attributes of the 2-grams. The entrainment of adverbs suggests that the bankruptcy language is more complex and convoluted. This language style is associated to narrative style, usually used by firm involved in irresponsible business conduct. Moreover, we can extend our analysis of language applying the extensions of the model proposed in sec 4.3. We can consider the sentences as groups of variable size and other word attributes. Other applications within a more sociological framework also prove the effectiveness of the proposed model to identify and model attribute concentration in small groups. The analysis supports our model's effectiveness as an appropriate null hypothesis to test whether children affected by neurological disorders tended to group. The application to Italian households shows a sort of imitation effect of children in replicating the family background model. Finally, the concentration of children's gender in households with two or three children suggests that the choice of families to have a third child is not indifferent to the genders of their first two children.

# Conclusions

This dissertation aims to develop and explore new statistical methods for textual analysis. We present several approaches to face different tasks and frameworks. Co-occurrences of words have a central role in the experiments proposed and we show how the context of words has a crucial role in text analysis. In Chapter 1, we observe the fuzziness of words in accounting reports, facing the problem of Word Sense Disambiguation (WSD) in terms of sentiment. We propose the application of bootstrap t-test to measure the semantic similarity among words, showing how a word is used in the "fail" or "health" context and proving the limitations of sentiment word lists. Chapters 2 and 3 concern the study of meaning of words by analysing their distributional semantic properties. In Chapter 2, we focus on interpreting, exploring and understanding the semantic space of words. The work's main contribution is to provide a robust, statistically rigorous method to evaluate the outputs of a topic model measuring the semantic coherence of estimated topics in terms of human interpretability. We propose a rigorous statistical approach based on hypothesis testing to develop a new topic-coherence measure, $Coh_{SVN}$, that approximates human ratings better than state-of-the-art methods. Indeed, the proposed measure ranges between [0, 1], providing a more readable framework for evaluating the coherence of the topics. The proposed approach allows one to distinguish between high-quality and low-quality topics using a battery of statistical tests. Then, in Chapter 3, we moving the research to interpretative and modellistic framework. The construction of statistically-validated word co-occurrence networks is generalized to study the main semantic similarities of words, moving the focus of the research to face the tasks of document clustering and topic extraction. The proposed methodology is at the core of an NLP toolkit to help researchers to perform Systematic Literature Reviews (SLR), as demonstrated through an application of the methodology to the themes of cobranding and coopetition. The method allows the selection of relevant studies on a specific topic and effectively extracts sub-topics considered in the (automatically selected) collection of papers, reducing the authors' bias with timing

advantages. We propose an unsupervised approach that does not need hyperparameters fine-tuning and is suitable for applications involving short texts (abstracts), overcoming some issues of topic modelling. In Appendix C, we also present an application of the method proposed to construct the network of documents. The results show the method's efficacy in capturing the key semantic relationships among words and representing similarities among documents. Furthermore, we study the short sequences ($n$-grams) as small words' context and the concentration of attributes in such local contexts. In Chapter 1, we face the problem of extracting linguistic features from a bankruptcy language, focusing the analysis on sequences of words and how words co-occur in short windows ($n$-grams). We highlight how the meaning of words varies with the context, narration and stories in which they are embedded. Indeed, our results demonstrate that we could effectively construct a statistical language model for predicting the corporate default, providing interpretable outputs that could give insight into why a company went into bankruptcy, allowing the investigation of the sentences classified as "negative" and moving away from fixed word lists. Finally, in Chapter 4, we present a new discrete probability distribution that aims to describe the concentration of word attributes in short sentences. The results show significant linguistic differences between the annual reports of healthy companies and companies that will go bankrupt in the near future (less than two years). Other applications within a more sociological framework also prove the effectiveness of the proposed model to identify and model attributes concentration in small groups. To conclude, the Statistically Validated Networks method has suitable properties to text analysis tasks. It allows the construction of word co-occurrence network representing the semantic similarities among words. Moreover, the test statistics provide a filter for the least informative connections since it considers the marginals occurrence of words. The results show how we can retrieve the semantic space of words and documents through the network representation. Indeed, our approach is not stochastic and needs less computational effort than probabilistic topic models. These characteristics are suitable for facing unsupervised clustering tasks. We believe this dissertation contributes to the text analysis research, providing insight into future research development. Moreover, our approaches meet the expectation of further development in improving automated textual analysis and demonstrate that we could effectively bridge the performance gap between deep learning, topic models and dictionary-based approaches.

# Appendixes

## Appendix A

**Table A1:** Description of Industrial Sectors (SIC codes)

| SIC | Category description |
|---|---|
| 1221 | bituminous coal & lignite surface mining |
| 1311 | crude petroleum & natural gas |
| 1381 | drilling oil & gas wells |
| 1382 | oil & gas field exploration services |
| 1389 | oil & gas field services, nec |
| 2020 | dairy products |
| 2300 | apparel & other finishd prods of fabrics & similar matl |
| 2621 | paper mills |
| 2670 | converted paper & paperboard prods (no contaners/boxes) |
| 2821 | plastic materials, synth resins & nonvulcan elastomers |
| 2834 | pharmaceutical preparations |
| 2870 | agricultural chemicals |
| 3330 | primary smelting & refining of nonferrous metals |
| 3334 | primary production of aluminum |
| 3533 | oil & gas field machinery & equipment |
| 3720 | aircraft & parts |
| 4400 | water transportation |
| 4412 | deep sea foreign transportation of freight |
| 4512 | air transportation, scheduled |
| 4522 | air transportation, nonscheduled |
| 4813 | telephone communications (no radiotelephone) |
| 4832 | radio broadcasting stations |
| 4911 | electric services |
| 4922 | natural gas transmission |
| 4931 | electric & other services combined |
| 5051 | wholesale-metals service centers & offices |
| 5063 | wholesale-electrical apparatus & equipment, wiring supplies |
| 5122 | wholesale-drugs, proprietaries & druggists' sundries |
| 5311 | retail-department stores |
| 5411 | retail-grocery stores |
| 5600 | retail-apparel & accessory stores |
| 5712 | retail-furniture stores |
| 5731 | retail-radio, tv & consumer electronics stores |
| 5812 | retail-eating places |
| 5900 | retail-miscellaneous retail |
| 5945 | retail-hobby, toy & game shops |
| 5960 | retail-nonstore retailers |
| 7320 | services-consumer credit reporting, collection agencies |
| 7371 | services-computer programming services |
| 7380 | services-miscellaneous business services |
| 8011 | services-offices & clinics of doctors of medicine |
| 8062 | services-general medical & surgical hospitals, nec |
| 8200 | services-educational services |

**Table A2:** Description of Companies

| Group | Company | CIK | SIC | Date of Failure (Y/M/D) | State |
|---|---|---|---|---|---|
| 1 | Sanchez Energy Corp. | 1528837 | 1311 | 2019/08/11 | TX |
| 1 | Frank's International N.V. | 1575828 | 1389 | - | NH |
| 2 | Key Energy Services Inc. | 318996 | 1389 | 2016/10/24 | TX |
| 2 | Comstock Resources Inc. | 23194 | 1311 | - | TX |
| 3 | Paperweight Development Corp. | 1166365 | 2670 | 2017/10/01 | WI |
| 3 | Neenah Inc. | 1296435 | 2621 | - | GA |
| 4 | Parker Drilling Co. | 76321 | 1381 | 20181212 | TX |
| 4 | Diamondback Energy, Inc. | 1539838 | 1311 | - | TX |
| 5 | Fairway Group Holdings Corp | 1555492 | 5411 | 2016/05/02 | NY |
| 5 | Fresh Market, Inc. | 1489979 | 5411 | - | NC |
| 6 | Ciber Inc | 918581 | 7371 | 2017/04/09 | CO |
| 6 | Syntel Inc | 1040426 | 7371 | - | MI |
| 7 | Cloud Peak Energy Inc. | 1441849 | 1221 | 2019/05/10 | WY |
| 7 | Alliance Holdings GP, L.P. | 1344980 | 1221 | - | OK |
| 8 | Breitburn Energy Partners LP | 1357371 | 1311 | 2016/05/15 | CA |
| 8 | Atwood Oceanics Inc. | 8411 | 1381 | - | TX |
| 9 | iHeartMedia, Inc. | 1400891 | 4832 | 2018/03/14 | TX |
| 9 | iHeartCommunications, Inc. | 739708 | 4832 | - | TX |
| 10 | Tops Holding Ii Corp. | 1584701 | 5411 | 2018/02/21 | NY |
| 10 | Weis Markets Inc. | 105418 | 5411 | - | PA |
| 11 | Battalion Oil Corp. | 1282648 | 1311 | 2016/07/27 | TX |
| 11 | Vaalco Energy Inc. | 894627 | 1311 | - | TX |
| 12 | Gulfmark Offshore Inc. | 1030749 | 3533 | 2017/05/17 | TX |
| 12 | Oil States International, Inc. | 1121484 | 3533 | - | TX |
| 13 | Grizzly Energy, LLC | 1384072 | 1311 | 2017/02/01 | TX |
| 13 | Laredo Petroleum, Inc. | 1528129 | 1311 | - | OK |
| 14 | Approach Resources Inc. | 1405073 | 1311 | 2019/11/18 | TX |
| 14 | Tetra Technologies Inc. | 844965 | 1311 | - | TX |
| 15 | SquareTwo Financial Corp. | 1505966 | 7320 | 2017/03/19 | CO |
| 15 | Synchronoss Technologies Inc. | 1131554 | 7371 | - | NJ |
| 16 | Ultra Petroleum Corp. | 1022646 | 1311 | 2016/04/29 | CO |
| 16 | Helmerich & Payne, Inc. | 46765 | 1381 | - | OK |
| 17 | Phi Inc. | 350403 | 4522 | 2019/03/14 | LA |
| 17 | Spirit Airlines, Inc. | 1498710 | 4512 | - | FL |
| 18 | Basic Energy Services, Inc. | 1109189 | 1389 | 2016/10/25 | TX |
| 18 | SemGroup Corp. | 1489136 | 1389 | - | OK |
| 19 | Bonanza Creek Energy, Inc. | 1509589 | 1311 | 2017/01/04 | CO |
| 19 | Willbros Group, Inc. | 1449732 | 1389 | - | TX |
| 20 | Energy XXI Ltd | 1343719 | 1382 | 2016/04/14 | Bermuda |
| 20 | SM Energy Co. | 893538 | 1311 | - | CO |
| 21 | Melinta Therapeutics, Inc. | 1461993 | 2834 | 2019/12/27 | CT |
| 21 | Arqule Inc. | 1019695 | 2834 | - | MA |
| 22 | A. M. Castle & Co. | 18172 | 5051 | 2017/06/18 | IL |
| 22 | Olympic Steel Inc. | 917470 | 5051 | - | OH |
| 23 | Global Geophysical Services Inc. | 1311486 | 1382 | 2016/08/03 | TX |
| 23 | Petroquest Energy Inc. | 872248 | 1311 | - | LA |
| 24 | Venoco, Inc. | 1313024 | 1311 | 2016/03/18 | CO |
| 24 | Resolute Energy Corp. | 1469510 | 1311 | - | CO |
| 25 | Cenveo, Inc. | 920321 | 2670 | 2018/02/02 | CT |
| 25 | Glatfelter Corp. | 41719 | 2621 | - | NC |
| 26 | Illinois Power Generating Co. | 1135361 | 4911 | 2016/12/09 | TX |
| 26 | El Paso Electric Co. | 31978 | 4911 | - | TX |
| 27 | Emerald Oil, Inc. | 1283843 | 1311 | 2016/03/22 | CO |
| 27 | Zion Oil & Gas Inc. | 1131312 | 1382 | - | TX |

| | | | | | |
|---|---|---|---|---|---|
| 28 | Rex Energy Corp. | 1397516 | 1311 | 2018/05/18 | PA |
| 28 | Goodrich Petroleum Corp. | 943861 | 1311 | - | TX |
| 29 | Aralez Pharmaceuticals Inc. | 1660719 | 2834 | 2018/08/10 | ON |
| 29 | PLx Pharma Inc. | 1497504 | 2834 | - | NJ |
| 30 | Aeropostale Inc | 1168213 | 5600 | 2016/05/04 | NY |
| 30 | Express, Inc. | 1483510 | 5600 | - | OH |
| 31 | Roan Resources, Inc. | 1326428 | 1311 | 2016/05/11 | OK |
| 31 | Southwestern Energy Co | 7332 | 1311 | - | TX |
| 32 | EP Energy Corp | 1584952 | 1311 | 2019/10/03 | TX |
| 32 | Wpx Energy, Inc. | 1518832 | 1311 | - | OK |
| 33 | Sandridge Energy Inc | 1349436 | 1311 | 2016/05/16 | OK |
| 33 | Range Resources Corp | 315852 | 1311 | - | TX |
| 34 | Gastar Exploration Inc. | 1431372 | 1311 | 2018/10/31 | TX |
| 34 | RSP Permian, Inc. | 1588216 | 1311 | - | TX |
| 35 | Penn Virginia Corp | 77159 | 1311 | 2016/05/12 | TX |
| 35 | W&T Offshore Inc | 1288403 | 1311 | - | TX |
| 36 | Itt Educational Services Inc | 922475 | 8200 | 2016/09/16 | IN |
| 36 | Stride, Inc. | 1157408 | 8200 | - | VA |
| 37 | Cumulus Media Inc. | 1058623 | 4832 | 2017/11/29 | GA |
| 37 | Cincinnati Bell Inc. | 716133 | 4813 | - | OH |
| 38 | Gymboree Corp. | 786110 | 2300 | 2017/06/11 | CA |
| 38 | Carters Inc. | 1060822 | 2300 | - | GA |
| 39 | Nobilis Health Corp. | 1409916 | 8062 | 2019/10/21 | TX |
| 39 | Rennova Health, Inc. | 931059 | 8062 | - | FL |
| 40 | Seventy Seven Energy Inc. | 1532930 | 1389 | 2016/06/07 | OK |
| 40 | Pdc Energy, Inc. | 77877 | 1311 | - | CO |
| 41 | Rentech, Inc. | 868725 | 2870 | 2017/12/19 | CA |
| 41 | Nektar Therapeutics | 906709 | 2834 | - | CA |
| 42 | New Source Energy Partners L.P. | 1560443 | 1311 | 2016/03/15 | OK |
| 42 | Atlas Resources Series 28-2010 L.P. | 1487561 | 1311 | - | PA |
| 43 | Triangle Petroleum Corp | 1281922 | 1311 | 2016/06/29 | CO |
| 43 | Erin Energy Corp. | 1402281 | 1381 | - | TX |
| 44 | Exco Resources Inc. | 316300 | 1311 | 2018/01/15 | TX |
| 44 | Bill Barrett Corp. | 1172139 | 1311 | - | CO |
| 45 | GenOn Energy, Inc. | 1126294 | 4911 | 2017/06/14 | NJ |
| 45 | Panhandle Eastern Pipe Line Company, Lp | 76063 | 4922 | - | TX |
| 46 | Tidewater Inc. | 98222 | 4400 | 2017/05/17 | TX |
| 46 | Seacor Holdings Inc. | 859598 | 4412 | - | FL |
| 47 | PG&E Corp. | 1004980 | 4931 | 2019/01/29 | CA |
| 47 | Nextera Energy Inc. | 753308 | 4911 | - | FL |
| 48 | Hexion Inc. | 13239 | 2821 | 2019/04/01 | OH |
| 48 | Avient Corp. | 1122976 | 2821 | - | OH |
| 49 | FTD Companies, Inc. | 1575360 | 5960 | 2019/06/03 | IL |
| 49 | Firstcash, Inc. | 840489 | 5900 | - | TX |
| 50 | Lri Holdings, Inc. | 1383875 | 5812 | 2016/08/08 | TN |
| 50 | Texas Roadhouse, Inc. | 1289460 | 5812 | - | KY |
| 51 | Nuverra Environmental Solutions, Inc. | 1403853 | 1389 | 2017/05/01 | AZ |
| 51 | Clayton Williams Energy Inc. | 880115 | 1311 | - | TX |
| 52 | Monitronics International Inc. | 1265107 | 7380 | 2019/06/30 | TX |
| 52 | Ascent Capital Group, Inc. | 1437106 | 7380 | - | CO |
| 53 | Amplify Energy Corp. | 1533924 | 1311 | 2016/04/30 | TX |
| 53 | Kosmos Energy Ltd. | 1509991 | 1311 | - | TX |
| 54 | Orexigen Therapeutics, Inc. | 1382911 | 2834 | 2018/03/12 | CA |
| 54 | Dynavax Technologies Corp. | 1029142 | 2834 | - | CA |
| 55 | hhgregg, Inc. | 1396279 | 5731 | 2017/03/06 | IN |
| 55 | Haverty Furniture Companies Inc. | 216085 | 5712 | - | GA |
| 56 | Southcross Energy Partners, L.P. | 1547638 | 4922 | 2019/04/01 | TX |

| | | | | | |
|---|---|---|---|---|---|
| 56 | Chugach Electric Association Inc. | 878004 | 4911 | - | AK |
| 57 | Claires Stores Inc. | 34115 | 5600 | 2018/03/19 | IL |
| 57 | Hanesbrands Inc. | 1359841 | 5600 | - | NC |
| 58 | Dean Foods Co. | 931336 | 2020 | 2019/11/12 | TX |
| 58 | Lifeway Foods, Inc. | 814586 | 2020 | - | IL |
| 59 | Horsehead Holding Corp. | 1385544 | 3330 | 2016/02/02 | PA |
| 59 | United States Antimony Corp. | 101538 | 3330 | - | MT |
| 60 | Real Industry, Inc. | 38984 | 5063 | 2017/11/17 | OH |
| 60 | Houston Wire & Cable Co. | 1356949 | 5063 | - | TX |
| 61 | Mattress Firm Holding Corp. | 1419852 | 5712 | 2018/10/05 | TX |
| 61 | Conns Inc. | 1223389 | 5731 | - | TX |
| 62 | Perfumania Holdings, Inc. | 880460 | 5900 | 2017/08/26 | NY |
| 62 | ZAGG Inc. | 1296205 | 5900 | - | UT |
| 63 | CHC Group Ltd. | 1586300 | 4522 | 2016/05/05 | BC |
| 63 | Hawaiian Holdings Inc. | 1172222 | 4512 | - | HI |
| 64 | Stone Energy Corp. | 904080 | 1311 | 2016/12/14 | LA |
| 64 | Oceaneering International Inc. | 73756 | 1389 | - | TX |
| 65 | Harvest Oil & Gas Corp. | 1361937 | 1311 | 2018/04/02 | TX |
| 65 | Carrizo Oil & Gas Inc. | 1040593 | 1311 | - | TX |
| 66 | Weatherford International plc | 1603923 | 3533 | 2019/07/01 | TX |
| 66 | NOV Inc. | 1021860 | 3533 | - | TX |
| 67 | westmoreland Coal Co. | 106455 | 1221 | 2018/10/09 | CO |
| 67 | Royal Energy Resources, Inc. | 1102392 | 1221 | - | SC |
| 68 | Peabody Energy Corp | 1064728 | 1221 | 2016/0413 | MO |
| 68 | Alliance Resource Partners Lp | 1086600 | 1221 | - | OK |
| 69 | Noranda Aluminum Holding CORP | 1422105 | 3334 | 2016/02/08 | TN |
| 69 | Century Aluminum Co | 949157 | 3334 | - | IL |
| 70 | International Shipholding Corp | 278041 | 4412 | 2016/08/01 | AL |
| 70 | Rand Logistics, Inc. | 1294250 | 4400 | - | NJ |
| 71 | Forbes Energy Services Ltd. | 1434842 | 1389 | 2017/01/22 | TX |
| 71 | Harvest Natural Resources, Inc. | 845289 | 1311 | - | TX |
| 72 | Orchids Paper Products Co. | 1324189 | 2621 | 2019/04/01 | OK |
| 72 | It Tech Packaging, Inc. | 1358190 | 2670 | - | China |
| 73 | 21st Century Oncology Holdings, Inc. | 1503518 | 8011 | 2017/05/25 | FL |
| 73 | Sunlink Health Systems Inc. | 96793 | 8062 | - | GA |
| 74 | Bristow Group Inc. | 73887 | 4522 | 2019/05/11 | TX |
| 74 | Era Group Inc. | 1525221 | 4522 | - | TX |
| 75 | Erickson Inc. | 1490165 | 3720 | 2016/11/08 | OR |
| 75 | Aar Corp. | 1750 | 3720 | - | IL |
| 76 | Hercules Offshore, Inc. | 1330849 | 1381 | 2015/08/13 | TX |
| 76 | Gran Tierra Energy Inc. | 1273441 | 1311 | - | AB |
| 77 | Chaparral Energy, Inc. | 1346980 | 1311 | 2016/05/09 | OK |
| 77 | Rpc Inc. | 742278 | 1389 | - | GA |
| 78 | Toys R Us Inc | 1005414 | 5945 | 2017/09/19 | NJ |
| 78 | Suburban Propane Partners Lp | 1005210 | 5900 | - | NJ |
| 79 | Bon Ton Stores Inc. | 878079 | 5311 | 2018/02/04 | PA |
| 79 | Sears Hometown & Outlet Stores, Inc. | 1548309 | 5311 | - | IL |
| 80 | Jones Energy, Inc. | 1573166 | 1311 | 2019/04/14 | TX |
| 80 | Northern Oil & Gas, Inc. | 1104485 | 1311 | - | MN |
| 81 | Aceto Corp. | 2034 | 5122 | 2019/02/19 | NY |
| 81 | Cosmos Holdings Inc. | 1474167 | 5122 | - | IL |
| 82 | Verso Corp. | 1421182 | 2621 | 2016/01/26 | OH |
| 82 | Kapstone Paper & Packaging Corp | 1325281 | 2621 | - | IL |

# Appendix B

**Table B1:** Coherence scores

| | $Coh_{SVN}$ | | | | | | state-of-the-art | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Topic | $J$ | $Dc$ | $SS$ | $FM$ | $D_\rho$ | $\tilde{R}$ | $\tilde{p}_v$ | PMI [158] | $UMass$ [156] | NPMI [124] | $CV$ [189] | $tf$-$idf$ [160] | HumanJ |
| $z_1$ | 0.006 | 0.012 | 0.003 | 0.137 | 0.076 | 0.037 | 0.133 | -2.619 | -5.988 | -0.119 | 0.326 | -296.42 | 2.332 |
| $z_2$ | 0.004 | 0.008 | 0.002 | 0.089 | 0.049 | 0.022 | 0.084 | -5.979 | -9.360 | -0.272 | 0.309 | -492.41 | 1.391 |
| $z_3$ | 0.010 | 0.018 | 0.005 | 0.291 | 0.159 | 0.060 | 0.265 | 0.926 | -1.965 | 0.084 | 0.663 | -87.84 | 3.743 |
| $z_4$ | 0.007 | 0.014 | 0.004 | 0.156 | 0.086 | 0.042 | 0.144 | -6.258 | -9.391 | -0.208 | 0.392 | -498.77 | 1.599 |
| $z_5$ | 0.006 | 0.011 | 0.003 | 0.140 | 0.078 | 0.037 | 0.131 | -3.533 | -6.818 | -0.148 | 0.321 | -344.72 | 2.416 |
| $z_6$ | 0.070 | 0.128 | 0.037 | 0.937 | 0.545 | 0.422 | 0.966 | 1.632 | -0.900 | 0.257 | 0.899 | -5.06 | 3.847 |
| $z_7$ | 0.021 | 0.039 | 0.011 | 0.345 | 0.194 | 0.118 | 0.317 | 0.864 | -1.365 | 0.127 | 0.627 | -62.09 | 2.688 |
| $z_8$ | 0.009 | 0.017 | 0.005 | 0.228 | 0.125 | 0.058 | 0.219 | 0.846 | -1.826 | 0.004 | 0.562 | -98.04 | 2.351 |
| $z_9$ | 0.024 | 0.043 | 0.013 | 0.356 | 0.206 | 0.148 | 0.354 | -1.721 | -5.016 | -0.067 | 0.293 | -247.09 | 3.178 |
| $z_{10}$ | 0.018 | 0.033 | 0.009 | 0.277 | 0.159 | 0.127 | 0.281 | -1.674 | -4.393 | -0.102 | 0.303 | -237.55 | 2.416 |
| $z_{11}$ | 0.019 | 0.037 | 0.010 | 0.397 | 0.223 | 0.140 | 0.394 | 0.977 | -1.738 | 0.063 | 0.622 | -93.95 | 3.381 |
| $z_{12}$ | 0.017 | 0.032 | 0.009 | 0.359 | 0.201 | 0.094 | 0.348 | 0.804 | -1.437 | 0.046 | 0.587 | -80.95 | 2.851 |
| $z_{13}$ | 0.061 | 0.111 | 0.033 | 0.886 | 0.506 | 0.341 | 0.848 | 2.437 | -1.716 | 0.365 | 0.911 | -10.61 | 3.431 |
| $z_{14}$ | 0.007 | 0.013 | 0.004 | 0.178 | 0.098 | 0.051 | 0.178 | -0.579 | -3.907 | -0.013 | 0.484 | -190.96 | 2.233 |
| $z_{15}$ | 0.015 | 0.028 | 0.007 | 0.425 | 0.234 | 0.085 | 0.408 | 0.586 | -1.306 | 0.093 | 0.590 | -79.56 | 3.356 |
| $z_{16}$ | 0.041 | 0.078 | 0.021 | 0.819 | 0.460 | 0.271 | 0.815 | 1.351 | -1.030 | 0.229 | 0.845 | -34.27 | 3.406 |
| $z_{17}$ | 0.155 | 0.014 | 0.505 | 0.500 | 0.247 | 0.048 | 0.026 | -2.040 | -4.797 | 0.007 | 0.384 | -258.17 | 2.901 |
| $z_{18}$ | 0.047 | 0.002 | 0.081 | 0.080 | 0.008 | 0.009 | 0.005 | -2.305 | -4.582 | -0.117 | 0.234 | -277.77 | 2.084 |
| $z_{19}$ | 0.028 | 0.051 | 0.015 | 0.623 | 0.343 | 0.152 | 0.603 | 1.654 | -0.990 | 0.230 | 0.867 | -32.36 | 3.535 |
| $z_{20}$ | 0.006 | 0.012 | 0.003 | 0.103 | 0.059 | 0.050 | 0.105 | -7.400 | -11.466 | -0.304 | 0.378 | -575.41 | 1.579 |
| $z_{21}$ | 0.023 | 0.042 | 0.012 | 0.419 | 0.235 | 0.142 | 0.396 | -0.762 | -4.153 | 0.089 | 0.579 | -191.81 | 3.460 |
| $z_{22}$ | 0.014 | 0.026 | 0.007 | 0.292 | 0.163 | 0.085 | 0.261 | -3.339 | -6.392 | -0.090 | 0.328 | -323.97 | 2.465 |
| $z_{23}$ | 0.023 | 0.043 | 0.012 | 0.445 | 0.247 | 0.130 | 0.418 | 0.818 | -1.308 | 0.125 | 0.661 | -75.73 | 3.644 |
| $z_{24}$ | 0.012 | 0.022 | 0.006 | 0.224 | 0.126 | 0.048 | 0.184 | 0.298 | -1.236 | 0.023 | 0.413 | -97.63 | 2.856 |
| $z_{25}$ | 0.028 | 0.053 | 0.014 | 0.600 | 0.333 | 0.165 | 0.564 | 1.183 | -1.350 | 0.169 | 0.781 | -47.16 | 3.396 |
| $z_{26}$ | 0.021 | 0.039 | 0.011 | 0.356 | 0.201 | 0.111 | 0.351 | 0.515 | -1.683 | 0.061 | 0.544 | -97.11 | 2.772 |
| $z_{27}$ | 0.007 | 0.012 | 0.003 | 0.132 | 0.075 | 0.043 | 0.135 | -7.618 | -12.673 | -0.285 | 0.379 | -589.28 | 1.837 |
| $z_{28}$ | 0.026 | 0.047 | 0.014 | 0.467 | 0.264 | 0.139 | 0.447 | 1.287 | -1.086 | 0.155 | 0.740 | -58.45 | 3.223 |
| $z_{29}$ | 0.018 | 0.034 | 0.010 | 0.396 | 0.219 | 0.104 | 0.370 | 0.825 | -1.457 | 0.127 | 0.674 | -72.54 | 3.307 |
| $z_{30}$ | 0.005 | 0.010 | 0.003 | 0.181 | 0.100 | 0.048 | 0.150 | -3.518 | -7.015 | -0.108 | 0.312 | -346.57 | 1.837 |

**Table B2:** Ranking coherence scores

| | $Coh_{SVN}$ | | | | | | | state-of-the-art | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Topic | $J$ | $Dc$ | $SS$ | $FM$ | $D_\rho$ | $\tilde{R}$ | $\tilde{p}_v$ | PMI [158] | $UMass$ [156] | NPMI [124] | $CV$ [189] | $tf\text{-}idf$ [160] | HumanJ |
| $z_1$ | 25 | 25 | 25 | 26 | 26 | 28 | 26 | 23 | 23 | 25 | 24 | 23 | 23 |
| $z_2$ | 30 | 30 | 30 | 29 | 29 | 30 | 29 | 27 | 27 | 28 | 27 | 27 | 30 |
| $z_3$ | 20 | 20 | 20 | 18 | 18 | 19 | 18 | 8 | 16 | 12 | 8 | 12 | 2 |
| $z_4$ | 22 | 22 | 22 | 24 | 24 | 27 | 24 | 28 | 28 | 27 | 19 | 28 | 28 |
| $z_5$ | 27 | 27 | 27 | 25 | 25 | 29 | 27 | 26 | 25 | 26 | 25 | 25 | 20 |
| $z_6$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 1 | 2 | 2 | 1 | 1 |
| $z_7$ | 11 | 12 | 11 | 16 | 16 | 13 | 16 | 9 | 9 | 8 | 10 | 7 | 18 |
| $z_8$ | 21 | 21 | 21 | 20 | 21 | 20 | 20 | 10 | 15 | 18 | 15 | 16 | 22 |
| $z_9$ | 8 | 8 | 8 | 15 | 13 | 7 | 13 | 20 | 22 | 20 | 29 | 20 | 13 |
| $z_{10}$ | 15 | 15 | 15 | 19 | 19 | 12 | 17 | 19 | 19 | 22 | 28 | 19 | 20 |
| $z_{11}$ | 13 | 13 | 13 | 11 | 11 | 9 | 11 | 7 | 14 | 13 | 11 | 13 | 9 |
| $z_{12}$ | 16 | 16 | 16 | 13 | 15 | 16 | 15 | 13 | 10 | 15 | 13 | 11 | 16 |
| $z_{13}$ | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 13 | 1 | 1 | 2 | 6 |
| $z_{14}$ | 23 | 23 | 23 | 23 | 23 | 21 | 22 | 17 | 17 | 19 | 17 | 17 | 24 |
| $z_{15}$ | 17 | 17 | 17 | 9 | 10 | 17 | 9 | 14 | 6 | 10 | 12 | 10 | 10 |
| $z_{16}$ | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 3 | 4 | 4 | 4 | 7 |
| $z_{17}$ | 7 | 6 | 7 | 6 | 6 | 5 | 6 | 21 | 21 | 17 | 20 | 21 | 14 |
| $z_{18}$ | 29 | 29 | 29 | 30 | 30 | 25 | 30 | 22 | 20 | 24 | 30 | 22 | 25 |
| $z_{19}$ | 4 | 5 | 4 | 4 | 4 | 6 | 4 | 2 | 2 | 3 | 3 | 3 | 4 |
| $z_{20}$ | 26 | 26 | 26 | 28 | 28 | 22 | 28 | 29 | 29 | 30 | 22 | 29 | 29 |
| $z_{21}$ | 10 | 10 | 9 | 10 | 9 | 8 | 10 | 18 | 18 | 11 | 14 | 18 | 5 |
| $z_{22}$ | 18 | 18 | 18 | 17 | 17 | 18 | 19 | 24 | 24 | 21 | 23 | 24 | 19 |
| $z_{23}$ | 9 | 9 | 10 | 8 | 8 | 11 | 8 | 12 | 7 | 9 | 9 | 9 | 3 |
| $z_{24}$ | 19 | 19 | 19 | 21 | 20 | 24 | 21 | 16 | 5 | 16 | 18 | 15 | 15 |
| $z_{25}$ | 5 | 4 | 5 | 5 | 5 | 4 | 5 | 6 | 8 | 5 | 5 | 5 | 8 |
| $z_{26}$ | 12 | 11 | 12 | 14 | 14 | 14 | 14 | 15 | 12 | 14 | 16 | 14 | 17 |
| $z_{27}$ | 24 | 24 | 24 | 27 | 27 | 26 | 25 | 30 | 30 | 29 | 21 | 30 | 26 |
| $z_{28}$ | 6 | 7 | 6 | 7 | 7 | 10 | 7 | 5 | 4 | 6 | 6 | 6 | 12 |
| $z_{29}$ | 14 | 14 | 14 | 12 | 12 | 15 | 12 | 11 | 11 | 7 | 7 | 8 | 11 |
| $z_{30}$ | 28 | 28 | 28 | 22 | 22 | 23 | 23 | 25 | 26 | 23 | 26 | 26 | 26 |

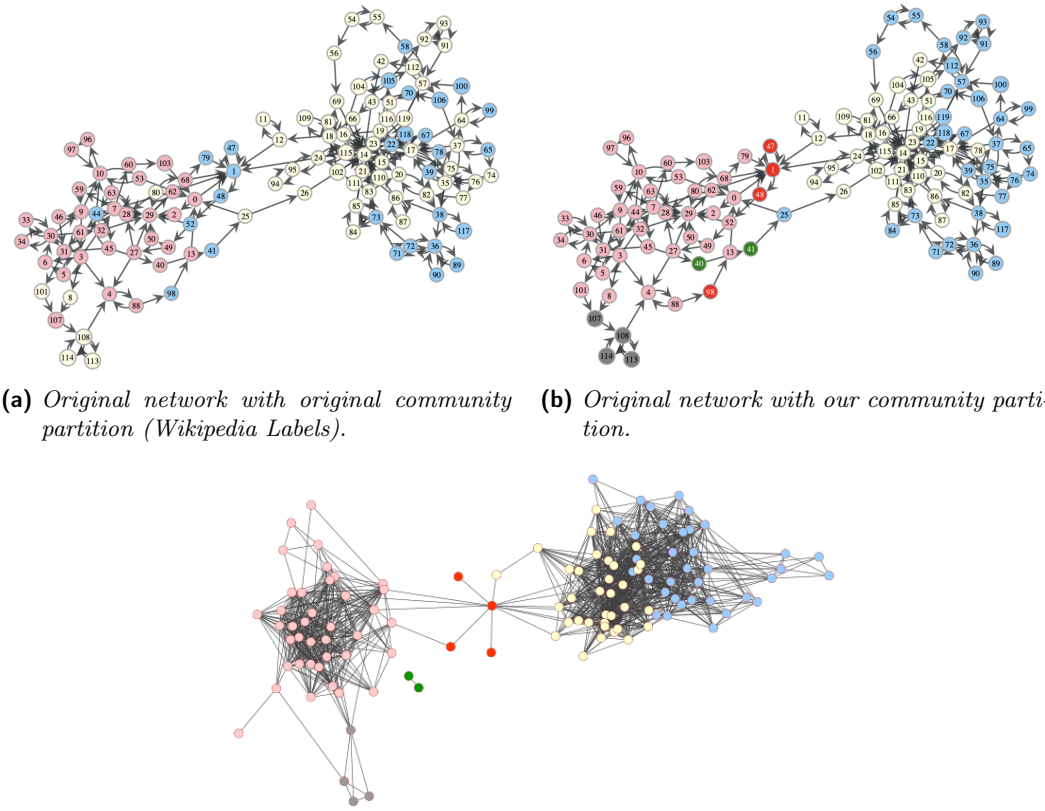
**Table B3:** Spearman rank correlation coefficient and Pearson correlation coefficient with human judgments for metrics without noise

| | *Correlation coefficient without noise* | |
|---|---|---|
| Method | Spearman | Pearson |
| $J$ | 0.81 | 0.67 |
| $Dc$ | 0.81 | 0.68 |
| $SS$ | 0.81 | 0.67 |
| $FM$ | 0.86 | **0.78** |
| $D_\rho$ | **0.87** | 0.77 |
| $\tilde{R}$ | 0.79 | 0.66 |
| $\tilde{p}_v$ | 0.86 | 0.77 |
| PMI [158] | 0.80 | 0.84 |
| $UMass$ [156] | 0.75 | 0.81 |
| NPMI [124] | **0.88** | **0.87** |
| $CV$ [189] | 0.77 | 0.76 |
| $tf\text{-}idf$ [160] | 0.81 | 0.85 |

# Appendix C

The dataset consists of 120 Wikipedia articles. Each article has a tag related to one of the following scientific areas: Subfields of physics (28 documents), Branches of biology (35 documents) and Fields of mathematics (57 documents). The dataset can be represented as a network where documents are nodes, and the hyperlinks are directed edges. The network has 120 nodes and 309 edges (hyperlinks). We use the Statistically Validated Networks method in 2 steps. In the first step, we represent words in a network. We split each document into sentences with a fixed length of 30. Then, we construct a bipartite network in which words and sentences make the two sets of nodes. A link is set between a word and a sentence if the word belongs to that sentence. Then, in the second step, we reconstruct a new bipartite network where the two sets of nodes are the documents and the validated pairs of words, where a link is set between a word and a document if the word belongs to that document. Again, we apply the SVN method but are now projecting the set of documents and constructing a network of documents. The weights of links are computed following Eq. (2.13). We use in both applications the Bonferroni correction with $alpha = 0.05$. Finally, we infer the communities of documents on the weighted network of documents. The Figure 4.3, we compare the partition of the original network and the partition of the SVN of documents. Figure (a) represents the original network with the original partition; Figure (b) represents the original network with our partition in SVN of documents; Figure (c) represents the SVN of documents with its partition. Moreover, we prove the efficacy of the SVN method as a filtering tool. As in Hyland et al. (2021) [111], we use the "*Maximum partition overlap*" to measure the similarity among partitions. We compare the original partition of Wikipedia and the ones provided by the Stochastic Block Model (SBM) and the same model when considering only the validated words through the SVN method described above. Hyland et al. (2021) [111] used the hierarchical SBM (hSBM), including hyperlinks in the model. We apply the SVN method to extract the validated words, and then we apply the hSBM. Table C1 shows the means and standard deviations of the scores among 100 replicates. The results show that the SBM and hSBM achieve better if we consider only the words validated through the SVN. Moreover, our method described above achieves the same result as hSBM combined with SVN. However, our method is not stochastic and needs less computational effort. Finally, in Table C2 we

show the nodes with the highest degree centrality measure in the SVN of documents. The articles related to these nodes seem to be methodological and interdisciplinary.



**(a)** *Original network with original community partition (Wikipedia Labels).*

**(b)** *Original network with our community partition.*

**(c)** *Statistically validated network of articles.*

**Figure 4.3:** Network representations of Wikipedia's articles

**Table C1:** Maximum partition overlap of the consensus partitions between the model. The values correspond to the mean over 100 replicates; standard deviation in parenthesis.

| Model | Model Partition vs True Partition |
|---|---|
| SVN | 0.7(·) |
| SBM | 0.5(0.05) |
| SVN + SBM | 0.58(0.06) |
| hSBM | 0.6(0.05) |
| SVN + hSBM | 0.7(0.06) |

**Table C2:** Betweenness centrality of articles

| article names | betweenness centrality |
|---|---|
| X-ray crystallography | 0.48 |
| Quantum field theory | 0.16 |
| Macromolecule | 0.14 |
| Protein structure prediction | 0.14 |
| Macromolecular docking | 0.13 |
| Space group | 0.05 |
| Translation operator (quantum mechanics) | 0.04 |

# Appendix D - Software and libraries

Statistical Language Model in Section 1.5.2 was performed in `Mathematica`. LDA model in section 2.4 was implemented in `R 4.1.1`. The statistically Validated Networks proposed in sections 2.3 and 3.3 were implemented in `Python 3.7`. Statistical data analyses were performed in `Python 3.7`. The SBM and hSBM model in Appendix C were implemented in `Python 3.7`, codes are available at `https://topsbm.github.io`. The visualization of statistically validated networks were obtained by using `Cytoscape 3.8.2`. The Entrainment model presented in Chapter 4 was implemented in `Mathematica`.

# References


[1] Sequencepredict (2017). URL `https://reference.wolfram.com/language/ref/SequencePredict.html`

[2] Adame-Sánchez, C., Caplliure, E.M., Miquel-Romero, M.J.: Paving the way for coopetition: drivers for work–life balance policy implementation. Review of Managerial Science **12**(2), 519–533 (2018)

[3] Agarwal, S., Chen, V.Y., Zhang, W.: The information value of credit rating action reports: A textual analysis. Management Science **62**(8), 2218–2240 (2016)

[4] Ahmadi, Z., Martens, P., Koch, C., Gottron, T., Kramer, S.: Towards bankruptcy prediction: Deep sentiment mining to detect financial distress from business management reports. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pp. 293–302. IEEE (2018)

[5] Albano, A., Plaia, A.: Element weighted kemeny distance for ranking data. ELECTRONIC JOURNAL OF APPLIED STATISTICAL ANALYSIS, 14(1) pp. 117–145.s (2021)

[6] Aletras, N., Stevenson, M.: Evaluating topic coherence using distributional semantics. In: Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers, pp. 13–22 (2013)

[7] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B., Kochut, K.: A brief survey of text mining: Classification, clustering and extraction techniques. arXiv preprint arXiv:1707.02919 (2017)

[8] AlSumait, L., Barbará, D., Gentle, J., Domeniconi, C.: Topic significance ranking of lda generative models. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 67–82. Springer (2009)

[9] Altman, E.I.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The journal of finance **23**(4), 589–609 (1968)

[10] Altman, E.I., Haldeman, R.G., Narayanan, P.: Zetatm analysis a new model to identify bankruptcy risk of corporations. Journal of banking & finance **1**(1), 29–54 (1977)

[11] Altman, E.I., Iwanicz-Drozdowska, M., Laitinen, E.K., Suvas, A.: Financial distress prediction in an international context: A review and empirical analysis of altman's z-score model. Journal of International Financial Management & Accounting **28**(2), 131–171 (2017)

[12] Amata, R., Dagnino, G.B., Minà, A., Picone, P.M.: Managing coopetition in diversified firms: Insights from a qualitative case study. Long Range Planning **55**(4), 102128 (2022)

[13] Anand, V., Bochkay, K., Chychyla, R., Leone, A., et al.: Using python for text analysis in accounting research. Foundations and Trends® in Accounting **14**(3–4), 128–359 (2020)

[14] Andreini, D., Bettinelli, C., Pedeliento, G., Apa, R.: How do consumers see firms' family nature? a review of the literature. Family Business Review **33**(1), 18–37 (2020)

[15] Araci, D.: Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063 (2019)

[16] Arun, R., Suresh, V., Madhavan, C.V., Murthy, M.N.: On finding the natural number of topics with latent dirichlet allocation: Some observations. In: Pacific-Asia conference on knowledge discovery and data mining, pp. 391–402. Springer (2010)

[17] Bacon, E., Williams, M.D., Davies, G.: Coopetition in innovation ecosystems: A comparative analysis of knowledge transfer configurations. Journal of Business Research **115**, 307–316 (2020)

[18] Bastani, K., Namavari, H., Shaffer, J.: Latent dirichlet allocation (lda) for topic modeling of the cfpb consumer complaints. Expert Systems with Applications **127**, 256–271 (2019)

[19] Beattie, V.: Accounting narratives and the narrative turn in accounting research: Issues, theory, methodology, methods and a research framework. The British Accounting Review **46**(2), 111–134 (2014)

[20] Beattie, V., Davison, J.: Accounting narratives: storytelling, philosophising and quantification (2015)

[21] Beattie, V., McInnes, B., Fearnley, S.: A methodology for analysing and evaluating narratives in annual reports: a comprehensive descriptive profile and metrics for disclosure quality attributes. In: Accounting forum, 3, pp. 205–236. Elsevier (2004)

[22] Beaver, W.H., Correia, M., McNichols, M.F.: Do differences in financial reporting attributes impair the predictive ability of financial ratios for bankruptcy? Review of Accounting Studies **17**(4), 969–1010 (2012)

[23] Belford, M., Mac Namee, B., Greene, D.: Stability of topic modeling via matrix factorization. Expert Systems with Applications **91**, 159–169 (2018)

[24] Belk, R.W.: Handbook of qualitative research methods in marketing. Edward Elgar Publishing (2007)

[25] Beneish, M.D.: The detection of earnings manipulation. Financial Analysts Journal **55**(5), 24–36 (1999)

[26] Bengtsson, M., Kock, S.: " coopetition" in business networks—to cooperate and compete simultaneously. Industrial marketing management **29**(5), 411–426 (2000)

[27] Bengtsson, M., Kock, S.: Coopetition—quo vadis? past accomplishments and future challenges. Industrial marketing management **43**(2), 180–188 (2014)

[28] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) **57**(1), 289–300 (1995)

[29] Benjamini, Y., Yekutieli, D.: The control of the false discovery rate in multiple testing under dependency. Annals of statistics pp. 1165–1188 (2001)

[30] Berkhin, P.: A survey of clustering data mining techniques. In: Grouping multidimensional data, pp. 25–71. Springer (2006)

[31] Blei, D., Lafferty, J.: Correlated topic models. Advances in neural information processing systems **18**, 147 (2006)

[32] Blei, D.M.: Probabilistic topic models. Communications of the ACM **55**(4), 77–84 (2012)

[33] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)

[34] Borchert, P., Coussement, K., De Caigny, A., De Weerdt, J.: Extending business failure prediction models with textual website content using deep learning. European Journal of Operational Research (2022)

[35] Bouma, G.: Normalized (pointwise) mutual information in collocation extraction. Proceedings of GSCL pp. 31–40 (2009)

[36] Bouncken, R.B., Gast, J., Kraus, S., Bogers, M.: Coopetition: a systematic review, synthesis, and future research directions. Review of Managerial Science **9**(3), 577–601 (2015)

[37] Boyd-Graber, J.L., Hu, Y., Mimno, D., et al.: Applications of topic models, vol. 11. now Publishers Incorporated (2017)

[38] Branch, B.: The costs of bankruptcy: A review. International Review of Financial Analysis **11**(1), 39–57 (2002)

[39] Brandenburger, A.M., Nalebuff, B.J.: Universal lessons every manager can learn from andy grove's paranoia. Harvard business review p. 169 (1996)

[40] Brielmaier, C., Friesl, M.: The attention-based view: Review and conceptual extension towards situated attention. International Journal of Management Reviews (2022)

[41] Briner, R.B., Denyer, D., Rousseau, D.M.: Evidence-based management: concept cleanup time? Academy of management perspectives **23**(4), 19–32 (2009)

[42] Briner, R.B., Denyer, D., et al.: Systematic review and evidence synthesis as a practice and scholarship tool. Handbook of evidence-based management: Companies, classrooms and research pp. 112–129 (2012)

[43] Bris, A., Welch, I., Zhu, N.: The costs of bankruptcy: Chapter 7 liquidation versus chapter 11 reorganization. The journal of finance **61**(3), 1253–1303 (2006)

[44] Bundy, J., Pfarrer, M.D., Short, C.E., Coombs, W.T.: Crises and crisis management: Integration, interpretation, and research development. Journal of management **43**(6), 1661–1692 (2017)

[45] Buntine, W., Jakulin, A.: Discrete component analysis. In: International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection", pp. 1–33. Springer (2005)

[46] Calabrò, A., Vecchiarini, M., Gast, J., Campopiano, G., De Massis, A., Kraus, S.: Innovation in family firms: A systematic literature review and guidance for future research. International journal of management reviews **21**(3), 317–355 (2019)

[47] Cecchini, M., Aytug, H., Koehler, G.J., Pathak, P.: Making words work: Using financial text as a predictor of financial events. Decision support systems **50**(1), 164–175 (2010)

[48] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., Blei, D.M.: Reading tea leaves: How humans interpret topic models. In: Advances in neural information processing systems, pp. 288–296 (2009)

[49] Chatman, J.A., O'Reilly, C.A.: Paradigm lost: Reinvigorating the study of organizational culture. Research in Organizational Behavior **36**, 199–224 (2016)

[50] Cheah, I., Zainol, Z., Phau, I.: Conceptualizing country-of-ingredient authenticity of luxury brands. Journal of Business Research **69**(12), 5819–5826 (2016)

[51] Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling. Computer Speech & Language **13**(4), 359–394 (1999)

[52] Cheng, L.T., Zheng, J., Savova, G.K., Erickson, B.J.: Discerning tumor status from unstructured mri reports—completeness of information in existing reports and utility of automated natural language processing. Journal of digital imaging **23**(2), 119–132 (2010)

[53] Chiambaretto, P., Gurău, C.: David by goliath: what is co-branding and what is in it for smes. International Journal of Entrepreneurship and Small Business **31**(1), 103–122 (2017)

[54] Chiao, Y.C., Lin, C.C., Huang, C.J.: Competing and cooperating globally: how firms' multimarket contact relates to joint price elevation in coopetition networks. Journal of Business & Industrial Marketing (2020)

[55] Chowdhary, K.: Natural language processing. Fundamentals of artificial intelligence pp. 603–649 (2020)

[56] Chuang, J., Gupta, S., Manning, C., Heer, J.: Topic model diagnostics: Assessing domain relevance via topical alignment. In: International conference on machine learning, pp. 612–620. PMLR (2013)

[57] Chyung, S.Y., Roberts, K., Swanson, I., Hankinson, A.: Evidence-based survey design: The use of a midpoint on the likert scale. Performance Improvement **56**(10), 15–23 (2017)

[58] Corciolani, M., Nieri, F., Tuan, A.: Does involvement in corporate social irresponsibility affect the linguistic features of corporate social responsibility reports? Corporate Social Responsibility and Environmental Management **27**(2), 670–680 (2020)

[59] Crick, J.M.: Unpacking the relationship between a coopetition-oriented mindset and coopetition-oriented behaviours. Journal of Business & Industrial Marketing (2020)

[60] Crick, J.M., Crick, D.: The dark-side of coopetition: Influences on the paradoxical forces of cooperativeness and competitiveness across product-market strategies. Journal of Business Research **122**, 226–240 (2021)

[61] Croux, C., Dehon, C.: Influence functions of the spearman and kendall correlation measures. Statistical methods & applications **19**(4), 497–515 (2010)

[62] Dagnino, G.B., Minà, A.: Unraveling the philosophical foundations of co-opetition strategy. Management and Organization Review **17**(3), 490–523 (2021)

[63] Dagnino, G.B., Picone, P.M., Ferrigno, G.: Temporary competitive advantage: a state-of-the-art literature review and research directions. International Journal of Management Reviews **23**(1), 85–115 (2021)

[64] De Roeck, A., Sarkar, A., Garthwaite, P.H.: Defeating the homogeneity assumption. In: Proceedings of 7th International Conference on the Statistical Analysis of Textual Data (JADT), pp. 282–294. Citeseer (2004)

[65] Debellis, F., Rondi, E., Plakoyiannaki, E., De Massis, A.: Riding the waves of family firm internationalization: A systematic literature review, integrative framework, and research agenda. Journal of World Business **56**(1), 101144 (2021)

[66] Decker, C., Baade, A.: Consumer perceptions of co-branding alliances: Organizational dissimilarity signals and brand fit. Journal of brand management **23**(6), 648–665 (2016)

[67] Denny, M.J., Spirling, A.: Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. Political Analysis **26**(2), 168–189 (2018)

[68] Devece, C., Ribeiro-Soriano, D.E., Palacios-Marqués, D.: Coopetition as the new trend in inter-firm alliances: literature review and research patterns. Review of Managerial Science **13**(2), 207–226 (2019)

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

[70] Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)

[71] Dieng, A.B., Ruiz, F.J., Blei, D.M.: The dynamic embedded topic model. arXiv preprint arXiv:1907.05545 (2019)

[72] Dieng, A.B., Ruiz, F.J., Blei, D.M.: Topic modeling in embedding spaces. Transactions of the Association for Computational Linguistics **8**, 439–453 (2020)

[73] Dorfleitner, G., Priberny, C., Schuster, S., Stoiber, J., Weber, M., de Castro, I., Kammler, J.: Description-text related soft information in peer-to-peer lending–evidence from two leading european platforms. Journal of Banking & Finance **64**, 169–187 (2016)

[74] Dorn, S., Schweiger, B., Albers, S.: Levels, phases and themes of coopetition: A systematic literature review and research agenda. European Management Journal **34**(5), 484–500 (2016)

[75] Du Jardin, P.: Bankruptcy prediction using terminal failure processes. European Journal of Operational Research **242**(1), 286–303 (2015)

[76] Du Jardin, P.: Forecasting bankruptcy using biclustering and neural network-based ensembles. Annals of Operations Research **299**(1), 531–566 (2021)

[77] Dumais, S.T., Furnas, G.W., Landauer, T.K., Deerwester, S., Harshman, R.: Using latent semantic analysis to improve access to textual information. In: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 281–285 (1988)

[78] Durand, R., Grant, R.M., Madsen, T.L.: The expanding domain of strategic management research and the quest for integration. Strategic Management Journal **38**(1), 4–16 (2017)

[79] Eisenstein, J.: Natural language processing. Jacob Eisenstein (2018)

[80] Emond, E.J., Mason, D.W.: A new technique for high level decision support. Department of National Defence Canada, Operational Research Division (2000)

[81] Emond, E.J., Mason, D.W.: A new rank correlation coefficient with application to the consensus ranking problem. Journal of Multi-Criteria Decision Analysis **11**(1), 17–28 (2002)

[82] Feldman, R., Dagan, I.: Knowledge discovery in textual databases (kdt). In: KDD, vol. 95, pp. 112–117 (1995)

[83] Fischer, T., Krauss, C.: Deep learning with long short-term memory networks for financial market predictions. European journal of operational research **270**(2), 654–669 (2018)

[84] Fowlkes, E.B., Mallows, C.L.: A method for comparing two hierarchical clusterings. Journal of the American statistical association **78**(383), 553–569 (1983)

[85] von Friedrichs Grängsjö, Y.: Destination networking: Co-opetition in peripheral surroundings. International Journal of Physical Distribution & Logistics Management (2003)

[86] Gandhi, P., Loughran, T., McDonald, B.: Using annual report sentiment as a proxy for financial distress in us banks. Journal of Behavioral Finance **20**(4), 424–436 (2019)

[87] Genova, V.G., Tumminello, M., Enea, M., Aiello, F., Attanasio, M.: Student mobility in higher education: Sicilian outflow network and chain migrations. Electronic Journal of Applied Statistical Analysis **12**(4), 774–800 (2019)

[88] Gerlach, M., Peixoto, T.P., Altmann, E.G.: A network approach to topic models. Science advances **4**(7), eaaq1360 (2018)

[89] Gernsheimer, O., Kanbach, D.K., Gast, J.: Coopetition research-a systematic literature review on recent accomplishments and trajectories. Industrial Marketing Management **96**, 113–134 (2021)

[90] Gnyawali, D.R., Park, B.J.R.: Co-opetition between giants: Collaboration with competitors for technological innovation. Research policy **40**(5), 650–663 (2011)

[91] Goodman, J.T.: A bit of progress in language modeling. Computer Speech & Language **15**(4), 403–434 (2001)

[92] Greenhalgh, T.: How to read a paper: Papers that summarise other papers (systematic reviews and meta-analyses). Bmj **315**(7109), 672–675 (1997)

[93] Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National academy of Sciences **101**(suppl 1), 5228–5235 (2004)

[94] Grüning, M.: Artificial intelligence measurement of disclosure (aimd). European Accounting Review **20**(3), 485–519 (2011)

[95] Gruszczyński, M.: On unbalanced sampling in bankruptcy prediction. International Journal of Financial Studies **7**(2), 28 (2019)

[96] Haenlein, M., Huang, M.H., Kaplan, A.: Guest editorial: Business ethics in the era of artificial intelligence (2022)

[97] Haenlein, M., Kaplan, A., Tan, C.W., Zhang, P.: Artificial intelligence (ai) and management analytics. Journal of Management Analytics **6**(4), 341–343 (2019)

[98] Haibe-Kains, B., Adam, G.A., Hosny, A., Khodakarami, F., Waldron, L., Wang, B., McIntosh, C., Goldenberg, A., Kundaje, A., Greene, C.S., et al.: Transparency and reproducibility in artificial intelligence. Nature **586**(7829), E14–E16 (2020)

[99] Haji, A.A., Hossain, D.M.: Exploring the implications of integrated reporting on organisational reporting practice: Evidence from highly regarded integrated reporters. Qualitative Research in Accounting & Management (2016)

[100] Harris, Z.S.: Distributional structure. Word **10**(2-3), 146–162 (1954)

[101] Hassan, O.A., Marston, C.: Disclosure measurement in the empirical accounting literature-a review article. International Journal of Accounting (2019)

[102] Heafield, K., Pouzyrevsky, I., Clark, J.H., Koehn, P.: Scalable modified kneser-ney language model estimation. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 690–696 (2013)

[103] Hiebl, M.R.: Sample selection in systematic literature reviews of management research. Organizational research methods p. 1094428120986851 (2021)

[104] Higgins, C., Walker, R.: Ethos, logos, pathos: Strategies of persuasion in social/environmental reports. In: Accounting forum, vol. 36, pp. 194–208. Elsevier (2012)

[105] Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 50–57 (1999)

[106] Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 50–57 (1999)

[107] Hong, L., Davison, B.D.: Empirical study of topic modeling in twitter. In: Proceedings of the first workshop on social media analytics, pp. 80–88 (2010)

[108] Houvardas, J., Stamatatos, E.: N-gram feature selection for authorship identification. In: International conference on artificial intelligence: Methodology, systems, and applications, pp. 77–86. Springer (2006)

[109] Hoyle, A., Goel, P., Hian-Cheong, A., Peskov, D., Boyd-Graber, J., Resnik, P.: Is automated topic model evaluation broken? the incoherence of coherence. Advances in Neural Information Processing Systems **34**, 2018–2033 (2021)

[110] Huang, A.H., Lehavy, R., Zang, A.Y., Zheng, R.: Analyst information discovery and interpretation roles: A topic modeling approach. Management science **64**(6), 2833–2855 (2018)

[111] Hyland, C.C., Tao, Y., Azizi, L., Gerlach, M., Peixoto, T.P., Altmann, E.G.: Multilayer networks for text analysis with multiple data types. EPJ Data Science **10**(1), 33 (2021)

[112] Kaya, B.: Hotel recommendation system by bipartite networks and link prediction. Journal of Information Science **46**(1), 53–63 (2020)

[113] Keselj, F.P.D.S.V., Wang, S.: Language independent authorship attribution using character level language models. In: EACL 2003: 10th Conference of the European Chapter, Association for Computational Linguistics-Proceedings of the Conference, April 12th-17th 2003, Agro Hotel, Budapest, Hungary, p. 267. Citeseer (2003)

[114] Kešelj, V., Peng, F., Cercone, N., Thomas, C.: N-gram-based author profiles for authorship attribution. In: Proceedings of the conference pacific association for computational linguistics, PACLING, vol. 3, pp. 255–264 (2003)

[115] Khanna, T., Gulati, R., Nohria, N.: The dynamics of learning alliances: Competition, cooperation, and relative scope. Strategic management journal **19**(3), 193–210 (1998)

[116] Kim, J.H., Reifgerst, A., Rizzonelli, M.: Musical social entrainment. Music & Science **2**, 2059204319848991 (2019)

[117] King, D.R., Meglio, O., Gomez-Mejia, L., Bauer, F., De Massis, A.: Family business restructuring: A review and research agenda. Journal of Management Studies **59**(1), 197–235 (2022)

[118] Klein, K., Semrau, T., Albers, S., Zajac, E.J.: Multimarket coopetition: How the interplay of competition and cooperation affects entry into shared markets. Long Range Planning **53**(1), 101868 (2020)

[119] Kneser, R., Ney, H.: Improved backing-off for m-gram language modeling. In: 1995 international conference on acoustics, speech, and signal processing, vol. 1, pp. 181–184. IEEE (1995)

[120] Krasnov, F., Sen, A.: The number of topics optimization: Clustering approach. Machine Learning and Knowledge Extraction **1**(1), 416–426 (2019)

[121] Kraus, M., Feuerriegel, S.: Decision support from financial disclosures with deep neural networks and transfer learning. Decision Support Systems **104**, 38–48 (2017)

[122] Kraus, S., Breier, M., Dasí-Rodríguez, S.: The art of crafting a systematic literature review in entrepreneurship research. International Entrepreneurship and Management Journal **16**(3), 1023–1042 (2020)

[123] Kraus, S., Mahto, R.V., Walsh, S.T.: The importance of literature reviews in small business and entrepreneurship research (2021)

[124] Lau, J.H., Newman, D., Baldwin, T.: Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 530–539 (2014)

[125] Le Roy, F., Czakon, W.: Managing coopetition: the missing link between strategy and performance. Industrial Marketing Management **53**(1), 3–6 (2016)

[126] Lewis, M.W.: Exploring paradox: Toward a more comprehensive guide. Academy of Management review **25**(4), 760–776 (2000)

[127] Li, F.: Annual report readability, current earnings, and earnings persistence. Journal of Accounting and economics **45**(2-3), 221–247 (2008)

[128] Li, F., et al.: Textual analysis of corporate disclosures: A survey of the literature. Journal of accounting literature **29**(1), 143–165 (2010)

[129] Li, J.: Prediction of corporate bankruptcy from 2008 through 2011. Journal of Accounting and Finance **12**(1), 31–41 (2012)

[130] Li, W., McCallum, A.: Pachinko allocation: Dag-structured mixture models of topic correlations. In: Proceedings of the 23rd international conference on Machine learning, pp. 577–584 (2006)

[131] Liddy, E.D.: Natural language processing (2001)

[132] Lo Verde, F.M., Tumminello, M., Ciziceno, M.: Household expenditures and the status of children: An analysis of the italian case. Social Policies **9**(1), 61–88 (2022)

[133] Lopatta, K., Gloger, M.A., Jaeschke, R.: Can language predict bankruptcy? the explanatory power of tone in 10-k filings. Accounting Perspectives **16**(4), 315–343 (2017)

[134] Lopez-Lira, A.: Risk factors that matter: Textual analysis of risk disclosures for the cross-section of returns. Jacobs Levy Equity Management Center for Quantitative Financial Research Paper (2020)

[135] Loughran, T., McDonald, B.: When is a liability not a liability? textual analysis, dictionaries, and 10-ks. The Journal of finance **66**(1), 35–65 (2011)

[136] Loughran, T., McDonald, B.: The use of word lists in textual analysis. Journal of Behavioral Finance **16**(1), 1–11 (2015)

[137] Lundgren-Henriksson, E.L., Tidström, A.: Temporal distancing and integrating: Exploring coopetition tensions through managerial sensemaking dynamics. Scandinavian Journal of Management **37**(3), 101168 (2021)

[138] Luo, Y.: A coopetition perspective of global competition. Journal of world business **42**(2), 129–144 (2007)

[139] Mai, F., Tian, S., Lee, C., Ma, L.: Deep learning models for bankruptcy prediction using textual disclosures. European journal of operational research **274**(2), 743–758 (2019)

[140] Makadok, R., Burton, R., Barney, J.: A practical guide for making theory contributions in strategic management (2018)

[141] Malo, P., Sinha, A., Korhonen, P., Wallenius, J., Takala, P.: Good debt or bad debt: Detecting semantic orientations in economic texts. Journal of the Association for Information Science and Technology **65**(4), 782–796 (2014)

[142] Manning, C., Schutze, H.: Foundations of statistical natural language processing. MIT press (1999)

[143] Mariani, M.M.: Coopetition as an emergent strategy: Empirical evidence from an italian consortium of opera houses. International Studies of Management & Organization **37**(2), 97–126 (2007)

[144] Marshall, I.J., Wallace, B.C.: Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Systematic reviews **8**(1), 1–10 (2019)

[145] Matin, R., Hansen, C., Hansen, C., Mølgaard, P.: Predicting distresses using deep learning of text segments in annual reports. Expert Systems with Applications **132**, 199–208 (2019)

[146] Maula, M., Stam, W.: Enhancing rigor in quantitative entrepreneurship research (2020)

[147] Mayew, W.J., Sethuraman, M., Venkatachalam, M.: Md&a disclosure and the firm's ability to continue as a going concern. The Accounting Review **90**(4), 1621–1651 (2015)

[148] McGrath, J.E., Kelly, J.R.: Time and human interaction: Toward a social psychology of time. Guilford Press (1986)

[149] McGregor, A., Hall, M., Lorier, P., Brunskill, J.: Flow clustering using machine learning techniques. In: International workshop on passive and active network measurement, pp. 205–214. Springer (2004)

[150] M'Chirgui, Z.: The economics of the smart card industry: towards coopetitive strategies. Economics of Innovation and New Technology **14**(6), 455–477 (2005)

[151] Merkl-Davies, D.M.: Impression management, myth creation and fabrication in private social and environmental reporting: Insights from erving goffman. Social and Environmental Accountability Journal **34**(2), 126–126 (2014)

[152] Merkl-Davies, D.M., Brennan, N.M.: Discretionary disclosure strategies in corporate narratives: incremental information or impression management? Journal of accounting literature **27**, 116–196 (2007)

[153] Merkl-Davies, D.M., Brennan, N.M.: A conceptual framework of impression management: new insights from psychology, sociology and critical perspectives. Accounting and business research **41**(5), 415–437 (2011)

[154] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems **26** (2013)

[155] Miller, J.: Rg (1981): Simultaneous statistical inference

[156] Mimno, D., Wallach, H., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 262–272 (2011)

[157] Morstatter, F., Liu, H.: In search of coherence and consensus: Measuring the interpretability of statistical topics. Journal of Machine Learning Research **18**(169), 1–32 (2018)

[158] Newman, D., Karimi, S., Cavedon, L.: External evaluation of topic models. In: in Australasian Doc. Comp. Symp., 2009. Citeseer (2009)

[159] Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics, pp. 100–108 (2010)

[160] Nikolenko, S.I., Koltcov, S., Koltsova, O.: Topic modelling for qualitative studies. Journal of Information Science **43**(1), 88–102 (2017)

[161] Omar, M., On, B.W., Lee, I., Choi, G.S.: Lda topics: Representation and evaluation. Journal of Information Science **41**(5), 662–675 (2015)

[162] O'Dochartaigh, A.: No more fairytales: A quest for alternative narratives of sustainable business. Accounting, Auditing & Accountability Journal (2019)

[163] Paisley, J., Wang, C., Blei, D.M., Jordan, M.I.: Nested hierarchical dirichlet processes. IEEE transactions on pattern analysis and machine intelligence **37**(2), 256–270 (2014)

[164] Paranyushkin, D.: Identifying the pathways for meaning circulation using text network analysis. Nodus Labs **26** (2011)

[165] Paydas Turan, C.: Success drivers of co-branding: A meta-analysis. International Journal of Consumer Studies **45**(4), 911–936 (2021)

[166] Peixoto, T.P.: Nonparametric bayesian inference of the microcanonical stochastic block model. Physical Review E **95**(1), 012317 (2017)

[167] Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543 (2014)

[168] Petticrew, M., Egan, M., Thomson, H., Hamilton, V., Kunkler, R., Roberts, H.: Publication bias in qualitative research: what becomes of qualitative research presented at conferences? Journal of Epidemiology & Community Health **62**(6), 552–554 (2008)

[169] Petticrew, M., Roberts, H.: Systematic reviews in the social sciences: A practical guide. John Wiley & Sons (2008)

[170] Phillips, N., Lawrence, T.B., Hardy, C.: Discourse and institutions. Academy of management review **29**(4), 635–652 (2004)

[171] Pinello, C., Picone, P.M., Destri, A.M.L.: Co-branding research: where we are and where we could go from here. European Journal of Marketing (2022)

[172] Plaia, A., Buscemi, S., Sciandra, M.: Consensus measures among preference rankings: a new weighted correlation coefficient for linear and weak orderings. Submitted (2020)

[173] Pomerol, J.C.: Artificial intelligence and human decision making. European Journal of Operational Research **99**(1), 3–25 (1997)

[174] Porciello, J., Ivanina, M., Islam, M., Einarson, S., Hirsh, H.: Accelerating evidence-informed decision-making for the sustainable development goals using machine learning. Nature Machine Intelligence **2**(10), 559–565 (2020)

[175] Pornel, J.B., Saldaña, G.A.: Four common misuses of the likert scale. Philippine Journal of Social Sciences and Humanities University of the Philippines Visayas **18**(2), 12–19 (2013)

[176] Puccio, E., Vassallo, P., Piilo, J., Tumminello, M.: Covariance and correlation estimators in bipartite complex systems with a double heterogeneity. Journal of Statistical Mechanics: Theory and Experiment **2019**(5), 053404 (2019)

[177] Pussegoda, K., Turner, L., Garritty, C., Mayhew, A., Skidmore, B., Stevens, A., Boutron, I., Sarkis-Onofre, R., Bjerre, L.M., Hróbjartsson, A., et al.: Systematic

review adherence to methodological or reporting quality. Systematic reviews **6**(1), 1–14 (2017)

[178] Quiniou, S., Cellier, P., Charnois, T., Legallois, D.: What about sequential data mining techniques to identify linguistic patterns for stylistics? In: International Conference on Intelligent Text Processing and Computational Linguistics, pp. 166–177. Springer (2012)

[179] Quispe, L.V., Tohalino, J.A., Amancio, D.R.: Using virtual edges to improve the discriminability of co-occurrence text networks. Physica A: Statistical Mechanics and its Applications **562**, 125344 (2021)

[180] Rabetino, R., Kohtamäki, M., Federico, J.S.: A (re) view of the philosophical foundations of strategic management. International Journal of Management Reviews **23**(2), 151–190 (2021)

[181] Raddats, C., Kowalkowski, C., Benedettini, O., Burton, J., Gebauer, H.: Servitization: A contemporary thematic review of four major research streams. Industrial Marketing Management **83**, 207–223 (2019)

[182] Ramrakhiyani, N., Pawar, S., Hingmire, S., Palshikar, G.: Measuring topic coherence through optimal word buckets. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pp. 437–442 (2017)

[183] Raza-Ullah, T., Kostis, A.: Do trust and distrust in coopetition matter to performance? European Management Journal **38**(3), 367–376 (2020)

[184] Real, R., Vargas, J.M.: The probabilistic basis of jaccard's index of similarity. Systematic biology **45**(3), 380–385 (1996)

[185] Rees, A., Hardy, G.E., Barkham, M.: Covariance in the measurement of depression/anxiety and three cluster c personality disorders (avoidant, dependent, obsessive-compulsive). Journal of affective disorders **45**(3), 143–153 (1997)

[186] Riquelme-Medina, M., Stevenson, M., Barrales-Molina, V., Llorens-Montes, F.J.: Coopetition in business ecosystems: The key role of absorptive capacity and supply chain agility. Journal of Business Research **146**, 464–476 (2022)

[187] Ritala, P.: Coopetition strategy–when is it successful? empirical evidence on innovation and market performance. British Journal of management **23**(3), 307–324 (2012)

[188] Robledo, S., Grisales Aguirre, A.M., Hughes, M., Eggers, F.: "hasta la vista, baby"– will machine learning terminate human literature reviews in entrepreneurship? Journal of Small Business Management pp. 1–30 (2021)

[189] Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the eighth ACM international conference on Web search and data mining, pp. 399–408 (2015)

[190] Rousseau, D.M., Manning, J., Denyer, D.: 11 evidence in management and organizational science: assembling the field's full weight of scientific knowledge through syntheses. Academy of Management Annals **2**(1), 475–515 (2008)

[191] Rudman, J.: The state of authorship attribution studies: Some problems and solutions. Computers and the Humanities **31**(4), 351–365 (1997)

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

[193] Sbalchiero, S., Eder, M.: Topic modeling, long texts and the best number of topics. some problems and solutions. Quality & Quantity pp. 1–14 (2020)

[194] Schmidt, C.O., Kohlmann, T.: When to use the odds ratio or the relative risk? International journal of public health **53**(3), 165 (2008)

[195] Schönbucher, P.J.: Credit derivatives pricing models: models, pricing and implementation. John Wiley & Sons (2003)

[196] Schütze, H., Manning, C.D., Raghavan, P.: Introduction to information retrieval, vol. 39. Cambridge University Press Cambridge (2008)

[197] Shirata, C.Y., Takeuchi, H., Ogino, S., Watanabe, H.: Extracting key phrases as predictors of corporate bankruptcy: Empirical analysis of annual reports by text mining. Journal of emerging technologies in accounting **8**(1), 31–44 (2011)

[198] Shumway, T.: Forecasting bankruptcy more accurately: A simple hazard model. The journal of business **74**(1), 101–124 (2001)

[199] Simsek, Z., Fox, B., Heavey, C.: Systematicity in organizational research literature reviews: A framework and assessment. Organizational Research Methods p. 10944281211008652 (2021)

[200] Smith, W.K., Lewis, M.W.: Toward a theory of paradox: A dynamic equilibrium model of organizing. Academy of management Review **36**(2), 381–403 (2011)

[201] Snyder, H.: Literature review as a research methodology: An overview and guidelines. Journal of business research **104**, 333–339 (2019)

[202] Sokal, R.R., Sneath, P.H.A., et al.: Principles of numerical taxonomy. Principles of numerical taxonomy. (1963)

[203] Srivastava, A., Sutton, C.: Autoencoding variational inference for topic models. arXiv preprint arXiv:1703.01488 (2017)

[204] Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology **60**(3), 538–556 (2009)

[205] Stevenson, M., Mues, C., Bravo, C.: The value of text for small business default prediction: A deep learning approach. European Journal of Operational Research **295**(2), 758–771 (2021)

[206] Strese, S., Meuer, M.W., Flatten, T.C., Brettel, M.: Examining cross-functional coopetition as a driver of organizational ambidexterity. Industrial Marketing Management **57**, 40–52 (2016)

[207] Sutskever, I., Martens, J., Hinton, G.E.: Generating text with recurrent neural networks. In: ICML (2011)

[208] Taherdoost, H.: What is the best response scale for survey and questionnaire design; review of different lengths of rating scale/attitude scale/likert scale. Hamed Taherdoost pp. 1–10 (2019)

[209] Tennyson, B.M., Ingram, R.W., Dugan, M.T.: Assessing the information content of narrative disclosures in explaining bankruptcy. Journal of Business Finance & Accounting **17**(3), 391–410 (1990)

[210] Thiessen, E.D., Kronstein, A.T., Hufnagle, D.G.: The extraction and integration framework: a two-process account of statistical learning. Psychological bulletin **139**(4), 792 (2013)

[211] Tian, S., Yu, Y., Guo, H.: Variable selection and corporate bankruptcy forecasts. Journal of Banking & Finance **52**, 89–100 (2015)

[212] Tibshirani, R.J., Efron, B.: An introduction to the bootstrap. Monographs on statistics and applied probability **57**, 1–436 (1993)

[213] Tregidga, H., Milne, M.J., Kearins, K.: Ramping up resistance: Corporate sustainable development and academic research. Business & Society **57**(2), 292–334 (2018)

[214] Tumminello, M., Micciche, S., Lillo, F., Piilo, J., Mantegna, R.N.: Statistically validated networks in bipartite complex systems. PloS one **6**(3), e17994 (2011)

[215] Van Cuilenburg, J.J., Kleinnijenhuis, J., De Ridder, J.A.: Artificial intelligence and content analysis. Quality and Quantity **22**(1), 65–97 (1988)

[216] Waldherr, A., Heyer, G., Jähnichen, P., Niekler, A., Wiedemann, G.: Mining big data with computational methods. In: Political Communication in the Online World, pp. 201–217. Routledge (2015)

[217] Wallach, H.M., Murray, I., Salakhutdinov, R., Mimno, D.: Evaluation methods for topic models. In: Proceedings of the 26th annual international conference on machine learning, pp. 1105–1112 (2009)

[218] Wang, H., Liu, X.: Undersampling bankruptcy prediction: Taiwan bankruptcy data. Plos one **16**(7), e0254030 (2021)

[219] Wang, L., Wei, B., Yuan, J.: Topic discovery based on lda_col model and topic significance re-ranking. JCP **6**(8), 1639–1647 (2011)

[220] Wang, S.C., Soesilo, P.K., Zhang, D.: Impact of luxury brand retailer co-branding strategy on potential customers: A cross-cultural study. Journal of International Consumer Marketing **27**(3), 237–252 (2015)

[221] Watanabe, W.M., Felizardo, K.R., Candido Jr, A., de Souza, É.F., de Campos Neto, J.E., Vijaykumar, N.L.: Reducing efforts of software engineering systematic litera-

ture reviews updates using text classification. Information and Software Technology **128**, 106395 (2020)

[222] Webster, J., Watson, R.T.: Analyzing the past to prepare for the future: Writing a literature review. MIS quarterly pp. xiii–xxiii (2002)

[223] Wiener, M., Saunders, C.: Forced coopetition in it multi-sourcing. The Journal of Strategic Information Systems **23**(3), 210–225 (2014)

[224] Wilhelm, M., Sydow, J.: Managing coopetition in supplier networks–a paradox perspective. Journal of Supply Chain Management **54**(3), 22–41 (2018)

[225] Wu, W., Xiong, H., Shekhar, S.: Clustering and information retrieval, vol. 11. Springer Science & Business Media (2003)

[226] Wujec, M.: Analysis of the financial information contained in the texts of current reports: A deep learning approach. Journal of Risk and Financial Management **14**(12), 582 (2021)

[227] Xing, L., Paul, M.J., Carenini, G.: Evaluating topic quality with posterior variability. arXiv preprint arXiv:1909.03524 (2019)

[228] Yan, X., Guo, J., Lan, Y., Cheng, X.: A biterm topic model for short texts. In: Proceedings of the 22nd international conference on World Wide Web, pp. 1445–1456 (2013)

[229] Yang, F., Dolar, B., Mo, L.: Textual analysis of corporate annual disclosures: a comparison between bankrupt and non-bankrupt companies. Journal of Emerging Technologies in Accounting **15**(1), 45–55 (2018)

[230] Yao, C., Duan, Z., Baruch, Y.: Time, space, confucianism and careers: a contextualized review of careers research in china–current knowledge and future research agenda. International Journal of Management Reviews **22**(3), 222–248 (2020)

[231] Zhang, X., Zhao, J., LeCun, Y.: Character-level convolutional networks for text classification. Advances in neural information processing systems **28** (2015)

[232] Zhao, W., Chen, J.J., Perkins, R., Liu, Z., Ge, W., Ding, Y., Zou, W.: A heuristic approach to determine an appropriate number of topics in topic modeling. In: BMC bioinformatics, vol. 16, pp. 1–10. Springer (2015)

[233] Zhou, G.: Measuring investor sentiment. Annual Review of Financial Economics **10**, 239–259 (2018)

[234] Zhou, T., Kuscsik, Z., Liu, J.G., Medo, M., Wakeling, J.R., Zhang, Y.C.: Solving the apparent diversity-accuracy dilemma of recommender systems. Proceedings of the National Academy of Sciences **107**(10), 4511–4515 (2010)

[235] Zuo, Y., Zhao, J., Xu, K.: Word network topic model: a simple but general solution for short and imbalanced texts. Knowledge and Information Systems **48**(2), 379–398 (2016)

# CRediT Author Statement

- **Chapter 1** - ***Andrea Simonetti:*** conceptualization, software, formal analysis, investigation, data curation, writing - original draft, visualization. ***Rodolfo Damiano:*** conceptualization, software, formal analysis, investigation, data curation, writing - original draft, visualization. ***Michele Tumminello:*** methodology, software, formal analysis, writing - review and editing, supervision.

- **Chapter 2** - ***Andrea Simonetti:*** conceptualization, methodology, software, formal analysis, investigation, data curation, writing - original draft, visualization. ***Alessandro Albano:*** conceptualization, methodology, software, formal analysis, investigation, data curation, writing - original draft, visualization. ***Antonella Plaia:*** conceptualization, formal analysis, supervision. ***Michele Tumminello:*** conceptualization, methodology, formal analysis, writing - review and editing, supervision.

- **Chapter 3** ***Andrea Simonetti:*** conceptualization, methodology, software, formal analysis, investigation, data curation, writing - original draft, visualization.

  ***Pasquale Massimo Picone:*** conceptualization, methodology, investigation, data curation, writing - original draft, visualization. ***Anna Minà:*** conceptualization, investigation, data curation, writing - original draft, visualization. ***Michele Tumminello:*** conceptualization, methodology, software, formal analysis, investigation, data curation, writing - original draft, visualization.

- **Chapter 4** - ***Michele Tumminello:*** conceptualization, methodology, software, formal analysis, investigation, data curation, writing - original draft, visualization. ***Andrea Simonetti:*** methodology, software, formal analysis, data curation, writing - original draft. ***Tiziana Di Matteo:*** methodology, supervision.

# Outputs of the PhD research

During the PhD program I co-authored three papers, each one corresponding to a chapter of the thesis. Specifically:

i) Simonetti, A., Damiano, R. and Tumminello, M (2022). An NLP method to predict bankruptcy from the analysis of annual corporate reports. Journal of Information Science. **Accepted**

ii) Simonetti, A., Albano, A., Plaia, A. and Tumminello, M (2022). Ranking coherence in Topic Models using Statistically Validated Networks. **To be published**

iii) Simonetti, A., Picone, P. M., Miná, A. and Tumminello, M (2022). Networks and Text Mining approach to perform systematic literature reviews. **To be published**

iv) Tumminello, M., Simonetti, A., Di Matteo, T. (2022) Detecting and modeling excess of attribute similarity in small groups-with applications to Text Mining and Social Sciences. **To be published**

Moreover, I co-authored other publications:

i) Simonetti, A., Albano, A., Plaia, A., and Tumminello, M. (2022). Statistically Validated Networks for evaluating coherence in topic models, Book of Abstracts, In The 10th International Conference on Complex Networks and their Applications. ISBN: 978-2-9557050-5-6.

ii) Albano, A. and Simonetti, A. (2022). Statistically Validated Networks for assessing topic quality in LDA models, Book of Abstracts, In The 5th European Conference on Social Network EUSN 2021. ISBN: 978-88-90109-13-3.

iii) Simonetti A, D'Angelo, N. and Adelfio, G. (2022) Marked Hawkes processes for Twitter data. 16th International Conference on Statistical Analysis of Textual Data. ISBN: 979-12-80153-30-2

During the three years of the Ph.D. program, I participated in the following conferences where I presented some of the content of this thesis:

- Conference of Complex Systems - 17-21/10/2022, Virtual (Palma de Mallorca). **Contributed Talk**
- Associazione per la Matematica Applicata alle Scienze Economiche e Sociali (A.M.A.S.E.S.), Palermo, 22-24/10/2022. **Contributed Talk**
- 16th International Conference on Statistical Analysis of Textual Data - 06-08/07/2022, Virtual(Naples). **Contributed Talk**
- 10th International Conference on Complex Networks and their Applications - 30/11-02/12 /2021, Virtual(Madrid). **Contributed Talk**