



---

**UNIVERSITÀ  
DEGLI STUDI  
DI BRESCIA**

**DOTTORATO DI RICERCA IN  
MODELLI E METODI PER L'ECONOMIA E IL MANAGEMENT  
(ANALYTICS FOR ECONOMICS AND MANAGEMENT)**

---

SETTORE SCIENTIFICO DISCIPLINARE  
SECS-S/01 - STATISTICA  
CICLO XXXVI

---

# **Statistical Analysis of Grouped Text Documents**

*Author:*  
Riccardo RICCIARDI

*Supervisor:*  
Professor Marica MANISERA  
*Tutor:*  
Professor Maurizio CARPITA

December 23, 2023

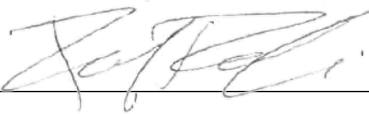


## Declaration of Authorship

I, Riccardo RICCIARDI, declare that this thesis titled, "Statistical Analysis of Grouped Text Documents" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Signed:



Date: 23/12/2023



# Abstract

## Statistical analysis of grouped text documents

Riccardo Ricciardi

The topic of this thesis is statistical models for the analysis of textual data, emphasizing contexts in which text samples are grouped.

When dealing with text data, the first issue is to process it, making it computationally and methodologically compatible with the existing mathematical and statistical methods produced and continually developed by the scientific community. Therefore, the thesis firstly reviews existing methods for analytically representing and processing textual datasets, including Vector Space Models, distributed representations of words and documents, and contextualized embeddings. It realizes this review by standardizing a notation that, even within the same representation approach, appears highly heterogeneous in the literature.

Then, two domains of application are explored: social media and cultural tourism. About the former, a study is proposed about self-presentation among diverse groups of individuals on the StockTwits platform, where finance and stock markets are the dominant topics. The methodology proposed integrated various types of data, including textual and categorical data. This study revealed insights into how people present themselves online and found recurring patterns within groups of users.

About the latter, the thesis delves into a study conducted as part of the "Data Science for Brescia - Arts and Cultural Places" Project, where a language model was trained to classify Italian-written online reviews into four distinct semantic areas related to cultural attractions in the Italian city of Brescia. The model proposed allows for the identification of attractions in text documents, even when not explicitly mentioned in document metadata, thus opening possibilities for expanding the database related to these cultural attractions with new sources, such as social media platforms, forums, and other online spaces.

Lastly, the thesis presents a methodological study examining the group-specificity of words, analyzing various group-specificity estimators proposed in the literature. The study considered grouped text documents with both outcome and group variables. Its contribution consists of the proposal of modeling the corpus of documents as a multivariate distribution, enabling the simulation of corpora of text documents with predefined characteristics. The simulation provided valuable insights into the relationship between groups of documents and words. Furthermore, all its results can be freely explored through a web application, whose components are also described in this manuscript.

In conclusion, this thesis has been conceived as a collection of papers. It aimed to contribute to the field with both applications and methodological proposals, and each study presented here suggests paths for future research to address the challenges in the analysis of grouped textual data.



# Abstract (Italian)

## Analisi statistica di documenti di testo raggruppati

Riccardo Ricciardi

L'argomento di questa tesi sono i modelli statistici per l'analisi dei dati testuali, con particolare attenzione ai contesti in cui i campioni di dati sono raggruppati.

Quando si ha a che fare con dati testuali, il primo problema è quello di elaborarli, rendendoli compatibili dal punto di vista computazionale e metodologico con i metodi matematici e statistici esistenti prodotti e continuamente sviluppati dalla comunità scientifica. Per questo motivo, la tesi passa in rassegna i metodi esistenti per la rappresentazione analitica e l'elaborazione di insiemi di dati testuali, compresi i Vector Space Model, le rappresentazioni distribuite di parole e documenti e i contextualized embeddings. Questa rassegna si concretizza nella standardizzazione di una notazione che, anche nell'ambito dello stesso approccio di rappresentazione, appare molto eterogenea in letteratura.

Vengono poi esplorati due domini di applicazione: social media e turismo culturale. Riguardo al primo, viene proposto uno studio sull'autodescrizione di gruppi diversi di individui sulla piattaforma StockTwits, dove la finanza e i mercati azionari sono gli argomenti dominanti. La metodologia proposta ha integrato diversi tipi di dati, compresi quelli testuali e variabili categoriche. Questo studio ha rivelato intuizioni sul modo in cui le persone si presentano online e trovato comportamenti comuni all'interno di gruppi di utenti.

Per quanto riguarda il secondo dominio, si approfondisce uno studio nell'ambito del progetto "Data Science for Brescia - Arts and Cultural Places". Un modello linguistico è stato addestrato per classificare le recensioni online scritte in italiano in quattro aree semantiche distinte relative alle attrazioni culturali della città italiana di Brescia. Questo modello permette di identificare le attrazioni nei documenti di testo, anche quando non sono esplicitamente menzionate nei metadati, aprendo così la possibilità di espandere il database relativo a queste attrazioni culturali con nuove fonti, come piattaforme di social media, forum e altri spazi online.

Infine, uno studio metodologico esamina la specificità di gruppo delle parole, analizzando diversi stimatori proposti in letteratura. Sono presi in considerazione documenti testuali raggruppati con variabili di outcome e di gruppo. Il suo contributo consiste nella proposta di modellare il corpus di documenti come una distribuzione multivariata, consentendo la simulazione di corpora di documenti di testo con caratteristiche predefinite. La simulazione ha fornito preziose indicazioni sulla relazione tra gruppi di documenti e parole. Inoltre, tutti i risultati possono essere liberamente esplorati attraverso un'applicazione web, i cui componenti sono descritti anche in questo manoscritto.

In conclusione, questa tesi è stata concepita come una raccolta di lavori scientifici. Ogni studio presentato suggerisce percorsi di ricerca futuri per affrontare le sfide dell'analisi dei dati testuali raggruppati.





## Acknowledgements

Qualcuno dice che, come le parole nel linguaggio, una persona va conosciuta per la sua compagnia. Sono concettualmente vicino a questo pensiero, ma credo sia difficile generalizzare la cosa, perchè c'è chi riesce a farcela da sé, chi ha grandi compagnie, chi, pur in grandi compagnie, risente dell'influenza di pochissime persone,... Più facile elencare qui le persone e i momenti che vengono in mente a me, che senza quelle e quelli non avrei concluso molto.

Mia madre che, in quanto tale, ma soprattutto in quanto Carla, ha la rara capacità di ascoltare, e fare suoi i miei problemi.

Le divagazioni con il Prof, mio padre, solitamente tra le 19.00 e le 20.00. Chissà quante sciocchezze abbiamo detto, spacciandoci per intellettuali.

Chiara ed Eraldo. In qualche modo, non ben definito, ci aiutiamo sempre.

Margherita.

Sauro ed Anna, per essere stati una seconda famiglia.

Roberto, per quanto è stato determinante negli ultimi cinque anni della mia vita, nonostante sia convinto che quella piemontese sia una delle due cucine regionali più buone d'Italia. (Robbè, fa' un giro qui: <https://www.tasteatlas.com/italy>).

I ragazzi del CRAL Volley che, sotto la guida del Leader Michele, hanno evitato che il *Monday* fosse *blue*.

I primi due anni di didattica con Maurizio. Perchè le prime esperienze sono state uno strano cocktail di ansia e divertimento (senza ghiaccio).

Il collegamento più diretto con questo lavoro è senza dubbio Marica. Grazie per essere l'arma più forte contro la sindrome dell'impostore. Spero di avere occasione di imparare ancora da te come gestire le cose e le persone.

[www.youtube.com/watch?v=P1XsuaRhogA&ab\\_channel=ChoYoung-Wuk-Topic](http://www.youtube.com/watch?v=P1XsuaRhogA&ab_channel=ChoYoung-Wuk-Topic)



# Contents

<b>Declaration of Authorship</b>	<b>iii</b>
<b>Abstract</b>	<b>v</b>
<b>Abstract (Italian)</b>	<b>vii</b>
<b>Acknowledgements</b>	<b>ix</b>
<b>Introduction</b>	<b>1</b>
<b>1 Literature Review</b>	<b>5</b>
1.1 Vector Space Models . . . . .	5
1.1.1 Pre-processing and other data structures . . . . .	7
1.1.2 Representations with document variables . . . . .	7
1.2 Distributed representations of words . . . . .	9
1.2.1 Word and document embedding methodologies . . . . .	11
1.2.2 Embeddings with document variables . . . . .	12
1.3 Contextualized embeddings . . . . .	12
1.4 Research production . . . . .	14
<b>2 Users' self-description on social media</b>	<b>17</b>
2.1 Introduction . . . . .	17
2.1.1 Self-describing on social media . . . . .	17
2.2 Related works . . . . .	19
2.3 Methods and data . . . . .	20
2.3.1 Methods . . . . .	20
2.3.2 Data . . . . .	23
2.4 Results . . . . .	26
2.4.1 Seed words . . . . .	26
2.4.2 Neighbourhood exploration . . . . .	28
2.5 Conclusions . . . . .	30
2.6 Discussion . . . . .	32
<b>3 Visitors' reviews of cultural attractions</b>	<b>39</b>
3.1 The DS4BS Project . . . . .	39
3.1.1 Investigating visitors' online opinions . . . . .	40
3.2 Related works . . . . .	41
3.3 Methods . . . . .	43
3.3.1 The Language Model . . . . .	43
3.3.2 Extracting group-specific keywords . . . . .	46
3.4 Data . . . . .	47
3.5 Results . . . . .	48
3.5.1 The model performance . . . . .	48
3.5.2 Vector representations of reviews . . . . .	48

3.5.3	Cluster-specific keywords . . . . .	50
3.6	Conclusions and discussion . . . . .	53
<b>4</b>	<b>Group-specific term estimators: a simulation study</b>	<b>55</b>
4.1	Introduction . . . . .	55
4.2	Methods . . . . .	56
4.2.1	Multivariate Bernoulli sampling with Gaussian copula . . . . .	57
4.2.2	Parameters of the simulation . . . . .	58
4.2.3	Analytical formulation of GS measures . . . . .	59
4.3	Study design and research questions . . . . .	62
4.4	Results . . . . .	64
4.4.1	General tendency . . . . .	64
4.4.2	Two terms with different group specificity . . . . .	65
4.4.3	Role of group unbalancing . . . . .	66
4.4.4	Role of term rarity . . . . .	67
4.4.5	Outcome-term vs group-term relationships . . . . .	68
4.5	The web application . . . . .	71
4.6	Conclusions and discussion . . . . .	72
	<b>Conclusion</b>	<b>75</b>
	<b>Bibliography</b>	<b>79</b>

# List of Figures

1.1	Vector representations of 6 words projected into the 2-dimensional space defined by the first two principal components. . . . .	10
1.2	Time evolution of the number of research products published from 1985 to 2022, categorized by the approach to corpus representation (dashed lines); in the second plot, subthemes of research are represented by solid and colored lines. . . . .	16
2.1	Summary of the proposed methodology. . . . .	24
2.2	Scatterplots of IG and absolute value of modRF by trading variables for tokens between the 5-th and 10-th percentile of IDF; based on the sign of modRF, the colors indicate which category the term is specific to. . . . .	27
2.3	90% bootstrap confidence intervals (5-th and 95-th percentiles) for the proportion of fundamental traders, medium/long-term investors, and professional traders, in the neighbourhood of seed word vectors; neighbourhood composed of the 10% most similar users. . . . .	31
3.1	Mobile phone and Mastercard data. Left: Points of Interest (POIs) of type "squares". Right: Map of the 16 commercial systems constituting the Distretto Urbano del Commercio (DUC) (red) and overlapping or neighboring POIs (blue). . . . .	40
3.2	Input and Embeddings for the Transformer Encoder in the BERT Representation. . . . .	44
3.3	High-Level Architecture for the MLM Objective of the Pre-trained BERT. . . . .	45
3.4	BERT High-Level Architecture for Fine-Tuning on Text Classification Tasks. . . . .	45
3.5	Performance metrics (averaged over the categories of the target variable) on validation set by epochs (model trained on unbalanced set). . . . .	49
3.6	First two principal components of embeddings of training reviews with no fine-tuning (left) and after fine-tuning (right); percentage of variance explained by PCs in parenthesis on the axes. . . . .	50
3.7	First two principal components of embeddings of validation reviews (left) and assessment reviews (right), after fine-tuning; PCs computed on embeddings of training set (variance explained by PCs in parenthesis on the axes). . . . .	51
3.8	Two-dimensional representation of clusters of embeddings of validation reviews obtained with HDBSCAN before (left) and after noise removal (right); clustering algorithm applied on 10 PCs; PCs computed on embeddings of training set (variance explained by PCs in parenthesis on the axes). . . . .	51
3.9	Top 10 by-cluster ranking of nouns based on the value of the Keynes K (English translations of nouns in parenthesis). . . . .	52

4.1	<b>Illustration of the Gaussian Copula Method to draw 2000 samples from a Multivariate Bernoulli distribution, with <math>(\rho_{X_Y, X_G}; \rho_{X_Y, X_T}; \rho_{X_G, X_T}) = (0.9; 0.6; 0.25)</math>, and <math>(u_Y^*, u_G^*, u_T^*) = (p_Y; p_G; p_T) = (0.7; 0.5; 0.2)</math>.</b> . . . . .	59
4.2	<b>Mean values (in red) and boxplots of the estimates, across multiple subscenarios; scenario parameter values: <math>n = 2000, p_G = 0.5, p_T = 0.1</math>. <i>corr</i>(<math>X_G, X_T</math>) stands for <math>\rho_{X_G, X_T}</math>.</b> . . . . .	65
4.3	<b>Mean values (in red) and boxplots of the estimates of PMI for positive values of correlations, across multiple subscenarios; scenario parameter values: <math>n = 2000, p_G = 0.5, p_T = 0.1</math>.</b> . . . . .	66
4.4	<b>Mean values of the estimates under the scenarios defined by <math>n = 2000, p_T = 0.1</math> and three values of <math>p_G</math>, across multiple subscenarios.</b> . . . . .	67
4.5	<b>Mean values of the estimates under the scenarios defined by <math>n = 2000, p_G = 0.5</math> and three values of <math>p_T</math>, across multiple subscenarios.</b> . . . . .	68
4.6	<b>Mean values of the estimates of PMI for positive values of the correlations under the scenarios defined by <math>n = 2000, p_T = 0.1</math> and three values of <math>p_G</math> (left) and scenarios defined by: <math>n = 2000, p_G = 0.5</math> and three values of <math>p_T</math> (right), across multiple subscenarios.</b> . . . . .	69
4.7	<b>Estimates of the ratio <math>V</math> for IDFR (left) and RFR (right), across multiple subscenarios; scenario parameter values: <math>n = 2000, p_Y = 0.5, p_G = 0.5, p_T = 0.1</math>.</b> . . . . .	70
4.8	<b>Estimates of the ratio <math>V</math> for IDFR (left) and RFR (right), across multiple subscenarios; scenario parameter values: <math>n = 2000, p_Y = 0.75, p_G = 0.25, p_T = 0.1</math>.</b> . . . . .	71

# List of Tables

1.1	Top 5 terms most similar to the representations of <i>king</i> , <i>hello</i> , and <i>pasta</i> (cosine similarities in parenthesis). . . . .	10
1.2	Examples of word analogies (cosine similarities in parenthesis). . . . .	11
1.3	Examples of English words changing sense depending on context. . . . .	13
1.4	Queries on Scopus database by topics. . . . .	15
2.1	Toy example: categorical self-labels and textual self-descriptions in a social media for scholars. . . . .	18
2.2	Description of the pre-processing, from original to final sample. . . . .	25
2.3	Proportion of professional, technical and medium/long-term traders, computed on the binary variables. . . . .	26
2.4	List of domain-specific and statistically relevant seed tokens. . . . .	28
2.5	Most similar terms for a subset of seed words; cosine similarity (in parenthesis) used as similarity measure, and similar words found among the 10% most frequent words of the vocabulary. . . . .	32
2.6	Most similar terms for seed words (group 1); cosine similarity (in parenthesis) used as similarity measure, and similar words found among the 10% most frequent words of the vocabulary. . . . .	34
2.7	Most similar terms for seed words (group 2); cosine similarity (in parenthesis) used as similarity measure, and similar words found among the 10% most frequent words of the vocabulary. . . . .	35
2.8	90% bootstrap confidence intervals (5-th and 95-th percentiles) for the proportion of fundamental traders, medium/long-term investors, and professional traders, in the neighbourhood of seed word vectors (group 1); neighbourhood composed of the 10% most similar users. . . . .	36
2.9	90% bootstrap confidence intervals (5-th and 95-th percentiles) for the proportion of fundamental traders, medium/long-term investors, and professional traders, in the neighbourhood of seed word vectors (group 2); neighbourhood composed of the 10% most similar users. . . . .	37
3.1	Distribution of attractions in the training and validation sets. . . . .	48
3.2	Performance metrics (averaged over the categories of the target variable) on validation and assessment set; best models obtained on unbalanced training set and oversampled training set. . . . .	49
4.1	Selected parameter values for simulation. . . . .	63
4.2	Correlation matrix between average GS estimates across simulations. . . . .	63





# Introduction

About three years ago, I experienced a "road-to-Damascus" moment. As a cinema enthusiast, particularly fond of sci-fi movies, I was reading reviews of the film "Interstellar" on the renowned *Internet Movie Database*. In a sense, I was qualitatively analyzing users' opinions on the film. After some reading, I noticed two phenomena. On one hand, different ratings were often associated with seemingly similar opinions. For example, one user had given a rating of 10/10 to the phrase "Out of this world," while another user had rated the phrase "Mesmerizing" 8/10. On the other hand, I found that the reviews "What a massive disappointment" and "Good, but Overrated" were both given a rating of 6/10, although they appeared to express different opinions at first glance.

These phenomena naturally give rise to other issues concerning the varying perception of ratings by different respondents. At that moment, I realized that my lack of knowledge of textual data analysis techniques would limit my ability to conduct numerous interesting investigations, resulting in a relevant loss of potentially valuable information. I felt that this information loss was comparable, if not greater, than reducing the rating scale to "positive," "neutral," and "negative" ratings.

Therefore, as clearly suggested by its title, this thesis addresses the statistical analysis of textual data, with a focus on the common case in which text samples can be grouped.

Let us proceed gradually, guided by two keywords: "text" and "grouped."

Today, it no longer takes a great effort to imagine that data analysis extends beyond structured data in a tabular format to include unstructured data such as texts, images, videos, sounds, and more. These data types have led to the emergence of entire research fields aimed at developing methods to process unstructured data, making it computationally and methodologically compatible with the existing mathematical and statistical methods produced and continually developed by the scientific community.

Therefore, I used my doctoral program to learn, or at least initiate, the use of tools for processing textual data and transforming it into meaningful information.

Textual data analysis is not a single research field in itself. The first identifiable field when dealing with text is *Natural Language Processing*, which is the study of processing data created in natural language. Depending on the specific research objectives, various subfields can be identified. To name a few, there is *Information Retrieval*, which seeks to provide the best methods for gathering information based on a user input request; *Sentiment Analysis*, which aims to extract latent opinions from a text regarding a specific topic and translate them into a negative-to-positive scale, and *Emotion Recognition*, which translates opinions into a broader spectrum of emotions; *Text Summarization*, which involves condensing a text document into a compact version that preserves all its essential information; *Content Analysis*, which, in its quantitative branch, seeks words that summarize a text and the relationships between them; . . .

The second keyword in the title of this thesis is "grouped." Documents in a corpus are usually divided into groups, such as product reviews categorized by the

type of product, tweets from politicians with different political affiliations, abstracts of scientific publications categorized by research field, job openings categorized by industry, and medical reports categorized by medical department. When deciding on a technique for representing textual data, one must consider their division into groups, as group membership may (or may not) convey important information about the data. For example, when analyzing product reviews, it is useful to know that the language used in the reviews varies depending on the consumer category. Similarly, when analyzing politicians' tweets about a specific topic and aiming to extract the most commonly used words to describe the topic, the political affiliation of the politicians provides a more comprehensive and varied picture of the vocabulary used by the political class to discuss the topic.

Thus, the structure of this thesis is described as follows.

Chapter 1 offers a review of the existing methods to analytically represent and process datasets of text documents. The review addresses the following topics: (1) Vector Space Models, (2) distributed representations of words and documents, and traditional methods to train them, and (3) contextualized embeddings with a focus on BERT-based models.

Eventually, the chapter also shows the time evolution of research products related to the various outlined approaches to text representations.

Chapter 2 addresses the exploration of self-presentations among diverse groups of individuals. Its aim is to reveal intriguing insights about the ways people choose to describe themselves, such as whether there exist recurrent, shared patterns in self-presentation among individuals within the same social or identity group, or whether certain themes are common to all groups.

From a methodological point of view, this involves the integration of different types of data: textual data, and data represented by variables with predefined categories, distributing texts into groups.

This chapter provides a methodology for this integration in the domain of self-descriptions on social media. Specifically, it broadens the literature about the social media StockTwits, less explored than other platforms, such as Facebook, Twitter, and Reddit, and developed for discussions about stock markets.

Chapter 3 presents a study within the context of the "Data Science for Brescia - Arts and Cultural Places" Project, whose primary goal is to enhance our comprehension of how individuals engage with cultural sites.

The purpose of the specific study here is to build a language model on the Italian-written online reviews to classify them into four distinct semantic *areas*, defined by the main four attractions of the city of Brescia, in Italy.

The great utility of such a model stems from the fact that it can be used to identify the attraction in text documents, such as posts on social media platforms, forums, and other online spaces, when the attraction is not explicitly mentioned in the document metadata. Therefore, the proposed model can process and classify those texts, allowing for the expansion of the online discourse database concerning those cultural attractions.

Chapter 4 refers to a methodological study about the notion of *group-specificity* of a term, since it analyzes the various *group-specificity estimators* proposed in the literature. They are methods to capture the association between a group of documents and a word. However, when employing these estimators, there may be a temptation to utilize them as a measure of the intensity of the association between a word and a group, to compare the association of one word with multiple groups, or to compare the association of multiple words with the same group.

Through this study, it has been checked whether the corpus size, group imbalance, and the probability of a word appearing in a document play a role in determining the admissibility of using estimators in these manners.

The contribution of the study is to model a corpus of documents as a sample of a multivariate distribution, allowing the definition of a simulation study. A web application is made available to freely explore all the results of the simulation.

The reader should perceive this thesis as a collection of papers, arranged in chronological order based on their presentation sequence.

With this thesis, I hope to make my contribution to such an intriguing and rapidly evolving research field.



## Chapter 1

# Literature Review

*Natural Language Processing* (NLP) is a field of study with the aim of processing data created by *natural language*, which is language naturally developed by individuals.

It has become a very active research area since the 2000s, thanks to, on the one hand, the increasing capacity of computer and information systems, and, on the other hand, the proliferation of both social media and web applications, mainly commercial [1].

Based on the specific task, there are several levels of investigation of NLP. In this review, the focus is on the sentence/document level, and specifically on the methods proposed to analytically represent and process datasets of small coherent pieces of text, where words are arranged according to logical links.

Thus, the review addresses the following topics: (1) Vector Space Models, (2) distributed representations of words and documents, and traditional methods to train them, and (3) contextualized embeddings with a focus on BERT-based models.

### 1.1 Vector Space Models

The first occurrence of the term "*Vector Space Models*" (VSM) in the literature dates back to the manuscript "*A vector space model for automatic indexing*" [2]. The term refers to a technique for processing text fragments as vectors, enabling both computers and available mathematical-statistical methods to compare them in terms of content similarity and measure how well single words convey the concepts expressed in them. Before delving into the scientific debate on these topics, it is necessary to introduce some fundamental definitions.

It's worth mentioning that the notation used in this manuscript is the result of an effort to standardize a notation that, even within the same document representation approach, appears highly heterogeneous in the literature.

A text document  $d_i$  is any piece of text composed of smaller elements interconnected by syntactic and semantic relationships. These smaller elements, referred to as *tokens*, may be either single words (or *terms*) or sequences of adjacent words. Thus, a *corpus* of documents can be defined as the set  $\mathbb{D}$  of documents  $d_i$ , i.e.  $\mathbb{D} = \{d_i\}_{i=1}^D$ , arising from combinations of tokens  $w_t$  in the *vocabulary*  $\mathbb{V} = \{w_t\}_{t=1}^V$ .

The initial formalization of the Vector Space Models, and thus the representations of documents as vectors, originates within the Information Retrieval (IR) field, driven by the need to rank documents based on their relevance to a query document. From that perspective, many authors discussed the need for an appropriate way to describe documents, introducing the concepts of *exhaustivity* of a document description and *specificity* of a term in the vocabulary: the former is the degree to which a description is able to cover the various concepts expressed by the terms occurring in a document, while the latter indicates the level of detail to which a term describes a given concept [3] [4].

The first attempts to quantify those factors with document-level and corpus-level statistics are in [5] and [6]: exhaustivity relates to counting the occurrences of a term in a document, while specificity involves counting the number of documents in which the term appears. Hence the definition of *count-based* models.

With a notation inspired by the formulation in [2] and [7], a corpus of documents can be described by means of the so-called *Document-Term Matrix* **DTM**, that is given by:

$$\mathbf{DTM} = \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \\ \vdots \\ \mathbf{h}_D \end{bmatrix} \in \mathbb{R}^{VD} \quad (1.1)$$

whose  $D$  rows are vector representations of documents; therefore,  $\mathbf{h}_i$  is a vector representation of  $d_i$ , and the generic element  $h_{it}$  of  $\mathbf{h}_i$  is given by  $h(f_{it}, c_t)$ , i.e. a function  $h(\cdot)$  of two components:  $f_{it}$ , a measure of the frequency of the term  $w_t$  in the document  $d_i$ , and  $c_t$ , a measure of the frequency of the same term  $w_t$  in the entire corpus  $\mathbb{D}$ ;  $h(\cdot)$  usually includes a *normalisation* component, to avoid biases introduced by unequal number of tokens occurred in the documents. In the literature, a specification of  $h(f_{it}, c_t)$  is commonly known as a *term-weighting scheme* [8] [9]. Thus, a document representation depends on all the  $V$  terms of the vocabulary  $\mathbb{V}$  of the corpus  $\mathbb{D}$ .

For the sake of clarity, **DTM** can be expanded as follows:

$$\mathbf{DTM} = \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1V} \\ h_{21} & h_{22} & \dots & h_{2V} \\ \dots & \dots & \dots & \dots \\ h_{D1} & h_{D2} & \dots & h_{DV} \end{bmatrix} = \begin{bmatrix} h(f_{11}, c_1) & h(f_{12}, c_2) & \dots & h(f_{1V}, c_V) \\ h(f_{21}, c_1) & h(f_{22}, c_2) & \dots & h(f_{2V}, c_V) \\ \dots & \dots & \dots & \dots \\ h(f_{D1}, c_1) & h(f_{D2}, c_2) & \dots & h(f_{DV}, c_V) \end{bmatrix} \quad (1.2)$$

One of the most commonly used term-weighting schemes is the *Term Frequency-Inverse Document Frequency* (TFIDF) [7], whose general form is given by:

$$\text{TFIDF}_{it} = f_{it} \cdot \text{IDF}_t = f_{it} \cdot \log_2\left(\frac{D}{D_t}\right) \quad (1.3)$$

where  $D$  is the total number of documents as already defined,  $D_t$  is the number of documents in which  $w_t$  occurs at least once, and thus  $D_t \geq 1$ . The frequency  $f_{it}$  is usually computed as one of the following:

- absolute frequency, i.e. count of the occurrences of  $w_t$  in  $d_i$ ;
- relative frequency, i.e. absolute frequency divided by the number of tokens in  $d_i$ ;
- boolean frequency, which equals 1 if  $w_t$  occurs in  $d_i$  at least once, and 0, otherwise.

Based on the intuition that a word is more specific to a document the rarer it is in the corpus, the IDF was proposed in [5] to lower the weight of frequent words; indeed, given  $f_{it}$ :

- when  $D_t \rightarrow D$ , i.e. as the number of documents with  $w_t$  increases, then  $\text{TFIDF}_{it} \rightarrow 0$ ;
- the smaller  $D_t$ , i.e. the rarer  $w_t$ , the higher the weight of  $w_t$  in  $d_i$ .

This term weighting scheme has been widely used and is still applied in many non-methodological studies (see for example [10] [11] [12]).

### 1.1.1 Pre-processing and other data structures

In the context of VSM, the *processing* of a corpus implies the organization of documents into vectors and, therefore, the computation of the elements of data structures like the **DTM**. However, a corpus needs to be prepared for that, i.e. to pass through the *pre-processing* phase, including several steps. The *tokenization* divides a string of text into its smaller elements, based on separation by blanks, punctuation, and/or special characters (e.g. non-ASCII characters). At this point, the vocabulary is composed of single terms, but a common practice is to expand it with the more frequent two- and/or three-word sequences, i.e. with *bigrams* and/or *trigrams*.

To reduce the size of the vocabulary, and therefore the dimensions of the **DTM**, *stemming* and *lemmatization* derive base forms from each word: the former technique only derives the root for each word, whereas the latter aims to reduce each word to its lemma, i.e. the word under which it is written in the dictionary. Moreover, to reduce  $V$ , the most common words, defined as *stop words*, such as articles and prepositions, may be removed if they are considered to add no meaningful content to the text.

Except for tokenization, those steps are strictly case- and domain-dependent. Thus, pre-processing precedes the structuring of the textual data into the matrix **DTM**. Other common data structures are the *Term-Term Matrix TTM* and the *Document-Document Matrix DDM* [13] [14]. Both are obtained from the **DTM** with  $h_{it} = f_{it}$ , where  $f_{it}$  is the boolean frequency, which here we call  $\mathbf{DTM}^{(b)}$ . **TTM** is easily computed as  $\mathbf{DTM}^{(b)} \cdot \mathbf{DTM}^{(b)}$ , yielding the matrix of co-occurrences of terms, whose generic element is the number of documents in which  $w_t$  and  $w_{t'}$  co-occur (with  $t \neq t'$ ), and the diagonal elements are the column totals of  $\mathbf{DTM}^{(b)}$ , i.e. the number of documents in which the term  $w_t$  occurs. Note that by splitting documents into smaller sequences, **TTM** can offer different levels of granularity of term co-occurrences.

The matrix product  $\mathbf{DTM}^{(b)} \cdot \mathbf{DTM}^{(b)}$  generates the **DDM**, whose generic element is the number of terms shared by the documents  $d_i$  and  $d_{i'}$  (when  $i \neq i'$ ), and the diagonal elements are the row totals of  $\mathbf{DTM}^{(b)}$ , i.e. the number of terms occurring in the document  $d_i$ .

### 1.1.2 Representations with document variables

The TFIDF and its variants proposed in [7] [15] are instances of unsupervised term-weighting schemes, wherein the weight of a term in a document is entirely independent of the group to which a document may potentially belong. However, documents in a corpus are usually divided into groups: consider product reviews categorized by type of product, abstracts of scientific publications grouped by research field, job openings categorized by industry, medical reports by medical department, .... Thus, it should be considered that a term may have different weights for different groups. Furthermore, studying the importance of words for different groups of documents can be the primary goal of analysis, such as the search for group-specific dictionaries in the field of Automatic Term Recognition [16] [17] e Content Analysis [18].

In the literature on count-based models, methodologies proposed to study the relationship between terms and individual groups of documents involve *dimensionality reduction* and *supervised* term-weighting schemes. They are primarily used in the context of text classification, i.e., when predefined categories or group memberships must be assigned to the documents.

Firstly, representing text documents as previously described has a relevant drawback. Since the document representations in **DTM** have  $V$  columns, one for each word in the vocabulary, they reside in a high-dimensional space and the matrix is very sparse, leading to the problem of the *curse of dimensionality* for both classification tasks and tasks involving document similarities. Therefore, several methods of *feature selection* have been proposed to reduce the dimensionality of **DTM**, by exploiting the group membership information stored in a *document variable* [19].

They involve the following steps:

1. For each word, measuring its relationship with the document variable;
2. Ranking the words based on that relationship;
3. Extracting a subset of words by selecting the top-ranked words.

With a different approach, [20] introduced the idea of *supervised* term-weighting schemes, i.e. utilizing group membership information directly in the construction of the term-weighting scheme and several studies embraced this approach [21] [8].

In both cases, estimators of a term's discriminatory power across categories are proposed.

For example, if we denote with  $Y$  the document variable defining the document categories  $j$  with  $j = 1, \dots, m$ , formulations of the well-known *Information Gain* (IG) and *Pointwise Mutual Information* (PMI) are shown.  $IG_t$  provides a measure of reduction in entropy of the variable  $Y$  achieved by knowing the presence or absence of the word  $w_t$  [19] [22], and it is computed for the generic term  $w_t$  and the category  $j$ , with  $j = 1, \dots, m$ , as follows:

$$\begin{aligned} IG_t = & - \sum_{j=1}^m P(Y = j) \log_2 P(Y = j) + \\ & + P(f_t = 1) \sum_{j=1}^m P(Y = j | f_t = 1) \log_2 P(Y = j | f_t = 1) + \\ & + P(f_t = 0) \sum_{j=1}^m P(Y = j | f_t = 0) \log_2 P(Y = j | f_t = 0) \end{aligned} \quad (1.4)$$

where  $f_t$  is the boolean frequency of the word  $w_t$  in the document.  $PMI_{jt}$  of the term  $w_t$  and the category  $j$  is:

$$PMI_{jt} = \log_2 \frac{P(Y = j, f_t = 1)}{P(Y = j)P(f_t = 1)} \quad (1.5)$$

Based on the idea that only the documents that include a word are crucial for checking its relevance in group separation, [23] proposed the *Modified Relevance Frequency* (modRF). The formula of  $modRF_{jt}$  of the term  $w_t$  for the category  $j$  results:

$$modRF_{jt} = \log_2 \left( 2 + \frac{\alpha + 1}{\gamma + 1} \right) \quad (1.6)$$

where:



- $\alpha$  = number of documents with  $Y = j$  and the occurrence of  $w_t$ ;
- $\gamma$  = number of documents with  $Y \neq j$  and the occurrence of  $w_t$ .

Further proposals and some modifications will be introduced in the following chapters. Specifically, Chapter 4 will illustrate a simulation study to assess the behavior of such measures, under different scenarios defined, for example, by the probability of a word  $w_t$  occurring in a document and the distribution of document variables in the corpus.

## 1.2 Distributed representations of words

The illustrated document representations have two main drawbacks: first, as already mentioned, they map documents in a very high and sparse dimensional space, and, second, they do not consider the semantic relationships between words.

Therefore, new studies worked towards obtaining low-dimensional representations of documents, focusing on words, and starting from their vector representations inspired by the well-known Firth's *distributional hypothesis* [24]: "*you shall know a word by the company it keeps*"; it implies searching for word representations in such a way that words appearing in similar contexts have similar representations. Generally, the *distributed representations* of words are trained by means of an *embedding* function mapping the words into a  $\mathbb{R}^E$  *semantic space*, representing  $w_t$  as  $\mathbf{e}_{w_t}$ , in the following matrix  $\mathbf{W}$  of *word embeddings*:

$$\mathbf{W} = \begin{bmatrix} \mathbf{e}_{w_1} \\ \mathbf{e}_{w_2} \\ \vdots \\ \mathbf{e}_{w_V} \end{bmatrix} \in \mathbb{R}^{VE} \quad (1.7)$$

In this way, semantics becomes a kind of intrinsic value of the word, which only acquires meaning in relation to other words. In this sense, embeddings create *language models* because, starting from the data, they generate abstract representations of language elements in order to propose relationships between them.

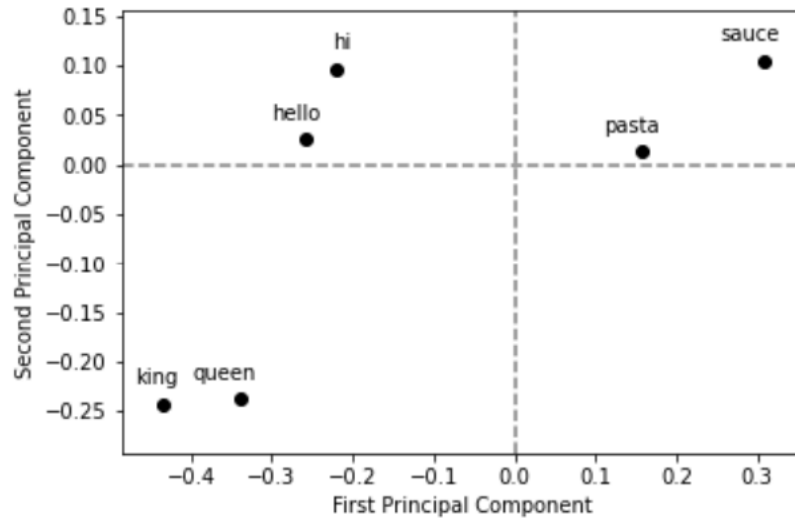
To give an intuition of those concepts, we introduce here an example, without delving into the specifics of vector training methodologies, which will be shown later.

In 2013, Google made available a language model of the type here discussed, i.e. a matrix of word embeddings creating a semantic space<sup>1</sup>. They used a massive corpus of Google News from a broad range of domains, to obtain a matrix  $\mathbf{W}$  with  $V = 3000000$  and  $E = 300$ . We then explored this model, and in the following we show the results of some calculations, to make the reader understand why this matrix  $\mathbf{W}$  creates a vector space that is *semantic*.

By projecting the semantic space into its first two principal components (see for example [25]), one can look at a word's position with respect to others. Six word vectors are shown in Figure 1.1. Despite the low percentage of variance explained by the first two components (about 7% of the total variance), the subspace they created is useful to look at some examples. The words *hi* and *hello* are close to each other, and far from the other two groups, consisting of *king* and *queen*, on the one hand, and *pasta* and *sauce*, on the other.

<sup>1</sup>Available here: <https://code.google.com/archive/p/word2vec/>

**Figure 1.1** Vector representations of 6 words projected into the 2-dimensional space defined by the first two principal components.



Neighbourhoods of these words can be further explored in an attempt to label their semantic field. To this end, if one considers a word vector, one can find those that are most *similar* to it. The similarity between two word vectors  $\mathbf{e}_{w_t}$  and  $\mathbf{e}_{w_s}$  can be measured by the *cosine similarity*, given by:

$$\cos_{w_t, w_s} = \cos(\mathbf{e}_{w_t}, \mathbf{e}_{w_s}) = \frac{\mathbf{e}_{w_t} \cdot \mathbf{e}_{w_s}}{\|\mathbf{e}_{w_t}\| \|\mathbf{e}_{w_s}\|} \in [-1; 1] \quad (1.8)$$

where  $\|\cdot\|$  represents the  $L_1$  norm of a vector.

In Table 1.1 the five most similar terms for *king*, *hello*, and *pasta* are reported.

**Table 1.1** Top 5 terms most similar to the representations of *king*, *hello*, and *pasta* (cosine similarities in parenthesis).

Word	5 Most similar terms
king	queen (0.65) monarch (0.64) prince (0.62) sultan (0.59) ruler (0.58)
hello	hi (0.65) goodbye (0.64) howdy (0.63) goodnight (0.59) greeting (0.59)
pasta	tomato_sauce (0.70) polenta (0.7) tortellini (0.7) ravioli (0.69) gnocchi (0.68)

One could thus identify three semantic fields: that of greetings and social interactions, that of royalty and titles, and that of culinary and food-related concepts (or perhaps Italian cuisine).

Semantics can be further explored by means of *analogies*. An analogy is a statement based on the words  $w_t, w_s, w_p, w_b$  of the type " $w_t$  is to  $w_s$  as  $w_p$  is to  $w_b$ ". Then  $w_b$  will be the word with the most similar vector to the one resulting from  $\mathbf{e}_{w_s} - \mathbf{e}_{w_t} + \mathbf{e}_{w_p}$  [26] [27], i.e.:

$$w_b = \operatorname{argmax}_{\mathbf{e}_{w_b}} \cos(\mathbf{e}_{w_b}; \mathbf{e}_{w_s} - \mathbf{e}_{w_t} + \mathbf{e}_{w_p}) \quad (1.9)$$

In Table 1.2 results of several analogies are shown. For example, the vector of *france* is the closest to  $\mathbf{e}_{italy} - \mathbf{e}_{rome} + \mathbf{e}_{paris}$ , and also the very famous analogy "*male* is to *king* as *female* is to *queen*" yields a good result.

**Table 1.2 Examples of word analogies (cosine similarities in parenthesis).**

Analogy	Result
rome : italy = paris : x	france (0.49)
rome : italy = berlin : x	germany (0.48)
male : king = female : x	queen (0.67)
boy : brother = girl : x	sister (0.82)
monkey : mammal = frog : x	amphibian (0.62)
monkey : mammal = crow : x	bird (0.52)
have : has = do : x	does (0.75)
cut : cutting = write : x	writing (0.72)
morning : sunrise = evening : x	sunset (0.63)

### 1.2.1 Word and document embedding methodologies

Approaches to generating word embeddings have evolved over the years, with an early technique being *Latent Semantic Analysis* (LSA) [28].

LSA employs *Singular Value Decomposition* (SVD) [29] on **DTM** to reduce the number of columns while maintaining the similarity structure, obtaining low-dense vector representations of words and documents. SVD allows the document-term matrix of the type in Equation (1.1) to be represented as:

$$\mathbf{DTM} = \mathbf{D} \cdot \Sigma \cdot \mathbf{T}' \quad (1.10)$$

where  $\mathbf{D}$  and  $\mathbf{T}$  are orthogonal matrices of dimensions  $D \times r$  and  $V \times r$ , respectively, and  $\Sigma$  is an  $r \times r$  diagonal matrix, defined as the matrix of *singular values*. If the generic column of **DTM** is denoted by  $h_{w_t}$ , by selecting the  $k$  largest singular values, the low-dense representations  $\mathbf{h}_i^{(k)}$ , and  $\mathbf{h}_{w_t}^{(k)}$  of the document  $d_i$  and the word  $w_t$  respectively, can be obtained from the followings:

$$\mathbf{h}'_i = \mathbf{h}_i^{(k)} \cdot \Sigma^{(k)} \cdot \mathbf{T}'^{(k)} \quad (1.11)$$

$$\mathbf{h}_{w_t} = \mathbf{D}^{(k)} \cdot \Sigma^{(k)} \cdot \mathbf{h}_{w_t}^{(k)} \quad (1.12)$$

By recalling the notation in (1.2),  $\mathbf{h}_{w_t}^{(k)} = \mathbf{e}_{w_t}$  is the embedding of word  $w_t$ , and  $\mathbf{h}'_i^{(k)} = \mathbf{e}_i$  is the embedding of document  $d_i$ , which takes into account the words occurring in it and the relationships among them.

LSA is thus an embedding technique based on VSM. More recent approaches rely on a different data structure to generate word embeddings: in its general form, it is a term co-occurrence matrix whose generic element counts the occurrences of the term  $w_t$  in the *context* of the term  $w_{t'}$ , where the context is a sequence composed of a predefined number of terms. From this conceptual starting point, numerous methods have been proposed.

It has been shown that neural network-based embeddings<sup>2</sup> outperform LSA, especially in the task of analogies [26]. Thus, a vast number of models of this kind have been proposed, but the most famous one today is known as the *Word2Vec* model [31]. It employs a simple architecture of a single-layer feedforward neural network. The

<sup>2</sup>[30] offers a comprehensive overview of neural networks and their application to NLP.

purpose of this network is to gradually analyze all the text by sliding a *context window*, and for each sequence of words either predicting a word given its context (in its *Continuous Bag-of-Words* variant) or predicting the surrounding context words for a given word (*Skip-gram* variant).

However, this model by itself does not enable the generation of document vectors. To address this limitation, a modification of Word2Vec was proposed [32], which has become well-known under the names *Doc2Vec* or *Paragraph2Vec*. This method will be discussed in Chapters 2 and 3.

Another widely used method for training word vectors is called *Global Vectors for Word Representations* (GloVe). As Word2Vec, it considers the term co-occurrence matrix counting the occurrences of  $w_t$  within the context of  $w_{t'}$ , but, as LSA, it makes use of matrix factorization to obtain the embeddings [33].

## 1.2.2 Embeddings with document variables

The methodologies illustrated in Section 1.2.1 for training (semantic) representations of words and documents treat the corpus as a whole. However, if the corpus consists of documents divided into groups, the final representations may be influenced by any imbalances between the groups, as the presence of a more numerous group could dominate the training phase [34]. Working with grouped document corpora is a rather common scenario. Consider, for instance, the study of the language used by social media user groups, politicians with different alignments, consumer groups writing reviews of products, job listings categorized by industry, medical reports categorized by pathology, and so on.

Although there is not extensive literature on this topic, some modifications of the methods mentioned above have been proposed to take into account document metadata.

To capture the semantic differences between U.S. states, [35] extends the *skip-gram* variant of Word2Vec. The goal is to obtain, for each word in the vocabulary, both a *global* representation and a *category-specific* one, i.e. a state-specific representation. This approach allows capturing the effect of a specific category on the final word representation.

To address demographic differences, [36] suggests training separate word vectors for each group. They also propose a method for comparing word representations derived from different groups. Meanwhile, [37] introduces two modifications to the model proposed by [35].

[38] applies Word2Vec to parliamentary corpora. They expand the architecture's input to consider both the party affiliation of the speaking member of parliament and a time variable.

Lastly, [39] enhances GloVe word vectors with contextual information from words: the input matrices consist of both a matrix of word co-occurrences and a matrix of contextual variables (topics, demographic variables, etc.). Working with a similar matrix factorization of word co-occurrences, [40] presents a method to obtain a matrix of *base* embeddings and another matrix for modeling how a specific document variable affects the *base* embeddings, similar to the output obtained by [35].

## 1.3 Contextualized embeddings

Traditional word embeddings like the ones illustrated in Section 1.2 assign a *static* vector representation to each word regardless of its context. This situation does not

always correspond to reality, since the word usage, and then word sense, does depend on the context. In Table 1.3 we propose some examples of English words that, depending on the context, assume different meanings. For instance, the representation of the word *park* given the sentence "I have to *park* my car" differs somewhat from its representation given the sentence "We are going to the *park* for a picnic".

**Table 1.3 Examples of English words changing sense depending on context.**

Word	Sentences
park	I have to <b>park</b> my car
	We are going to the <b>park</b> for a picnic
clip	She used a <b>clip</b> to secure her hair
	He attached the document with a paper <b>clip</b>
cell	Prisoners lived in <b>cells</b>
	Billions of <b>cells</b> make up the human body
bass	He caught a large <b>bass</b> while fishing
	She can play the electric <b>bass</b>
left	Turn <b>left</b> after the first block
	He <b>left</b> the keys at home
pitch	The baseball player threw a fast <b>pitch</b>
	He's able to hit the high <b>pitch</b> notes
row	<b>Row</b> across the river to the other bank
	Delete the first <b>row</b> of the matrix

This led to the proposal of *contextualized* word representations, to represent words with multiple vectors based on their contexts, and use them as the base of complex architectures to represent pieces of text [41].

*Transformers* are the most popular architectures in recent years that are based on the idea of contextualized embeddings. Their main characteristics are the ability to handle *long-term dependencies*, *high parallelizability*, and adaptability to *transfer learning*.

At their core, Transformers employ *self-attention* mechanisms, allowing each word in a sequence to attentively interact with all other words in the sequence. This capacity empowers the model to capture intricate, long-range dependencies and craft contextualized embeddings [42]. The relevance of these long-term dependencies lies in scenarios where the relevance of a word at the beginning of a sequence influences the interpretation of a word at the sequence's end. Before the advent of Transformers, language models may struggle with retaining this crucial contextual information. For instance, consider the sentence: "The young boy always carries his guitar with *him*"; without accounting for long-term dependencies, the model may not remember "boy" from the beginning, leading to ambiguity in the appropriate pronoun (him, her, it) at the end.

Transformers exhibit exceptional parallelizability, making them adept at processing extensive data swiftly. This attribute is particularly valuable for managing substantial datasets and harnessing the computational power of multiple *graphics processing units*.

Lastly, Transformers have opened the doors of *Transfer learning* to NLP practitioners. It is a two-step process: *pre-training* and *fine-tuning*. During pre-training, a transformer model is trained on a massive corpus of text. This phase allows the model to learn general language patterns from the data. Then the model's knowledge can be adapted (or fine-tuned) for specific downstream tasks on a task-specific text corpus and domain. One of the major advantages of this *modularity* of the process is that one can load an already available model and use it in the fine-tuning phase, without the need to train one's own model, as well as load several models and compare their performance to one's own case<sup>3</sup>.

One of the most influential NLP models has been proposed by a research team at Google AI: *Bidirectional Encoder Representations from Transformers* (BERT) [43]. It is a transformer-based model, and distinguishes itself by its bidirectional nature, considering both left and right contextual information when crafting embeddings, i.e. looking at the piece of text before and after a word to create its embeddings. BERT has been trained on a huge corpus consisting of books and English Wikipedia pages, totaling 3,3 billion words.

BERT has been applied and fine-tuned across a wide range of tasks and application domains [44] [45] [46] [47].

Furthermore, several BERT-like architectures have been proposed, e.g. RoBERTa, trained on larger corpus than BERT [48], DistilBERT and ALBERT, for applications with limited computational resources [49] [50], and SciBERT, pre-trained on 1,4 million scientific papers from Semantic Scholar [48].

## 1.4 Research production

In this section, we show the time evolution of research products related to the various approaches to text representations outlined in this chapter. The data source is Scopus<sup>4</sup>, one of the most important databases of peer-reviewed scientific literature. The term "research products" in this context refers to books, conference proceedings, and articles in scientific journals.

The queries used to search for papers in the database are formulated based on the titles, abstracts, and keywords of the manuscripts. Following the approach adopted in the chapter, three topics have been identified:

- **VSM**, including Vector Space Models with also a focus on term-weighting schemes and TFIDF;
- **LSA**, Word2Vec (**W2V**) and Doc2vec (**D2V**), **GloVe**, considering both separately and as a whole;
- **BERT**.

The boolean queries used for Scopus database searching are reported in Table 1.4.

---

<sup>3</sup>A great source of pre-trained language models is the Hugging Face community: <https://huggingface.co/models>.

<sup>4</sup><https://www.scopus.com/>

Table 1.4 Queries on Scopus database by topics.

Topic	Query on Scopus
VSM	TITLE-ABS-KEY("vector space model") OR TITLE-ABS-KEY("term-weighting scheme") OR TITLE-ABS-KEY("term weighting scheme") OR TITLE-ABS-KEY("TFIDF")
LSA, W2V & D2V, GloVe	TITLE-ABS-KEY("latent semantic analysis") OR TITLE-ABS-KEY("latent semantic indexing") OR TITLE-ABS-KEY("word2vec") OR TITLE-ABS-KEY("doc2vec") OR (TITLE-ABS-KEY("glove") AND TITLE-ABS-KEY(word embeddings))
BERT	TITLE-ABS-KEY("bert") AND TITLE-ABS-KEY("transformers")

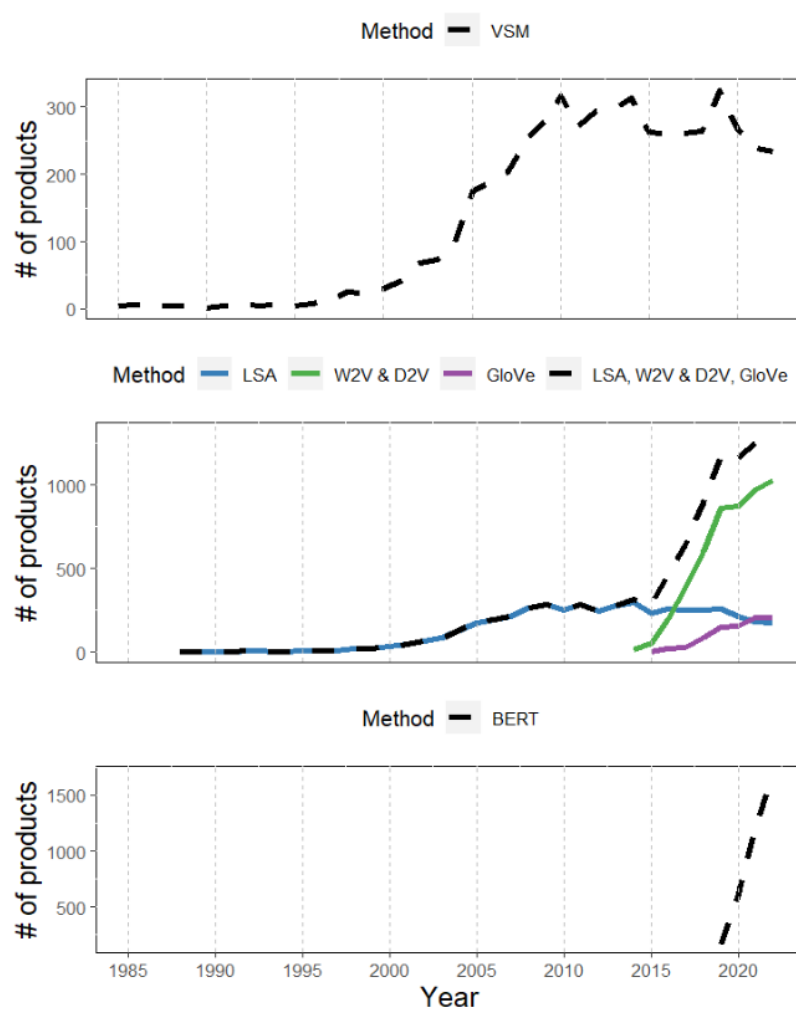
In Figure 1.2, the number of research products published from 1985 to 2022 is displayed, categorized by the approach to corpus representation. Dashed lines represent a theme, while subthemes of research are represented by solid and colored lines, in the second plot.

The entire field began to gain interest in the late 1990s, even though the first use of the term "*Vector Space Model*" dates back to 1975 [2]. VSMs and LSA share a similar increase in production until 2013-14, when Word2vec and Doc2vec algorithms are first introduced. The production on the latter topics rapidly increased, dominating the field until 2019, with the introduction of BERT, two years after the seminal paper proposing the self-attention mechanism and Transformers [42].

In a few years, BERT-based models have garnered attention in research, both methodologically, with the proposal of increasingly sophisticated modifications to its architecture, and applicatively, thanks to the opportunities offered by Transfer learning.

In conclusion, contextualized word embeddings lay the foundation, followed by transformers, which usher in a new era of NLP capabilities. BERT and other Transformers-based models, e.g. GPT and XLNet [51] [52], harness the power of contextualization, long-range dependencies, and pre-training to achieve remarkable performance in a multitude of natural language processing tasks.

Figure 1.2 Time evolution of the number of research products published from 1985 to 2022, categorized by the approach to corpus representation (dashed lines); in the second plot, subthemes of research are represented by solid and colored lines.





## Chapter 2

# Users' self-description on social media

## 2.1 Introduction

In the realm of human interaction, the manner individuals choose to present themselves to others is an intriguing aspect. The process of self-presentation not only comprises *which* information is shared, but is equally shaped by *how* that information is conveyed.

To illustrate this concept, let us consider a scenario that, while not the most likely (hopefully), serves as an illustrative example. Envision a job applicant in an interview setting, where the interviewer, rather than allowing the candidate the freedom to provide a narrative introduction, rigidly insists that the candidate respond to fundamental inquiries with *labels* - like "engineering" and "30" - in response to the questions, "What is your educational background?" and "How old are you?". This scenario raises the question of what nuanced aspects of self-presentation, such as the chosen words and themes, etc., could remain unexplored in the analysis of self-descriptions based solely on these simplistic labels.

Transcending the individual case, exploring self-presentations among diverse groups of individuals may reveal intriguing insights about the ways people choose to describe themselves, such as whether there exist recurrent, shared patterns in self-presentation among individuals within the same social or identity group, or whether certain themes are common to all groups.

From a methodological point of view, this involves the integration of different types of data: textual data, and data represented by variables with predefined categories, distributing texts into groups.

This study provides a methodology for this integration in the domain of self-descriptions on social media. This proposal includes, on the one hand, the semantic representation of words and entire self-descriptions in a common vector space, using one of the tools described in Section 1.2.1, and on the other hand, the identification of *important* words for groups of individuals, by means of count-based tools described in Section 1.1 for documents with membership information, which will be used as a *guide* to study the semantic space and then explore group differences in self-description.

### 2.1.1 Self-describing on social media

Nowadays social media platforms act like sites for people to show themselves, besides the interaction activity. On the one hand, people can share their new achievements, daily thoughts or favorite piece of arts, bad personal events, and other *microdetails* about themselves. On the other hand, people can take care of their own

profile, filled in with their jobs, location, sexual orientation, and generally, depending on the specific domain of the social media, whatever *macrodetails* contributing to draw a self-portrait. This work focuses on the latter, i.e., explicit self-descriptions, less dynamic than the former.

Among the information available on a user profile, we focused on (1) *categorical* self-labels and (2) *textual* self-descriptions. First, we looked at *categorical* self-labels, i.e. answers to *structured* questions allowing the user to choose among a fixed number of options. Consider the following examples: on LinkedIn, users can indicate their location and their job's industry; on Facebook, gender, relationship status, and sexual orientation are few examples of user self-labels; scholars may express their discipline of expertise, languages spoken, and their position, on ResearchGate; lastly, on Reddit, one is required to indicate its gender identity, *inter alia*.

However, limiting on this type of information to analyse social media's users may cause a non-negligible loss of information. Indeed, the pivotal point of this study is represented by *textual*, or *unstructured*, self-descriptions, i.e. what usually are called *bios*. All the platforms in the above example allow users to write a bio.

The toy example in Table 2.1 helps to glimpse why using both the aforementioned types of self-description leads to more complex analysis and results. From those bios of users in a social media for scholars, one would be able to state that young researchers are more focused than full professors on using their bios to show up and frame themselves as researchers. *Game theory* is specific to mathematicians, whereas *Ancient Greeks* are mentioned by humanists; eventually, new technologies are neither discipline nor position-specific.

**Table 2.1 Toy example: categorical self-labels and textual self-descriptions in a social media for scholars.**

Bio	Discipline	Position
<i>Enjoy your life, read sci-fi novels, explore new technologies, and study Game theory!</i>	Mathematics	Full Professor
<i>Ancient Greeks used to say: "Love prudence". This should be the way to judge the world's phenomena.</i>	Literature	Full Professor
<i>I'm a PhD Student in Mathematics. My research interests are optimization problems and game theory.</i>	Mathematics	PhD Student
<i>As a PhD Student in Literature, I'm studying how new technologies are affecting the way we read Ancient Greek classics.</i>	Literature	PhD Student

The purpose of this work is to study the language of self-descriptions on social media, by looking at how groups of users distribute around specific terms and semantic field of interest. In doing so, this study contributes to the literature by proposing a new methodology that leverages both categorical and textual self-descriptions: (1) it combines domain knowledge and statistical measures to find relevant words for grouped users, (2) trains a language model to study the distribution of groups of users around words found in (1), as well as to find similar terms, and (3) uses a bootstrap procedure to assess the variability of the results. Additionally, this work broadens the literature about the social media StockTwits, less explored than other platforms, such as Facebook, Twitter, and Reddit, and developed for discussions about stock markets.

This chapter is organized as follows: in Section 2.2, we illustrate relevant works about users' self-descriptions on social media, as well as about the social media

StockTwits; Section 2.3 provides a detailed explanation of the proposed methodology and the data on which it is applied; in Section 2.4, results are shown and commented<sup>1</sup>.

## 2.2 Related works

Despite the extensive literature about social media users [54] [55] [56] [57], to the best of our knowledge, there is a lack of studies looking at the relations between spontaneously self-declared descriptions, both textual and categorical, on social media.

On the one hand, there are studies based on textual self-descriptions.

*Word association thematic analysis* was used in [58] to detect gender differences in self-descriptions among Twitter users. However, since Twitter does not indicate gender identity, this information was inferred from the username, exploiting official tables of popular first names, and pronouns in the bio.

Starting from human annotations, [59] found out that self-descriptions contribute more than posts and usernames in inferring demographic attributes, namely age and ethnicity, from a list of *signals*.

On the other hand, there exists extensive literature about categorical self-labels.

In a study of mobility, [60] exploited the users' declaration of location, name, and surname on Twitter, to group people transiting Chicago. To investigate behavioural differences in the use of social media between rural and urban users, [61] looked at the users' profiles to extract the location, as independent variable, and gender and privacy settings (private or public profile), among several independent variables.

[62] proposed a model to predict education and job from users' tweets, by using, as ground truth, self-declared information on the related Google Plus users' accounts; they did not consider bios, assuming a lack of data.

Lastly, [63] proposed a classification framework based on the analysis of social networks; *inter alia*, they used data from the social media Flickr, where users share their photos and declare their interests by subscribing to groups of interest, that form the basis for the analysis of user interactions.

As mentioned, our study exploited both the textual and the categorical self-descriptions. In [64], several classification models have been trained to predict, *inter alia*, political alignment of Twitter's users classified themselves as either Democrats or Republicans in the Twitter directories WeFollow and Twellow; however, in leveraging the bio field, they only used regular expressions to detect explicit mentioning of political orientation.

Differently, the methodology of this study was based on using *Natural Language Processing* methodology to fit a language model on self-descriptions, and studying how groups of users distribute around specific words of interest, using the self-declared categorical labels to define those groups.

To propose a methodology to study users' self-descriptions, we considered users' data from StockTwits (ST). Founded in 2009, ST lets investors, traders, and finance enthusiasts share their ideas about the stock markets. When people join the platform, they may fill in their profile with a *bio*, and the clarification of which kind of traders they are, according to their *experience*, *approach* and *primary holding period*.

Regarding the literature on StockTwits users, on this platform people post tweets, and interact with other users' content; but particularly, depending on their confidence in stock price rising, users usually tag their tweets with *bullish* or *bearish*,

---

<sup>1</sup>This chapter is an extended version of the work presented at the 51st Scientific Meeting of the Italian Statistical Society on June, 2022 [53].

and/or with a specific stock symbol. This may explain why literature about ST have focused on exploiting the large volume of tweets to predict the stock market behavior [65] [66] and, generally, for *sentiment analysis* [67] [68] [69].

## 2.3 Methods and data

### 2.3.1 Methods

The analysis included the following steps:

1. Sample selection;
2. Seed word search;
3. Word and document vector training;
4. Neighbourhood exploration.

On a sample of text documents with balanced characteristics, an analysis of *important* words has been performed. On the same sample, a language model has been trained to obtain word and document vectors, and, as a final step, the characteristics of documents in the neighbourhoods of important words have been studied.

In this context, fitting a language model implies a decision about the way of representing text documents, i.e. the users' self-descriptions. In the bag-of-words models, a text document is represented as a vector of weights of the words occurred in the entire corpus of documents [7]. This approach has two main drawbacks: first, they do not consider the *semantic relationships* between words, and, moreover, they illustrate documents in a very high dimensional space. To overcome both issues, the well-known *Paragraph2Vec* algorithm was used to obtain low-dimensional representations of documents, by applying the intuition that words appearing in similar context have to assume similar representations, and documents should be as close as possible to the words that most represent them [32] [31].

In the following subsections, we expose the methodology in general notation.

**Sample selection** Since the purpose is to train a language model to study group differences in self-describing, we need the groups to be evenly represented in the training process, to prevent the model from overfitting specific linguistic signals [34].

Suppose a corpus of text documents  $\mathbb{C}$  and  $q$  document binary variables  $X_1, \dots, X_q$  are given.

According to all the possible combinations among the document variables,  $\mathbb{C}$  is divided into  $2^q$  strata, and the balanced sample of documents  $\mathbb{D} = \{d_1, \dots, d_n\}$  is extracted, such that each stratum contains  $\frac{n}{2^q}$  documents.

In the example in Section 2.1, where  $q = 2$ , the sample selection step would require to draw  $\frac{n}{4}$  samples from each stratum, i.e. from each combination of  $X_{Discipline}$  and  $X_{Position}$ , assuming the values *Mathematics* or *Literature*, and *Full Professor* or *PhD Student*, respectively.

**Seed word search** In a piece of text, a *token* may be either a single word or a sequence of adjacent words; a two-word token is called *bigram*. We define as *seed word* any token that is crucial for the exploration of a language.

A seed word may come from domain knowledge [70] [71], but here we also consider as a seed word any statistically relevant token in the corpus.

To explore group differences, we search statistically relevant tokens by means of two measures of *feature selection* (FS) in text categorization<sup>2</sup>. Among the methods based on boolean term-frequency, on the one hand, there are ones measuring the stochastic dependence between the document variable and the presence/absence of the term, and the most popular are *Mutual Information*,  $\chi^2$  statistic, and *Information Gain* [72] [22] [73] [74]. They assume only one comprehensive value for the document variable. We considered only one measure among them, that is the *Information Gain* (IG).

On the other hand, [75] and [76] proposed the *Relevance Frequency* and the *Relevance Frequency Ratio*, respectively, based on the consideration that, when weighting a term for a document, only the documents in which the term appears are *relevant* for text classification purposes, and then for separating the groups of documents. We considered a modified version of the two, i.e. the *modified Relevance Frequency* (modRF) [23].

IG provides a measure of reduction in entropy of the variable of interest  $X_j$  achieved by knowing the presence or absence of a term; modRF focuses on the documents containing a term of interest  $w_t$ , and compares the frequencies of the groups of documents defined by  $X_j$ .

Since we assume  $X_j$  binary, the IG of  $X_j$  based on the word  $w_t$  results [77]:

$$\text{IG}_{jt} = \text{IG}(X_j, w_t) = H(X_j) - H(X_j|f_{w_t}) \quad (2.1)$$

where

$$\begin{aligned} H(X_j) &= - \sum_{c=0}^1 P(X_j = c) \log_2 P(X_j = c) \\ H(X_j|f_{w_t}) &= -P(f_{w_t} = 1) \sum_{c=0}^1 P(X_j = c|f_{w_t} = 1) \log_2 P(X_j = c|f_{w_t} = 1) + \\ &\quad - P(f_{w_t} = 0) \sum_{c=0}^1 P(X_j = c|f_{w_t} = 0) \log_2 P(X_j = c|f_{w_t} = 0) \end{aligned}$$

and where  $f_{w_t}$  is the boolean frequency of the word  $w_t$  in the document, i.e. it is equal to 1 if the word occurs at least once in a document. The value of IG is 0 when there is no relationship between the variable and the token, and it increases as the intensity of this relationship increases.

The formula of modRF of the word  $w_t$  for the variable  $X_j$  results:

$$\text{modRF}_{jt} = \log_2 \left( \frac{\alpha + 1}{\gamma + 1} \right) \quad (2.2)$$

where:

- $\alpha$  = number of documents with  $f_{w_t} = 1$  such that  $X_j = c$
- $\gamma$  = number of documents with  $f_{w_t} = 1$  such that  $X_j \neq c$

Note that the argument of the logarithm is the *odds*( $X_j = c|f_{w_t} = 1$ ) where an add-one smoothing is performed to avoid zero division. Moreover, with respect to the original formula in [23], we avoided adding 2 to the ratio to make the absolute value of modRF the same for both categories of  $X_j$ . When  $\alpha = \gamma$ , modRF is equal to 0; it is

<sup>2</sup>Those measures use frequencies of words in the sample corpus, and thus *statistically relevant* does not refer to any inferential properties of words.

greater than 0 when  $w_t$  better represents documents of the group  $c$ , whereas modRF is lower than 0 when  $w_t$  better represents documents of the other group.

The two measures are computed on a subset of the entire vocabulary, because the probability of finding a word may affect the FS process. The purpose is to compute the two measures for words that are neither too rare nor too frequent. To identify this subset we considered the well-known *Inverse Document Frequency* (IDF), which weighs a word inversely with respect to the percentage of documents including it [5] [6]. In formula, the IDF for the term  $w_t$  results:

$$\text{IDF}_t = \text{IDF}(w_t) = \log_2 \frac{|\mathbb{D}|}{|\mathbb{D}_t|} \quad (2.3)$$

where  $\mathbb{D}_t$  is the set of documents including  $w_t$ .

**Word and document vector training** IG and modRF do not capture semantic relationships between words. In order to study the context of words and distributions of groups of users around seed words, we used the well-known *Paragraph2Vec* algorithm to obtain low-dimensional representations of documents. The algorithm outputs vector representations of words and documents, pushing words appearing in similar contexts to assume similar representations, and documents to be as close as possible to the words that most represent them [31] [32].

In particular, we exploited a version of the *Paragraph2Vec*, that is called *Distributed Memory of Paragraph Vector* (DM-PV) model [32]. After mapping each document and each word to a vector in matrix  $\mathbf{D}$  and  $\mathbf{W}$ , respectively, it uses a one-layer feedforward neural network that predicts the central word by taking as input the concatenation of document and word vectors.

Specifically, for  $w_{(1)}, w_{(2)}, \dots, w_{(T)}$  ordered words occurring in  $d_i$ , it maximizes

$$\frac{1}{T} \sum_{t=r}^{T-r} \log P(w_{(t)} | w_{(t-r)}, \dots, w_{(t+r)}, d_i) \quad (2.4)$$

where at each iteration  $w_t$  is predicted by using the *softmax* function:

$$P(w_{(t)} | w_{(t-r)}, \dots, w_{(t+r)}, d_i) = \frac{e^{y_{w_{(t)}}}}{\sum_m e^{y_{w_{(m)}}}} \quad (2.5)$$

defining

$$y_{w_{(t)}} = a + b f(w_{(t-r)}, \dots, w_{(t+r)}, d_i) \quad (2.6)$$

where  $a, b$  are *softmax* parameters and  $f(\cdot)$  is the function concatenating word and document vectors. [78].

At the end of training phase, the process uniquely maps  $w_t$  and  $d_i$  into the low-dimensional representations, i.e. *embeddings*,  $\mathbf{e}(w_t)$  and  $\mathbf{e}(d_i)$ , respectively, in the same vector space. Window's size  $r$  and the vectors' size are the hyperparameters of the algorithm.

**Neighbourhood exploration** The analysis of the embeddings pivoted around the embeddings of seed words.

Defining  $S_{ti}$  as the similarity between the embedding of the seed word  $w_t^*$  and the embedding of  $d_i$ , for each seed word the set of *neighbour documents* is defined as:

$$\mathbb{N} = \{i : P(S_{ti} \geq s_{ti}^*) = h\} \quad (2.7)$$

where  $s_{ti}^*$  is a predefined quantile of the distribution of  $S_{ti}$ . As a measure of similarity  $S_{ti}$ , the cosine similarity has been chosen (see Equation (1.8)), i.e.:

$$\text{cos}_{w_t, d_i} = \frac{\mathbf{e}(w_t^*) \cdot \mathbf{e}(d_i)}{\|\mathbf{e}(w_t^*)\| \|\mathbf{e}(d_i)\|} \quad (2.8)$$

that is the dot product of the word vector and the document vector divided by the product of their lengths.

The following proportion for the binary variable  $X_j$  is subsequently estimated:

$$\hat{p}_{tj} = \frac{1}{nh} \sum_{i \in \mathbb{N}} x_{ij} \quad (2.9)$$

In the application of this chapter, it means that the proportions of professional, fundamental, and medium/long-term traders are computed among the users in the neighbourhood of specific words.

Eventually, by drawing  $B$  bootstrap samples, the *bootstrap* distribution of  $\hat{p}_{tj}$  is estimated; it is used, on the one hand, to compute a 90% confidence interval, and, on the other hand, to calculate the standard errors of the estimates for tests on proportions and differences between bootstrap proportions [79].

The proposed methodology is summarized in Figure 2.1.

### 2.3.2 Data

We considered users' data from StockTwits (ST) to propose a pipeline to study users' self-descriptions. When people join ST, they may fill in their profile with a *bio* using a maximum of 250 characters, and the clarification of which kind of traders they are, according to their *experience*, *approach* and *primary holding period*. These trading characteristics are what we previously called *categorical self-labels*, and they will be used to compute the three binary document variables.

Through the *Application Programming Interface* (API) of the StockTwits platform<sup>3</sup>, all the tweets from 2010 to 2021 have been accessed, retaining all the available information about the posting users; then, following [53], pre-processing at the tweet and user level has been performed.

#### Pre-processing

Out of about 800 thousand users, 11% had both bio and trading information, and we retained the first record per user, to store only the first completion of the profile.

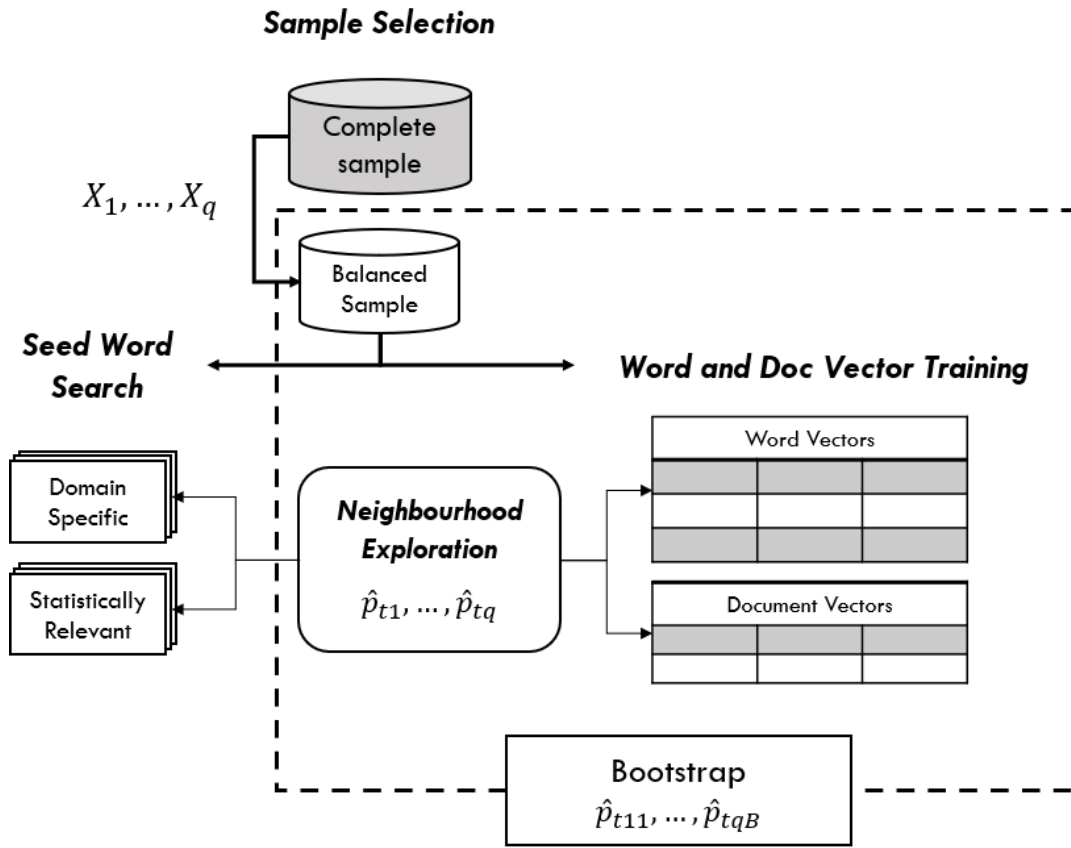
About the language, we used two versions of Google's *Compact Language Detector Algorithm* (CLD), i.e. CLD2 [80] and CLD3 [81], and retained users for whom both the algorithms estimated an English-written bio.

After removing e-mail addresses, URLs, punctuation, and special characters from bios, we removed users with text less than six words long<sup>4</sup>.

<sup>3</sup><https://firestream-portal.stocktwits.com/>

<sup>4</sup>Where 6 was the 25-th percentile of the distribution of the number of words per bio.

Figure 2.1 Summary of the proposed methodology.



Eventually, we eliminated a bio if it contained a six-gram that occurred 10 times or more in the corpus<sup>5</sup>; that's because very similar bios might belong to the same user with different usernames.

This pre-processing yielded a dataset of 29264 users with a textual self-description and self-declared trading characteristics, as illustrated in detail in Table 2.2.

### Trading features

As previously stated, three key self-declared characteristics of trading were taken into account: experience, approach, and holding period. From these labels, we obtained three binary variables, that are  $X_{Professional}$ ,  $X_{HoldingPeriod}$ , and  $X_{Fundamental}$ .

To determine the user's experience level, ST proposes an ordinal scale that includes options for *novice*, *intermediate*, and *professional*. The binary variable  $X_{Professional}$  has been coded to discriminate between professional and non-professional (novice or intermediate) users.

About the holding period, i.e. the amount of time between the purchase and the sale of a security, users can choose among the categories *day trader*, *swing trader*, *position trader*, and *long-term investor*. The binary variable  $X_{Medium/Long-Term}$  has been coded as 1, for *long-term* and *position* investors, and 0 for *day* and *swing* traders. *Day traders* buy and sell securities within the same day, while *swing traders* invest from days to some weeks. *Position traders* and *long-term investors* hold bonds from weeks to months and for over a year, respectively.

<sup>5</sup>Where 10 was the 99.99-th quantile of the distribution of the number of occurrences of the six-grams in the corpus



**Table 2.2 Description of the pre-processing, from original to final sample.**

	# Users	% Users w.r.t. previous step
Total	810446	
<i>Non-missing</i>		
Non-missing bio and trading info	89297	11.02%
<i>Language Detection</i>		
English language (both CLD2 and CLD3)	41339	46.29%
<i>Bio length</i>		
bio length $\geq 6$ unigrams	32704	79.11%
<i>Repeated bios</i>		
bio with 6-grams repeated $< 10$ times	<b>29264</b>	89.48%

The platform provides users with various options for *approach* in investments, including *fundamental*, *growth*, *value*, *technical*, *momentum*, and *global macro*. However, the binary variable  $X_{Fundamental}$  only differentiates between fundamental and technical traders, and it is motivated as follows.

The categories for *approach* do not represent alternative strategies. Indeed, technical analysis looks at future trends by analyzing changes in price, volume, and open interest, mainly through charts [82]. Some technical traders also focus on momentum indicators to measure the speed of price changes and make trading decisions. It explains why  $X_{Fundamental} = 0$  for *technical* and *momentum* approaches, and thus generally considered as technical analysis.

On the other hand, *fundamental analysis* is used to identify overvalued or undervalued securities by analyzing industry and company financial statements and characteristics. *Value* investors use this information to determine a stock's intrinsic value, compare it to its market price, and buy and hold until the intrinsic value is higher than the market price [83]. Conversely, *growth* investors focus on companies that are in the growth phase of their life cycle and have a growth rate higher than their industry average [84]. Therefore, these approaches are collected into the general group of *fundamental* approaches, and thus  $X_{Fundamental} = 1$  for all of them.

We should mention the correspondence between the approach and the holding period. The investment horizon strictly depends on the means for market analysis: technical traders use instruments released daily, or not more than monthly, and then invest for a short-term, whereas fundamental traders look at income statements, that are less frequently released, and therefore prefer a medium/long holding period.

One may note that the definition of  $X_{Fundamental}$  and  $X_{Medium/Long-Term}$ , as well as the relationship between the approach and the holding period, have been motivated by financial literature, but the same motivations are not obvious if considered users' self-descriptions. In [85] it is shown that they are still verified on StockTwits data.

Lastly, since the *Global Macro* category had a frequency of lower than 3%, and since it does not fit into the *fundamental/technical* dichotomy, users applying this approach have been removed from the dataset.

**Table 2.3 Proportion of professional, technical and medium/long-term traders, computed on the binary variables.**

	# Users	% Users
Professional traders	9221	0.32
Technical traders	17012	0.60
Medium/Long-Term traders	10560	0.37
# Users (No Global Macro)	28416	

## 2.4 Results

As motivated in Section 2.3, we wanted to prevent the language model from being biased towards certain linguistic groups. For example, when studying the meaning of word *finance* for a certain population, focusing on the *fundamental* traders, as well as on *professionals*, would most likely lead to a biased definition of what finance is for the population.

Since we considered  $q = 3$  binary document variables, our original corpus was composed of  $2^3 = 8$  strata, and for each stratum 700 users have been drawn, that is the 75% of the minimum among the stratum sizes. The resulting balanced sample has been used for seed-word searching, embedding training, and neighbourhood exploration.

### 2.4.1 Seed words

After tokenization, numbers in bios were replaced with the placeholder  $\langle \textit{number} \rangle$ , as usual, and frequent bigrams have been also considered as tokens.

The IDF was used to identify a subset of the vocabulary for the seed-word search, to select tokens that were neither too rare nor too frequent. Above the 10-th percentile of its distribution, the IDF was higher than 10: it means that the ratio between the number of documents and the number of documents including a word was higher than 1024:1. On the other side, below the 5-th percentile, IDF was around 8, which results in a ratio  $\frac{|D_{w_t}|}{|D|}$  lower than 0.4%, and then in a too rare word. For these reasons, we decide to limit the search to the words between the 5-th and 10-th percentiles of the IDF distribution.

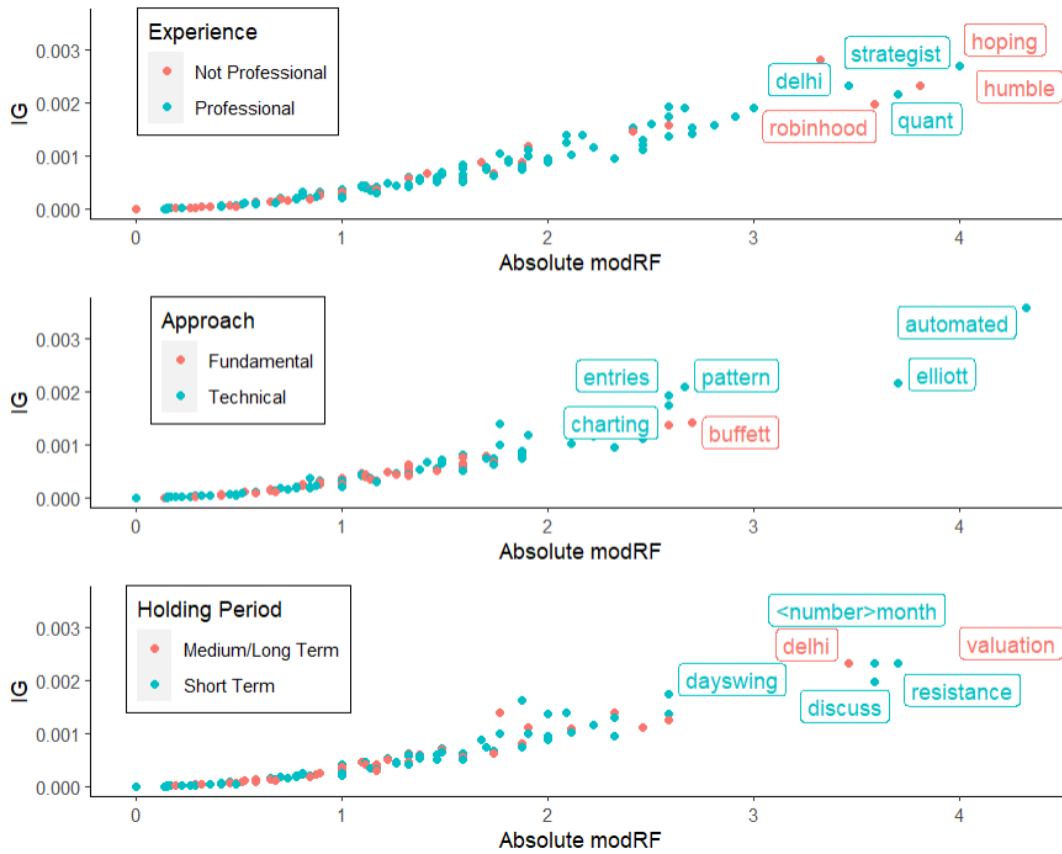
Then, for each binary variable, we calculated the IG and modRF of the tokens. In the end, we considered statistically relevant the tokens whose IG and absolute value modRF were both in the highest 1%. About modRF, it is worth recalling that its absolute value gives a measure of the term's ability to differentiate groups, while its sign tells us which particular group that term is specific to.

Figure 2.2 shows the scatterplots of IG and the absolute value of modRF by trading variables, and textual labels appear for statistically relevant tokens, i.e. seed tokens.

Applying feature selection methods gives already interesting and reasonable insights into the group differences in self-describing. Regarding the experience, domain-specific words set the professionals apart from the non-professionals: one may expect *delhi*, *strategist*, and *quant*<sup>6</sup> to be used by people in the trading world for a non-negligible amount of time, while *hoping* and *humble* recall inexperience.

<sup>6</sup>*Quants* is used to refer either to *quantitative traders*, or *quantitative funds*, that use computational and automatic approaches for stock recommendations [86].

Figure 2.2 Scatterplots of IG and absolute value of modRF by trading variables for tokens between the 5-th and 10-th percentile of IDF; based on the sign of modRF, the colors indicate which category the term is specific to.



About the approach, on the one hand, words like *automated charting*, and *pattern* evoke a technical strategy. Moreover, *elliott* most likely refers to the "Elliott Wave Theory" of the 1920s, a milestone for the technical analysis field, proposing a method to identify patterns in the movement of prices [87]. On the other hand, *buffett*, specific to the fundamental traders, cites Warren Buffett, the most famous *value* investor, i.e. an example of a fundamental trader.

Lastly, about the holding period, in the last plot in Figure 2.2 *< number >\_month* and *dayswing* may probably allude specifically to short-term investors. Additionally, the *resistance* level is a key aspect of technical analysis, namely of short-term trading, since it establishes the level beyond which the price no longer rises during a bullish trend.

About the other words highlighted in the graph, we will explore word and document vectors, to obtain insights into them.

In Table 2.4 selected seed tokens are listed; notice that domain-specific ones are detected with the aim of representing the trading subcategories: e.g., *momentum*, *value* and *growth*, for technical and fundamental trading, respectively, *intermediate* and *novice* for non-professional traders, etc.

**Table 2.4 List of domain-specific and statistically relevant seed tokens.**

Domain-specific			Statistically relevant	
professional	technical	trader	dayswing	robinhood
intermediate	technical_analysis	stock_market	valuation	quant
novice	momentum		resistance	pattern
long_term	value		<i>&lt; number &gt;_month</i>	entries
short_term	growth		delhi	automated
position	<i>&lt; number &gt;_years</i>		discuss	charting
swing	experience		hoping	buffett
day	years_experience		strategist	elliott
fundamental	trading		humble	

## 2.4.2 Neighbourhood exploration

By using a window size  $r$  of 5 terms, and 100 as the word and document vectors' size, the DM-PV version of *Paragraph2Vec* model has been fitted on the sample of users' bios. Then, by means of cosine similarity, for each seed word:

- similar words have been searched among the sampling estimates of vectors of the most frequent words, and results are shown in Table 2.5;
- the bootstrap distributions of  $\hat{p}_{tj}$  with  $X_j = \{X_{Professional}, X_{Medium/Long-Term}, X_{Fundamental}\}$  have been computed via 5000 bootstrap samples, by setting  $h = 0.1$ , and their 5-th and 90-th percentiles are illustrated in Figure 2.3.

Additionally, tests on proportions and differences between proportions were performed, by using the bootstrap estimates to compute the standard errors of the estimates [79].

In this section, results for a subgroup of seed words are commented on, and complete tables of results are reported in the Appendix to this chapter.

Figure 2.3 illustrates the 90% bootstrap confidence intervals for the proportion of fundamental traders, medium/long-term investors, and professional traders, in the

neighbourhood of seed word vectors, where neighbourhood is composed of the 10% most similar users.

Consider the following example to facilitate the readability of the results presented in Figure 2.3 and Table 2.5. On the one hand, in the graph, the first column provides the 90% confidence interval for the proportion of fundamental traders, medium/long-term investors, and professional traders, respectively, in the neighbourhood of the word *professional*. On the other hand, for the same word, we can look at the table for its most similar terms; the cosine similarity between the word *professional* and similar terms is shown in parenthesis.

The neighbourhoods of the words *professional* and *intermediate* are characterized by percentages of fundamental traders statistically lower than 50 (p-values equal to 0.04 and 0.05, respectively). Moving from the *professional* neighbourhood to the *intermediate* and then to the *novice* one results in a decrease in the average percentage of professional traders and an increase in variability. There is no significant difference in the percentage of medium- and long-term investors among the three neighbourhoods (lowest p-value is greater than 0.2), and it is not significantly different from 50% (lowest p-value is greater than 0.3).

In the neighbourhoods of *valuation* and *strategist*, in 90% of the samples, the percentages of professional traders are between 66-76% and 61-74%, respectively, and then sometimes even higher than those in the neighbourhood of *professional*. In fact, if we look at Table 2.5, *valuation* and *strategist* fall into a context of words that are specific to the ST domain and are more likely to be used by users in the trading world. The same is true for the words *professional*, *intermediate*, and *novice*, which seem to identify semantic fields that go from more to less focused on the topic.

Moreover, the percentages of medium- and long-term investors around *valuation* and *strategist* are statistically higher than 50% (p-values lower than 0.001).

The words from *technical analysis* to *position* (excluding *automated* and *elliott*) are those that most recall the link between the approach in trading and the primary holding period: *value*, *growth*, and *buffett* are mostly characteristic of fundamental traders and therefore medium- and long-term investors, while the others recall technical analysts and therefore short-term investors. In particular, moving from the *day* to the *long-term* neighbourhoods gradually increases the percentage of medium- and long-term investors.

It is very interesting to note that the words from *value* to *long-term* in the plot have neighbourhoods characterized by percentages of professionals significantly lower than 50% (the highest p-value is 0.06, for *position*). This is particularly true in the case of *Buffett* (p-value lower than 0.001), who is the most famous value investor in the world. This could be related to the fact that these are very famous words and/or also directly proposed by the ST platform.

Despite the link between the two trading dimensions, *automated* and *elliott* are very useful to distinguish the approach in trading, but not the holding period. Moreover, even if we look at similar words in Table 2.5, they are related to very specific terms in the domain of stock trading and, indeed, have high percentages of professional traders around them: in 90% of the samples, the percentages are between 68 and 81% for *automated*, and between 60 and 75% for *elliott*.

In Section 2.4.1, FS measures suggested that *robinhood* was a useful word to distinguish between professional and non-professional traders. In reality, the percentage of professionals around *robinhood* has the highest variability in the graph, and hovers around 54%. Instead, the percentage of medium- and long-term investors is statistically lower than 50% (p-value lower than 0.05), while among similar words,

*watchlist*, *dm* (most likely an abbreviation of *direct message*), and *receive* appear. Robinhood is actually a trading app, and by reading bios with this term, there seem to be users spamming and recommending this service.

FS measures also identified *resistance* as capable of distinguishing between short and long-term trading. However, the embeddings do not: among the users closest to *resistance*, the percentage of long-term traders is significantly lower than 50%, and even the nearby word vectors in Table 2.5 do not recall technical analysis or short-term investments. In general, it does not seem to belong to any specific category of traders (lowest p-value among the three categories is higher than 0.2), and the similar words in the sample do not recall any semantically interpretable fields.

In conclusion, we can summarize the insights found as follows:

- The words related to trading characteristics expressly proposed by the platform are actually used by users to self-describe, and therefore they are useful to distinguish one group of traders from another, for each trading feature considered separately;
- What distinguishes the professionals from the non-professionals is the use of domain-specific words, but still more searched words. In fact, specific words that identify a trading category proposed by the platform do not specifically belong to a professional audience.
- Technical traders have a very different vocabulary from fundamental traders, often referring to the tools they use to assess the profitability of their current and potential investments;
- The relationship between the approach in trading and the primary holding period also emerges from the bios, while the self-labeled experience transversely influences the self-describing.
- FS measures identify some words as powerful in group discriminating, but the methodology proposed shed light on the high variability that those results may have.

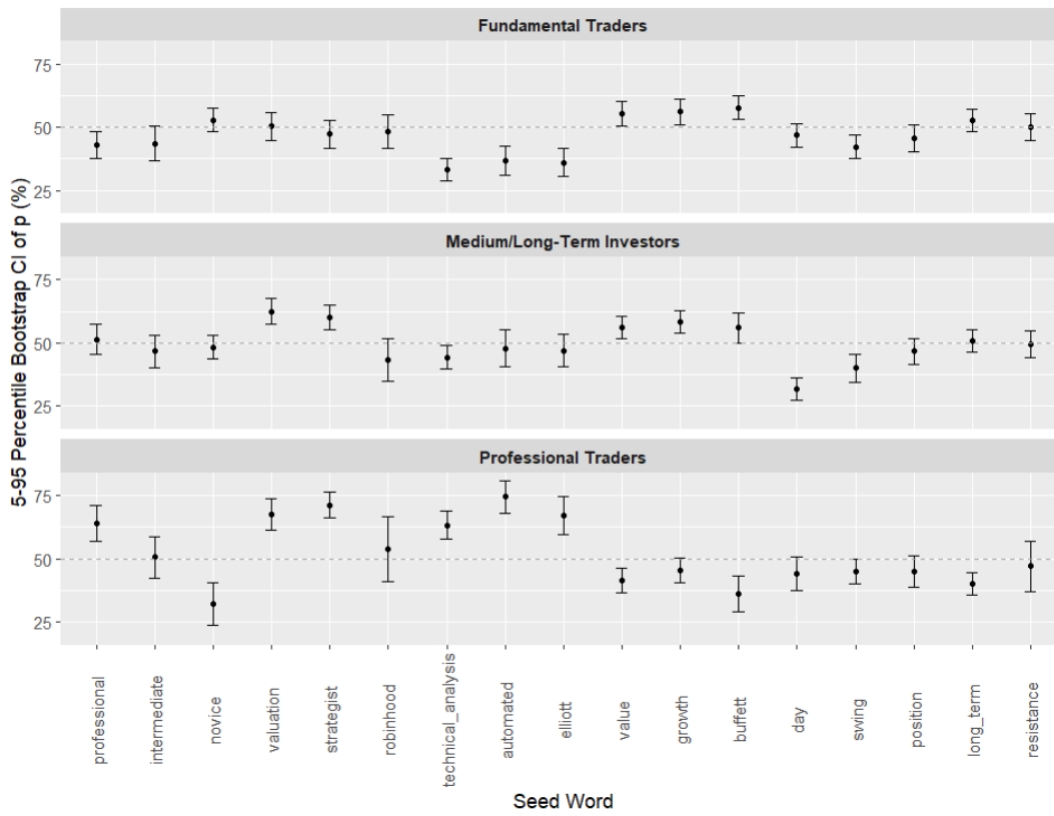
## 2.5 Conclusions

In this chapter, we focused on *categorical* and *textual* self-descriptions to study how different groups of people present themselves on social media, by exploring the case of the StockTwits platform. Using this dual source of data is motivated by the fact that focusing only on the former, or on the latter, may cause a non-negligible loss of information.

Specifically, StockTwits users write a short bio, and specify whether they are either technical or fundamental traders, about their approach, either short-term or long-term investors, about their primary holding period, and either professionals or non-professionals, about their experience in trading.

The methodology proposed consisted of training a language model on a sample of text documents. We worked on a sample with balanced categorical characteristics, to prevent the model from overfitting specific linguistic signals. Then, by using a list of both domain-specific and statistically relevant words as a guide, similarities between word and document representations were explored to analyze group differences in self-describing. Eventually, a bootstrap procedure was leveraged to assess the validity of the results.

**Figure 2.3 90% bootstrap confidence intervals (5-th and 95-th percentiles) for the proportion of fundamental traders, medium/long-term investors, and professional traders, in the neighbourhood of seed word vectors; neighbourhood composed of the 10% most similar users.**



**Table 2.5 Most similar terms for a subset of seed words; cosine similarity (in parenthesis) used as similarity measure, and similar words found among the 10% most frequent words of the vocabulary.**

Seed word	Similar terms
professional	options_trader (0.9) systems (0.88) fx (0.88) trader (0.87) developer (0.87)
intermediate	swingday (0.99) occasional (0.97) dayswing (0.96) specialize (0.96) primary (0.95)
novice	parttime (0.9) amateur (0.89) avid (0.88) self (0.87) taught (0.86)
valuation	analyzing (0.99) models (0.98) strategist (0.98) chief (0.98) strategic (0.98)
strategist	publisher (0.99) chief (0.99) extensive (0.98) specializes (0.98) valuation (0.98)
robinhood	watchlist (0.99) dm (0.98) receive (0.98) track (0.97) consistently (0.97)
technical_analysis	using (0.96) fundamental (0.93) indicators (0.91) chart (0.91) strategies (0.9)
automated	charting (0.96) report (0.96) indices (0.95) custom (0.95) via (0.95)
elliott	wave (0.98) trending (0.98) algorithms (0.98) pattern (0.98) timing (0.97)
value	long_term (0.94) growth (0.93) fundamentals (0.91) technicals (0.91) investments (0.9)
growth	dividend (0.96) investments (0.94) value (0.93) term (0.92) biotech (0.92)
buffett	serious (0.99) buffet (0.99) warren (0.99) gambling (0.98) fool (0.98)
day	every (0.91) trade (0.86) swing (0.84) day_trading (0.83) time (0.8)
swing	swing_trades (0.92) mostly (0.92) trade (0.91) swings (0.9) plays (0.9)
position	short_term (0.98) term (0.97) swing_trades (0.96) momentum (0.96) mostly (0.95)
long_term	value (0.94) short_term (0.93) term (0.93) position (0.92) longterm (0.92)
resistance	ride (0.98) set (0.97) wont (0.95) rarely (0.95) winners (0.95)

This study suggested that the words related to trading characteristics expressly proposed by the platform are used by users to self-describe, and therefore they are useful to distinguish one group of traders from another, for each trading feature considered separately. Generally, bios reflect the relationship between the approach and the primary holding, while are transversely influenced by the experience in trading. Particularly, technical traders have very different semantics from fundamental traders, often referring to the tools they use to assess the profitability of their current and potential investments, like charting, predictive and automated models, etc. In addition, what distinguishes the professionals from the non-professionals is the use of domain-specific words, but still very specific words.

About the bootstrap procedure, it was useful to assess the validity of what was suggested by the feature selection measures: some words were initially detected as able to discriminate between groups of users, but the proposed methodology shed light on the high variability of some of those results.

In conclusion, the proposed procedure can prove beneficial in other contexts as well, that is whenever one wants to explore the heterogeneity of a language across grouped text documents.

## 2.6 Discussion

This study aims to contribute to the analysis of the relationships between labels that summarize characteristics and the texts related to them. Apart from checking



whether the two correspond, it would be interesting to understand how individuals understand a given characteristic, and to study the heterogeneity of the language of people belonging to a certain group, as well as to compare the heterogeneity of groups. For instance, it can be insightful to understand how politicians with different political alignments use the term *democratic* and whether the term itself takes on varying nuances within different political groups.

From a methodological point of view, the purpose of further research is twofold. First, the aim is to improve the sample selection step, to obtain a balanced sample. When working with text data, it is worth considering that statistical units, i.e. words, are organized in a three-level hierarchy, where if usually documents belong to only one group, as in the context of this chapter, words instead may appear in more than one document. In this case, specific techniques of multilevel stratification should be used to create homogeneous groups of statistical units that share the same characteristics at multiple levels. Last, an analysis of sensitivity to the parameter  $h$  should be performed, in order to further validate the estimates of the proportions.

Regarding the search for statistically relevant tokens, Chapter 4 will address a simulation study that analyzes the ability of well-known count-based methods to capture the association between a group of documents and a word. This will serve to assess the role of corpus characteristics, such as the potential different sizes of groups, or the probability that a word appears in a document, in the comparability of the values of measures, like Information Gain and modRF, for different words within the same group and for the same word but across different groups.

## Appendix: Word similarities and bootstrap confidence intervals

In this Appendix word similarities and bootstrap confidence intervals are reported for all the seed words shown in Table 2.4. Due to space issues, the seed words are divided into two groups.

**Table 2.6 Most similar terms for seed words (group 1); cosine similarity (in parenthesis) used as similarity measure, and similar words found among the 10% most frequent words of the vocabulary.**

Seed word	Similar terms
professional intermediate	options_trader (0.9) systems (0.88) fx (0.88) trader (0.87) developer (0.87) swingday (0.99) occasional (0.97) dayswing (0.96) specialize (0.96) primary (0.95)
novice long_term short_term	parttime (0.9) amateur (0.89) avid (0.88) self (0.87) taught (0.86) value (0.94) short_term (0.93) term (0.93) position (0.92) longterm (0.92) position (0.98) term (0.97) momentum (0.97) technicals (0.96) swing_trades (0.94)
position	short_term (0.98) term (0.97) swing_trades (0.96) momentum (0.96) mostly (0.95)
swing day fundamental	swing_trades (0.92) mostly (0.92) trade (0.91) swings (0.9) plays (0.9) every (0.91) trade (0.86) swing (0.84) day_trading (0.83) time (0.8) technical (0.97) patterns (0.97) quantitative (0.94) macro (0.93) technical_analysis (0.93)
technical	fundamental (0.97) patterns (0.95) quantitative (0.95) macro (0.9) technical_analysis (0.9)
technical_analysis momentum value	using (0.96) fundamental (0.93) indicators (0.91) chart (0.91) strategies (0.9) short_term (0.97) position (0.96) mostly (0.95) focus (0.94) term (0.94) long_term (0.94) growth (0.93) fundamentals (0.91) technicals (0.91) investments (0.9)
growth < number > _years	dividend (0.96) investments (0.94) value (0.93) term (0.92) biotech (0.92) full_time (0.85) since_< number > (0.85) past (0.85) yrs (0.85) trading_stocks (0.84)
experience	< number > _years (0.83) professional (0.83) markets (0.82) trading (0.79) past (0.79)
years_experience trading trader	ai (0.99) special (0.99) network (0.99) focuses (0.99) technologies (0.99) forex (0.88) active (0.88) successful (0.87) experienced (0.84) currencies (0.83) mainly (0.92) options_trader (0.92) day_trader (0.91) dayswing (0.9) fulltime (0.89)

**Table 2.7 Most similar terms for seed words (group 2); cosine similarity (in parenthesis) used as similarity measure, and similar words found among the 10% most frequent words of the vocabulary.**

Seed word	Similar terms
stock_market	teaching (0.85) enjoy (0.84) beginner (0.83) new (0.83) successful (0.82)
dayswing	swingday (0.96) intermediate (0.96) occasional (0.95) contrarian (0.94) mainly (0.93)
valuation	analyzing (0.99) models (0.98) strategist (0.98) chief (0.98) strategic (0.98)
resistance	ride (0.98) set (0.97) wont (0.95) rarely (0.95) winners (0.95)
< number > _month	limited (0.97) pm (0.97) rate (0.96) inside (0.95) webull (0.93)
delhi	india (0.97) app (0.97) usa (0.97) helps (0.97) across (0.97)
discuss	posting (0.96) enter (0.95) talk (0.95) robinhood (0.94) receive (0.94)
hoping	works (0.99) aim (0.98) boys (0.98) lots (0.98) gambling (0.98)
strategist	publisher (0.99) chief (0.99) extensive (0.98) specializes (0.98) valuation (0.98)
humble	yet (0.99) challenge (0.98) words (0.98) baby (0.98) havent (0.98)
robinhood	watchlist (0.99) dm (0.98) receive (0.98) track (0.97) consistently (0.97)
quant	metals (0.99) individual (0.97) precious (0.97) specializing (0.97) attorney (0.97)
pattern	algorithm (0.99) various (0.99) driven (0.98) elliot (0.98) algo (0.98)
entries	exits (0.99) premium (0.94) full (0.94) enter (0.92) updates (0.92)
automated	charting (0.96) report (0.96) indices (0.95) custom (0.95) via (0.95)
charting	custom (0.97) newsletter (0.97) automated (0.96) via (0.96) basis (0.95)
buffett	serious (0.99) buffet (0.99) warren (0.99) gambling (0.98) fool (0.98)
elliott	wave (0.98) trending (0.98) algorithms (0.98) pattern (0.98) timing (0.97)

**Table 2.8 90% bootstrap confidence intervals (5-th and 95-th percentiles) for the proportion of fundamental traders, medium/long-term investors, and professional traders, in the neighbourhood of seed word vectors (group 1); neighbourhood composed of the 10% most similar users.**

Seed word	Estimated Proportion of Traders		
	Fundamentals	Medium/Long-terms	Professionals
professional	(37.5,48.4)	(45.5,57.3)	(57,70.9)
intermediate	(36.8,50.5)	(40.4,52.9)	(42.5,58.6)
novice	(48.4,57.5)	(43.8,53)	(23.8,40.5)
long_term	(48.2,57.1)	(46.4,55.4)	(35.5,44.3)
short_term	(41.8,52)	(43.8,53.4)	(40.5,52.1)
position	(40.2,50.9)	(41.6,51.8)	(38.9,51.1)
swing	(37.5,47)	(34.6,45.5)	(40.2,50)
swing_trader	(42.1,51.6)	(39.8,48.9)	(36.8,47.5)
swing_trading	(34.3,44.8)	(34.1,44.5)	(43.9,58.4)
day	(42.3,51.4)	(27.3,36.4)	(37.3,50.7)
day_trader	(42.1,52.1)	(39.3,49.5)	(36.2,48.9)
fundamental	(34.3,43.9)	(47.7,56.2)	(58.2,69.1)
technical	(34.6,44.1)	(47.3,56.1)	(57,68.2)
technical_analysis	(28.7,37.5)	(39.6,48.9)	(57.7,68.8)
momentum	(38.6,47.7)	(42.3,51.1)	(44.3,54.1)
value	(50.5,60.4)	(51.6,60.7)	(36.6,46.4)
growth	(51.1,61.1)	(54.1,63)	(40.5,50.2)
< number > _years	(42,50.2)	(44.1,52.1)	(45.4,55.5)
experience	(38.4,47.9)	(43.6,53.6)	(51.6,65.4)
years_experience	(40.5,55.2)	(49.8,65.7)	(53.6,72.1)
trading	(31.1,39.5)	(33.4,41.8)	(62.3,72.7)
trader	(36.2,44.8)	(41.2,49.6)	(48.8,58.9)

**Table 2.9 90% bootstrap confidence intervals (5-th and 95-th percentiles) for the proportion of fundamental traders, medium/long-term investors, and professional traders, in the neighbourhood of seed word vectors (group 2); neighbourhood composed of the 10% most similar users.**

Seed word	Estimated Proportion of Traders		
	Fundamentals	Medium/Long-terms	Professionals
stock_market	(42,54.8)	(41.2,54.6)	(26.2,58.4)
dayswing	(37.1,48.6)	(38.9,49.8)	(41.8,57)
valuation	(45,55.7)	(57.5,67.5)	(61.3,73.6)
resistance	(45,55.5)	(44.3,54.8)	(37.1,57)
< number > _month	(40.7,52.5)	(34.5,47.9)	(49,66.4)
delhi	(47,56.4)	(54.8,67.7)	(54.3,75.2)
discuss	(38.4,52.9)	(32,48.2)	(39.6,68.2)
hoping	(52.1,60)	(47.5,56.4)	(25,37.7)
strategist	(41.8,52.7)	(55.4,65.2)	(66.4,76.2)
humble	(51.2,59.5)	(46.1,55.5)	(27.3,40.7)
robinhood	(41.6,55)	(35,51.5)	(41.1,66.6)
quant	(38.6,52.1)	(50.9,62.7)	(60,73.6)
pattern	(33.4,48.9)	(43.6,57.5)	(50,71.6)
entries	(35,45.9)	(33,47.3)	(50.4,67.3)
automated	(31.2,42.5)	(40.5,55.4)	(68.2,80.7)
charting	(32,43.6)	(41.1,55.2)	(68,80.5)
buffett	(53,62.3)	(50,62.1)	(28.9,43.4)
elliott	(30.4,41.6)	(40.7,53.4)	(59.5,74.5)



## Chapter 3

# Visitors' reviews of cultural attractions

### 3.1 The DS4BS Project

This chapter presents a research project<sup>1</sup> that is part of a larger research initiative, that is the "Data Science for Brescia - Arts and Cultural Places" (DS4BS) Project, funded by Fondazione Cariplo<sup>23</sup> with the support of Fondazione Brescia Musei<sup>4</sup>, and proposed by DMS StatLab and BODaI-Lab of the University of Brescia (Italy)<sup>5</sup>. The project aims to improve the understanding of how people visit cultural sites (museums, theaters, monuments, and historic buildings) in the Italian city of Brescia. It combines the use of big data, new technologies, and complex statistical methods to achieve this goal; particularly, special consideration has been given to experimenting new methods for public detection and engagement, investigating cultural attitudes and perceptions, and creating new forms of access to culture, especially in the context of cultural tourism.

A Data Science approach is developed under the lens of two integrated perspectives, and the output of the research will be integrated with multimedia contents to define artworks as smart objects.

The first research line concerns the monitoring of presences and crowding in artistic and cultural places. The study uses mobile phone data, which are currently among the best data sources for the study of social phenomena in urban areas [88]. Indeed, they allow for observing the presence and movement of individuals at a high geographical (i.e., small area) and temporal (i.e., short time intervals) level of detail. The mobile phone data at disposal have been provided by Olivetti S.p.A.<sup>6</sup> with the support of FasterNet S.r.l.<sup>7</sup>. The database refers to a selection of *Points Of historical, artistic, cultural, or social Interest* (POI) in the city of Brescia observed during the year 2022. Overall, we have information about 25 POIs, divided into 5 macro-categories (4 monuments, 6 museums, 4 squares, 3 theatres, 8 other typologies), and represented as circles with a 100-meter radius (see the left map of Figure 3.1). For each POI, the database reports the statistical presence, which is defined

---

<sup>1</sup>The project has been presented at the *Statistics for Data Science and Artificial Intelligence Conference* in Pavia (Italy), on April 28, 2023, after the acceptance of the manuscript *Statistics and Data Science for Arts and Culture: an application to the city of Brescia.*, Ricciardi R., Carpita M., Perazzini S., Zuccolotto P., Manisera M., 2023."

<sup>2</sup><https://www.fondazionecariplo.it/>

<sup>3</sup>This work has been supported by Fondazione Cariplo, grant n. 2020-4334, project Data Science for Brescia – Arts and Cultural Places (DS4BS).

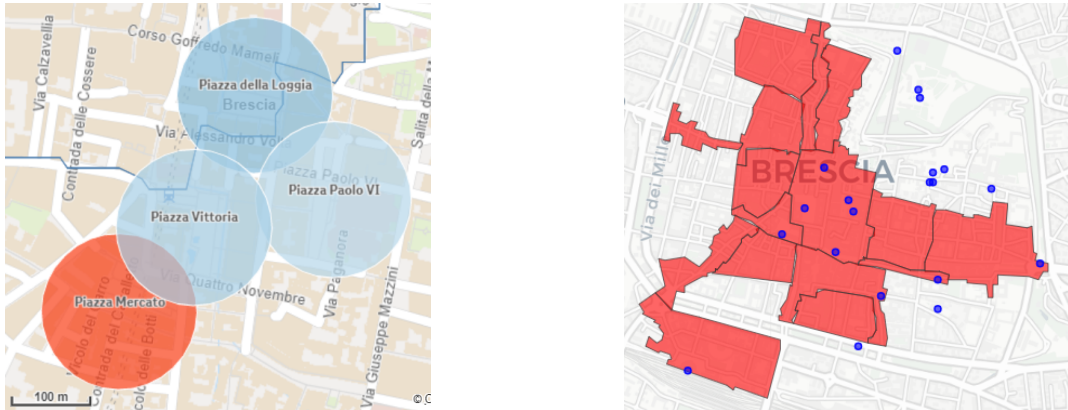
<sup>4</sup><https://www.bresciamusei.com>

<sup>5</sup><https://bodai.unibs.it/ds4bs>

<sup>6</sup>[www.olivetti.com](http://www.olivetti.com)

<sup>7</sup>[www.fasternet.it](http://www.fasternet.it)

**Figure 3.1 Mobile phone and Mastercard data. Left: Points of Interest (POIs) of type "squares". Right: Map of the 16 commercial systems constituting the Distretto Urbano del Commercio (DUC) (red) and overlapping or neighboring POIs (blue).**



as the average number of individuals during a 15-minute interval, along with some information about the individuals (e.g., age and gender).

To further explore the multifaceted impacts of the events, this research line is analyzing people's presences in conjunction with other data sources. In particular, some expenditure indices on transactions registered on Mastercard data have been provided by the Municipality of Brescia for the year 2022, capturing variations in different aspects of expenditure, such as the number of purchases, the total value of purchasing, and the average value of a purchase, and therefore allow us to investigate the economic impact of cultural events. Our database refers to 9 industries (e.g., eating places, accommodations, total apparel, ...) and 3 payment circuits (i.e., international, domestic, overall) in the 16 commercial systems constituting the *Distretto Urbano del Commercio* (DUC) of the city of Brescia (see the right map of Figure 3.1).

The second research line investigates the visitors' experience and is based on both offline and online opinions.

The visitors' experience is evaluated offline by means of questionnaires aimed at evaluating their visit, in terms of expectations and satisfaction with several aspects. In addition, visitors were asked to describe their experience in terms of sight, hearing, touch, smell and taste, even if the only sense concretely used is sight, or hearing. Focusing on this part of the questionnaire, among different response scale formats, the study proposes multi-point semantic differential scales, which required the respondent to position himself/herself on a rating between two bipolar adjectives.

In order to analyse those data with appropriate statistical models, we resorted to the CUM model (Combination of a discrete Uniform and a - linearly transformed - Multinomial random variable [89]), recently proposed in the framework of the CUB (Combination of discrete Uniform and shifted Binomial random variables [90]) class of models.

Moving on to online opinions, this chapter will showcase the results achieved so far in the analysis of Google reviews, provided by Fondazione Brescia Musei.

### 3.1.1 Investigating visitors' online opinions

The aim of this research is to build a language model trained on online reviews posted on Google by people who have visited cultural attractions in Brescia. These reviews are distributed across the city's four main cultural attractions, all managed



by Fondazione Brescia Musei: the *Castle*, the picture gallery *Pinacoteca Tosio-Martinengo*, the archeological site of *Roman Brixia*, and the *Santa Giulia Museum*.

Given that classification, the language model here proposed seeks to obtain a semantic representation of the reviews while categorizing them into four distinct semantic *areas*, which are precisely defined by these four attractions. In this regard, if the four attractions are considered as topics, the research problem here becomes a supervised topic modeling framed as a document multi-classification.

This classification model is designed using reviews for which the attraction is known, but its great utility stems from the fact that it can be used to identify the attraction in text documents, including both reviews and non-reviews, that do not provide this information. The importance of this lies in its ability to bring order to an otherwise overwhelming influx of textual data. Data on social media platforms, forums, and other online spaces constitute a source of information for those managing the attractions. It can help in understanding people's sentiments and the topics that matter most to them, both in terms of general cultural offerings and specific attractions; moreover, it provides insights into whether an initiative, a recent renovation, etc. at an attraction, is a topic of online discussion or not. However, text documents coming from that source lack explicit labels identifying the cultural attraction they refer to. Traditionally, the process of comprehending and categorizing these reviews would entail an arduous manual reading of each review, scrupulously determining which attraction it relates to, and this laborious and time-consuming task becomes harder as the volume of reviews continues to rise. The proposed model can process and classify those texts, allowing for the expansion of the online discourse database concerning those cultural attractions, saving both time and effort.

From a methodological point of view, the analysis presented here is an application of the methodologies generally described in Section 1.3, as it involves fine-tuning a BERT model for a multiclassification task.

The chapter is structured as follows. Section 3.2 is dedicated to reviewing scientific works that apply NLP methodologies in the field of Tourism and Cultural Tourism. In Section 3.3, the high-level architecture of BERT and the specific case of its fine-tuning in a multi-classification context are described, and methods to extract keywords from group of fine-tuned vectors of reviews. Section 3.4 will detail the available data and its random subsets. Eventually, Section 3.5 will show the results of this study.

## 3.2 Related works

The NLP methodologies have not yet dominated the field of *Tourism* research. In this section, we will describe some studies of interest that utilize this group of methodologies to gain insights into people's opinions on tourist-related topics and entities. Additionally, there will be a focus on *Cultural Tourism*, the specific field in which the DS4BS project operates.

The vast majority of studies that utilize NLP techniques in the field of Tourism analyze user reviews, with Tripadvisor being by far the most widely used platform for collecting these reviews.

Methodologically, the techniques used include *Sentiment Analysis* (SA) to explore the sentiment polarity of users, *Topic Modeling* to identify discussion themes, and *Content Analysis* (CA) to extract keywords and their relationships. It is worth noting that these techniques are not always easily distinguishable, but rather they are

mostly mixed in hybrid approaches: they may either combine techniques to more comprehensively address a task or use one technique instrumentally for another.

In tourism research, when reviewing recent contributions chronologically, there have been studies that employed a hybrid approach for SA. For instance, in [91], exploration of tourist reviews published on *Virtualtourist.com* was conducted using the well-known Stanford Sentiment Treebank [92] to extract a sentiment *polarity score* from the reviews. Additionally, co-occurrences between transport-related keywords were investigated, such as *expensive-taxi* and *dirty-metro/bus*.

A lexicon-based approach has been utilized in [93] to perform SA by senses. They employed the *Vocabulary of the Senses* [94] to classify Yelp reviews based on the sensory experiences conveyed by the words in the reviews. They further computed sentiment polarity using the *SentiStrength* lexicon-based tool [95].

Topic modeling and aspect extraction methods were applied in [96] to discover the sentiment and *emotions* of Airbnb users, categorized by discussion themes. They extracted reviews from "green" users using a dictionary related to *sustainability*.

There are also studies that use BERT-based language models for sentiment analysis. For example, [97] built a recommender system based on Tripadvisor reviews, using various approaches, including the utilization of a fine-tuned BERT to group and classify reviews by sentiment. Another study used BERT to extract different *aspects* of interest and analyze rating distributions by aspect [98].

Text documents from surveys and reviews have been subject to CA, aided by the proliferation of CA softwares. For example, [99] used *RapidMiner Studio*<sup>8</sup> to analyze open-ended responses from a questionnaire directed at residents of the Italian region of Puglia, concerning the reasons they believe tourists should visit their region. [100], on the other hand, used KH Coder<sup>9</sup> to analyze the content of Airbnb reviews, in which users anonymously share their reviews after a peer-to-peer accommodation experience with Airbnb.

Finally, highly sophisticated language models like BERT also find applications in tasks other than those mentioned above, as in the research presented in [101]. They proposed a BERT-based methodology tested on reviews from Tripadvisor, Traveloka, and Hotels.com to extract names, locations, and facilities from reviews.

Regarding the specific field of Cultural Tourism, the literature is relatively scarce. In [102], term-weighting methods were applied to represent user reviews of cultural sites and UNESCO squares in Macau to identify user types.

Finally, a study proposed in [103] will be considered for future developments of the DS4BS project. The research is part of a comprehensive study on the quality evaluation of cultural tourism offered by Italian museums. In this specific work, two approaches are proposed to find topics of interest in online Tripadvisor reviews. On one hand, they suggest a top-down approach: domain knowledge is used to identify topics of interest, and both a rule-based classifier and a BERT-based classification model are used to categorize reviews by topic. On the other hand, an unsupervised topic modeling approach with *Latent Dirichlet allocation* is proposed to discover topics in a data-driven (bottom-up) manner.

---

<sup>8</sup><http://www.rapidminer.com/>

<sup>9</sup><https://khcoder.net/>

## 3.3 Methods

### 3.3.1 The Language Model

To train a language model on the reviews of the attractions, a pre-trained version of the BERT model has been selected.

BERT exploits a Transformers-based architecture to create document embeddings and contextualized word embeddings, whose importance has been introduced in Section 1.3.

Utilizing a pre-trained language model entails employing its architecture and the final weights obtained through the original training of the model to obtain a representation of the text corpus. This becomes the input to a neural network with several layers, and at each layer, a new transformation of a text sequence is generated, culminating in the last layer, namely the *last hidden state*, which represents the final output and then the final document embedding.

However, since BERT is not originally trained on a specific domain corpus, it learns very general language patterns and may not be well-suited to represent text documents in a highly specific context. In our case, to obtain more specific representations, we add a classification layer to the architecture with the objective of predicting which of the four attractions the review refers to. This involves a new training phase and, consequently, the updating of the network's weights. At the end of this phase, the last hidden state will yield a new vector representation of our reviews.

A two-dimensional visualization of the review vectors will be presented to highlight the difference between the representation obtained without fine-tuning and those fine-tuned with BERT.

### BERT and Multilingual BERT

The specific model used in this study is the BERT *Base Multilingual Cased*<sup>10</sup>. It is referred to as "base" because it is a neural network with fewer layers than the "large" model<sup>11</sup> and yields lower-dimensional vector representations. It is "multilingual" because it is trained on a corpus of Wikipedia pages written in 104 languages<sup>12</sup>. Lastly, it is "cased" because it retains the original casing of the text: during tokenization, it does not transform the words to lowercase.

The general architecture of this BERT model is described below. The sources of the following explanation are [43], [97], and [104].

To gain a high-level understanding of the BERT architecture, we show it as divided into three phases: (1) document preprocessing to create an input for the final embedding process, (2) the Transformer encoder, which generates the vector-based semantic representation of the document and its constituent tokens, and (3) the prediction objective.

First, Figure 3.2 illustrates how a document is transformed into input for the network and its initial representation before entering the Transformer encoder. The illustration considers an example document consisting of two sentences [43]:

- Sentence A: *my dog is cute;*
- Sentence B: *he likes playing.*

<sup>10</sup>Available at <https://huggingface.co/bert-base-multilingual-cased>

<sup>11</sup>An example of large BERT model is available at <https://huggingface.co/tftransformers/bert-large-cased>

<sup>12</sup>Details of the dataset available at <https://huggingface.co/datasets/wikipedia>

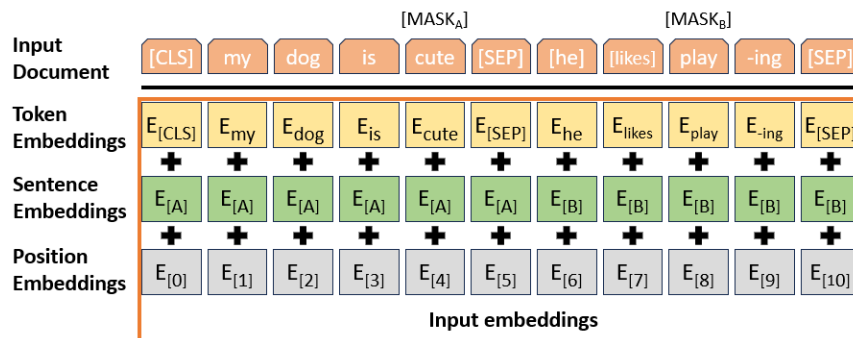
A text document divided into sentences serves as the network's input. These sentences are tokenized, where tokens are not necessarily words but can also be subwords, as in the case of the word *playing* being split into *play* and *-ing*.

This model is trained using a "Masked Language Modeling" (MLM) objective, where a token is "masked" within the sentence and, based on the other tokens, the model attempts to predict the masked token. "Masking" a token involves replacing it with the placeholder *[MASK]* when a sentence enters the network. Specifically, 15% of the tokens are masked, and, among these, a random subsample of 10% is replaced with a random token. Additionally, two placeholders are used at the beginning of the network: *[CLS]* for the start of the first sentence and *[SEP]* between sentences.

Each token is initially embedded as the sum of three different embeddings:

- *token embedding*, which learns to represent each vocabulary element as a vector;
- *sentence embedding*, indicating whether the token belongs to sentence A or B;
- *positional embedding*, which learns to represent a token's position in a document.

**Figure 3.2 Input and Embeddings for the Transformer Encoder in the BERT Representation.**



Author's illustration inspired by [43].

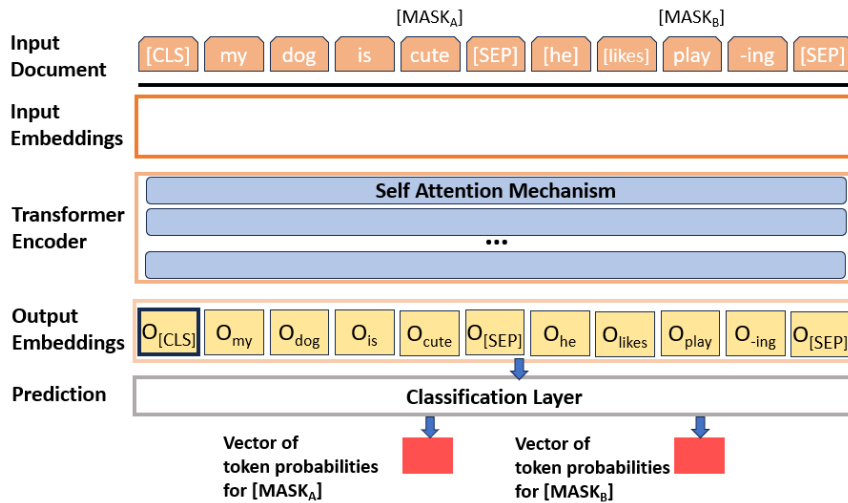
Second, the sum of the three embeddings serves as the input to the Transformer encoder, which returns the output embeddings. These output embeddings are the final semantic representations of each individual token, as well as the entire document,  $\mathbf{O}_{[CLS]}$ , as illustrated in Figure 3.3. The base BERT model returns a vector of size 768.

At its core, this encoder employs self-attention mechanisms, enabling each word in a sequence to attentively interact with all other words in the document. This capability allows the model to capture intricate, long-range dependencies and create contextualized embeddings [42]. For a detailed explanation of how the Transformer algorithms work, refer to [104].

Finally, a classification layer is used to predict the masked tokens using the output embeddings.

This architecture is quite *deep*. The number of parameters, or weights, that the model has to learn during the training phase depends on the number of tokens in the vocabulary, the length and the number of the documents in the corpus, the number of layers in the network, and self-attention operations in the Transformer encoder, and the dimensions of the output embeddings. The *Multilingual Base Cased BERT* model learns the values of 179 million parameters on the Wikipedia corpus.

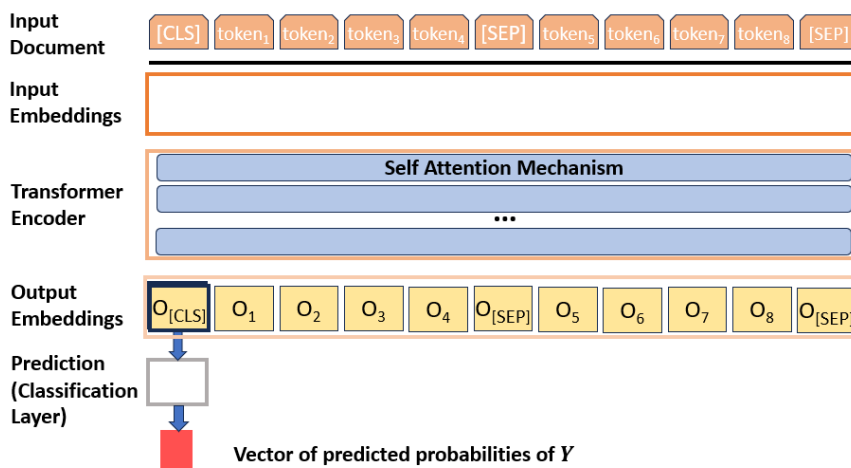
Figure 3.3 High-Level Architecture for the MLM Objective of the Pre-trained BERT.



Simply using the pre-trained model means providing a corpus to the network, which pre-processes the text documents, encodes them, and returns the output embeddings, without any further model training.

In our case, however, we fine-tuned the pre-trained Multilingual BERT. This involves adding fully-connected classification layers to the vector  $O_{[CLS]}$ , i.e. the vector representation of the review in our case, as shown in Figure 3.4. In this way, the model addresses a multi-classification task: given an input document  $d_i$  it tries to predict its value  $y_i$  of a *target* variable  $Y$ . In our application, where the reviews are the text documents,  $Y$  is represented by the attraction, and then it has four possible categories.

Figure 3.4 BERT High-Level Architecture for Fine-Tuning on Text Classification Tasks.



To test the model performance, we used the *validation set approach* as described in [105]: we randomly split the observations into a training set, for model training, and a validation set, on whose observations we use the trained model to predict the review's attraction.

In the training step, the model operates in *epochs*. During an epoch, the model goes through the entire training set, and performs a forward pass to make predictions and calculate a loss, which quantifies how far off the predictions are from the actual categories. Considering  $c = 1, \dots, 4$ , the categories of the target variable  $Y$ , the model computes the sum of *CrossEntropy* losses over the  $n_{tr}$  documents in the training set [106] [107], given by:

$$L = \sum_{i=1}^{n_{tr}} H_i = \sum_{i=1}^{n_{tr}} \left[ - \sum_{c=1}^4 I_{ic} \log \hat{p}_i(c) \right] \quad (3.1)$$

where  $H_i$  is the CrossEntropy loss for the  $i$ -th document,  $\hat{p}_i(c)$  is the predicted probability for the document  $i$  of being of category  $c$ , and  $I_{ic}$  is an indicator variable that equals 1, if  $y_i = c$ , and 0, otherwise.

After calculating  $L$ , the model performs a backward pass to adjust its internal parameters using the AdamW optimization algorithm [108].

At the end of each epoch, the model performance is further evaluated on the validation set, by comparing the predictions with the actual value of the target variable. If we consider  $\hat{I}_c$  indicating whether the prediction of the target variable is  $c$ , the *confusion matrix* between actual and predicted values of  $Y$  for the category  $c$  is given by:

	$\hat{I}_c = 1$	$\hat{I}_c = 0$
$I_c = 1$	True Positive (TP)	False Negative (FN)
$I_c = 0$	False Positive (FP)	True Negative (TN)

The performance metrics we used to compute the model at each epoch are *Precision*, *Recall*, and *F1-score* F1 [109], since they are highly prevalent in the literature [36] [110] [111] [101]. They are given by:

$$Precision = \frac{TP}{TP + FP} \quad (3.2)$$

$$Recall = TP / (TP + FN) \quad (3.3)$$

$$F1 = 2 * (Precision * Recall) / (Precision + Recall) \quad (3.4)$$

We took the average of those metrics over the four categories to compare model performance on the validation set in 30 epochs.

However, since the model is chosen from among 30 models evaluated on the validation set, the choice is influenced by the performance on this set, ultimately affecting the selection of the final model. To account for this, before the training-validation split, we randomly select 20 reviews for each attraction, which we refer to as the *assessment* set. This set is used to gather additional performance statistics entirely separate from the process of estimating the model's final parameters.

### 3.3.2 Extracting group-specific keywords

To further validate the model, we employed count-based methods to extract group-specific keywords from the groups of vectors potentially identified in the semantic space, i.e. among vectors of document embeddings.

The objectives are as follows:

1. Identifying groups of reviews in the validation set based on their embeddings.

2. Extracting group-specific keywords and determine if they have a direct interpretation in the domain of cultural attractions in the city of Brescia.

First, to detect groups of vectors, we utilized a clustering algorithm. We aim to extract keywords from well-defined groups, avoiding the inclusion of words from "noisy" reviews in the extraction process. For this purpose, we employed the *Hierarchical Density-Based Clustering of Applications with Noise* (HDBSCAN) algorithm [112] [113], a hierarchical density-based clustering algorithm that is robust against noise.

To extract cluster-specific keywords, we considered a method based on the boolean frequency of a term in text documents. If  $T$  is the binary variable for the term presence in a document, and  $G$  is the binary variable indicating whether a document belongs to a certain cluster or not, we have the following contingency table:

	T = 0	T = 1	
G = 1	$\alpha$	$\beta$	
G = 0	$\gamma$	$\delta$	

where  $\alpha, \beta, \gamma, \delta$  indicate the absolute joint frequencies.

If  $g$  and  $t$  are generic values assumed by  $G$  and  $T$ , respectively, then the *Chi-square statistic*  $\chi^2$  results [19]:

$$\chi^2 = \sum_{g=0}^1 \sum_{t=0}^1 \frac{(n_{gt} - \hat{n}_{gt})^2}{\hat{n}_{gt}} \quad (3.5)$$

where  $n_{gt}$  is the absolute frequency for the combination  $G = g, T = t$ , and  $\hat{n}_{gt}$  is the expected frequency under the assumption of stochastic independence between  $G$  and  $T$ .

Then the *Keyness*  $K$  operates the following transformation on the  $\chi^2$  statistic [114]:

$$K = \begin{cases} -\chi^2, & \text{if } \alpha < \gamma \\ 0, & \text{if } \alpha = \gamma \\ \chi^2, & \text{if } \alpha > \gamma \end{cases} \quad (3.6)$$

Therefore,  $K$  allows for discriminating between positive association between a cluster and a word and negative association. To extract the cluster-specific keywords, for each cluster, we compute  $K$  between the cluster and all the words in the vocabulary and then rank those words based on the value of  $K$ .

### 3.4 Data

The available reviews are distributed across the city's four main cultural attractions, all managed by Fondazione Brescia Musei: the *Castle*, the picture gallery *Pinacoteca Tosio-Martinengo*, the archeological site of *Roman Brixia*, and the *Santa Giulia Museum*.

The study has so far focused on Italian reviews, but it will be extended to all other languages, fully harnessing the potential of Multilingual BERT. To detect the reviews in the Italian language, we used two versions of Google's *Compact Language Detector Algorithm* (CLD), i.e. CLD2 [80] and CLD3 [81], and retained reviews for which both the algorithms estimated an Italian-written text, resulting in a total of 5753 reviews.

As already mentioned, we randomly sampled 20 reviews for each attraction to create an assessment set, and then randomly split the remaining reviews into training and validation sets. 70% of reviews have been used for model training.

Table 3.1 displays the distribution of attractions in the training and validation sets. It highlights the high imbalance among the attractions in the training set. Therefore, the model trained on this set was compared against a balanced set achieved through *oversampling*: reviews from the minority categories were resampled with replacement to match the number of observations of the majority category, i.e., *Castle* with 2,658 reviews. The oversampling thus returned a balanced training set of 10,632 reviews.

**Table 3.1 Distribution of attractions in the training and validation sets.**

Attraction	Training set		Validation set	
	$n_{tr}$	(%)	$n_{val}$	(%)
<i>Castle</i>	2658	58.6	667	58.8
<i>Santa Giulia Museum</i>	963	21.2	245	21.6
<i>Roman Brixia</i>	728	16.0	176	15.5
<i>Pinacoteca Tosio Martinengo</i>	189	4.2	47	4.1
Total observations	4538		1135	

## 3.5 Results

### 3.5.1 The model performance

The multiclassification model has been trained for 30 epochs, and then, based on the performance metrics on the validation set, the best model is selected out of the 30. Due to the imbalance in the distribution of attractions, the same methodology was applied using a balanced training set of 10,632 reviews. Therefore, the entire procedure resulted in two best models, i.e., one trained on the unbalanced set and one on the oversampled one. Figure 3.2 shows their performance on the validation and assessment sets. Precision, Recall, and F1 have been averaged over the 4 categories of the target variable.

These two models perform very similarly on the validation set. For this reason, in the following analyses, we will consider the best model obtained on the unbalanced set to take advantage of using the original data. Figure 3.5 shows the progression of the average Precision, Recall, and F1-score across epochs. The best model out of the 30 is the last one. However, it's unlikely that the training was stopped at an early stage because there is no clear increasing trend.

Furthermore, Table 3.2 illustrates that the same performance achieved on the validation set is maintained on the assessment set, which remained separate from both the training and best-model selection phases.

The model exhibits good performance, given the context of multiclassification with four categories.

### 3.5.2 Vector representations of reviews

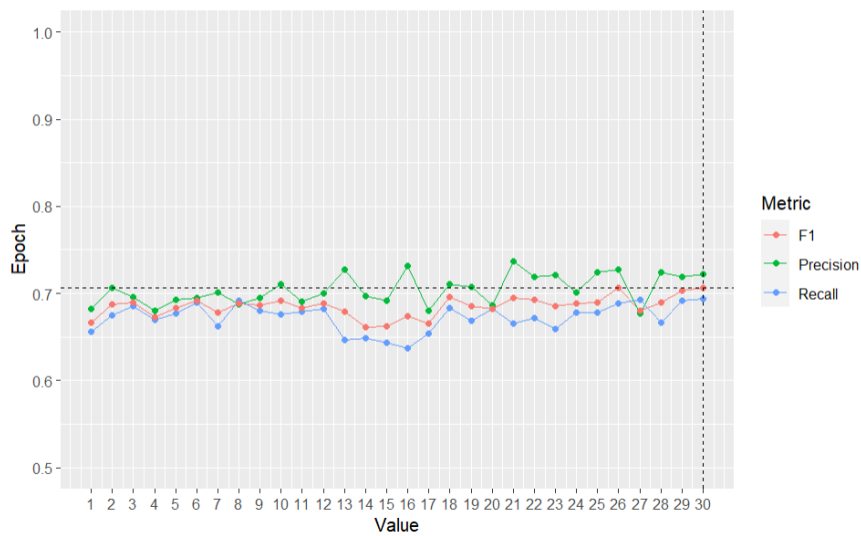
A two-dimensional visualization of the 768-dimensional vector of reviews is used here to highlight the difference between the representations of reviews obtained



**Table 3.2 Performance metrics (averaged over the categories of the target variable) on validation and assessment set; best models obtained on unbalanced training set and oversampled training set.**

Dataset	Observations	Training set	Precision	Recall	F1
Validation	1135	Unbalanced	0.722	0.694	0.707
		Oversampled	0.730	0.701	0.713
Assessment	80	Unbalanced	0.746	0.713	0.714
		Oversampled	0.739	0.713	0.717

**Figure 3.5 Performance metrics (averaged over the categories of the target variable) on validation set by epochs (model trained on unbalanced set).**

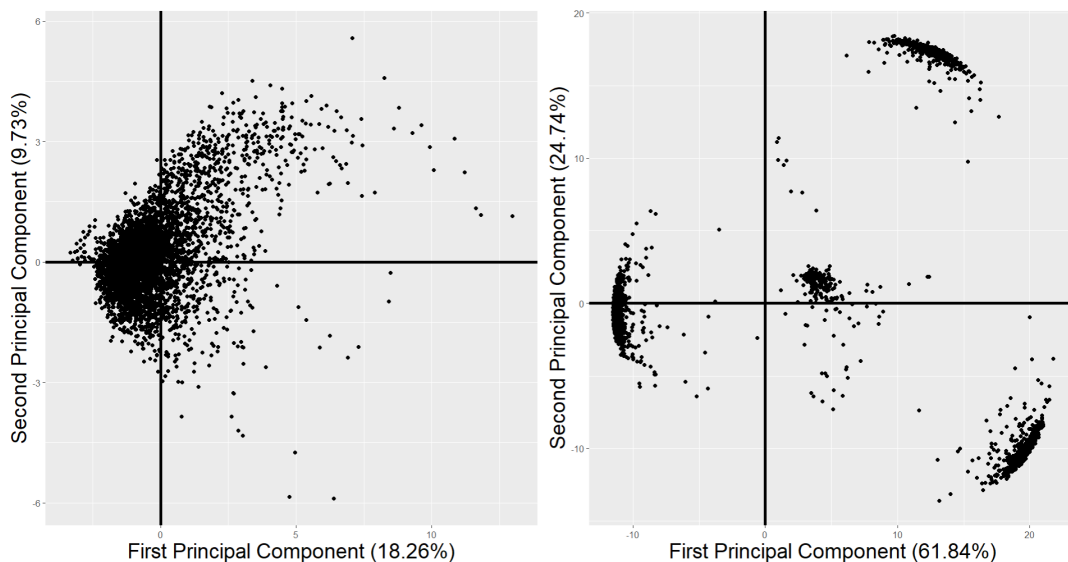


without fine-tuning BERT and those after fine-tuning. This comparison is shown in Figure 3.6, by means of the plot of the vector representations of reviews on the subspace defined by the first two *Principal Components* (PCs).

The two graphs show that the first two PCs obtained from the embeddings before fine-tuning explain less than 30% of the total variance, which is much less variability compared to what they capture with fine-tuned embeddings (approximately 87% of the total variance). Moreover, fine-tuning allows a clear separation of reviews into four distinct clusters.

A similar structure is also apparent for the validation and assessment reviews (Figure 3.7), where their embeddings are projected into the 2-dimensional subspace defined by the first PCs computed on the training embeddings.

**Figure 3.6** First two principal components of embeddings of training reviews with no fine-tuning (left) and after fine-tuning (right); percentage of variance explained by PCs in parenthesis on the axes.



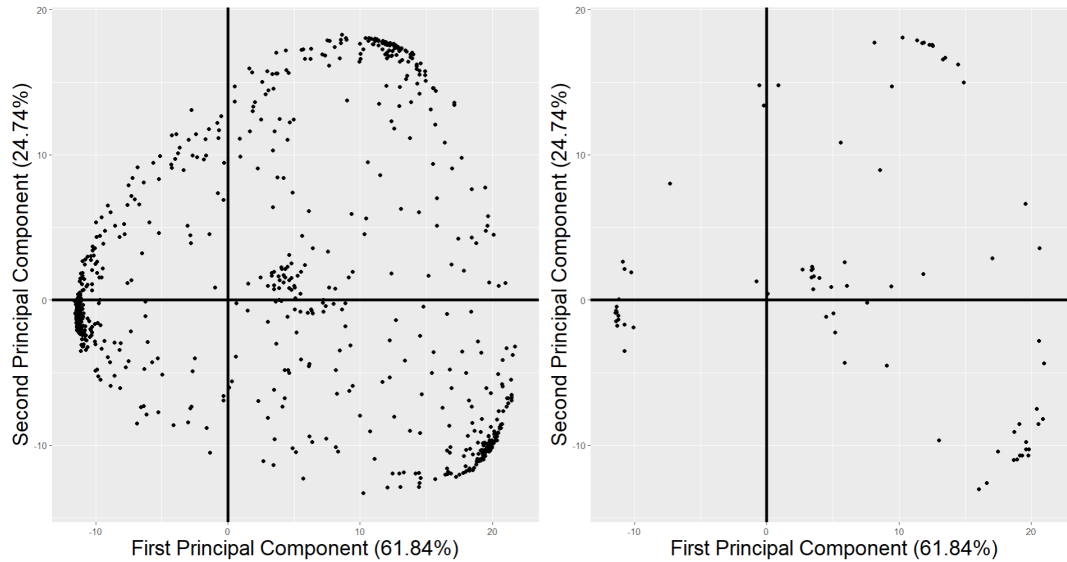
### 3.5.3 Cluster-specific keywords

To further validate the model, we employed the *Keyness*  $K$ , i.e. a transformation of the  $\chi^2$  statistic, to extract group-specific keywords from the clusters of vectors identified by the point clouds in the vector space. By means of  $K$ , we computed the association between a cluster and all the words in the vocabulary. Therefore, each word has a cluster-specific rank based on its value of  $K$ .

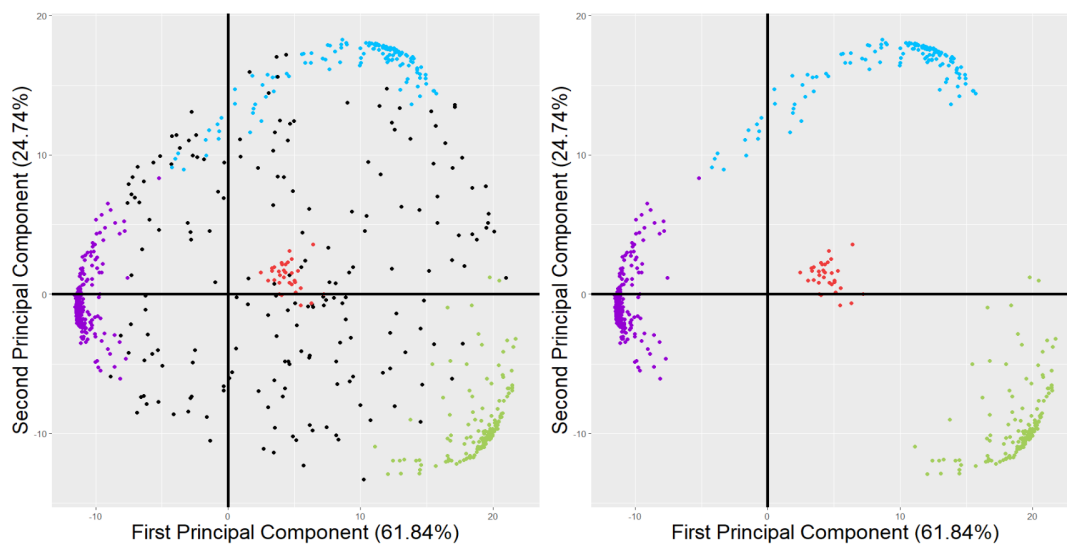
The first ten PCs calculated on the training embeddings capture more than 99% of the total variance. Therefore, we projected the validation set embeddings into the 10-dimensional subspace defined by these PCs.

The application of the HDBSCAN clustering algorithm on those 10-dimensional vectors found 4 groups of reviews, enabling the removal "noisy" reviews. Figure 3.8 highlights both the existence of clusters and that of 'noisy' reviews.

**Figure 3.7** First two principal components of embeddings of validation reviews (left) and assessment reviews (right), after fine-tuning; PCs computed on embeddings of training set (variance explained by PCs in parenthesis on the axes).



**Figure 3.8** Two-dimensional representation of clusters of embeddings of validation reviews obtained with HDBSCAN before (left) and after noise removal (right); clustering algorithm applied on 10 PCs; PCs computed on embeddings of training set (variance explained by PCs in parenthesis on the axes).



Then, we collected documents into clusters, retaining only the nouns from each document. To tag each word based on its part of speech, we utilized the *Udpipe* model<sup>13</sup> [115].

Figure 3.9 shows the top 10 ranking of words within clusters based on their value of *Keyness K*. It highlights the words with the strongest association with a cluster, helping in the identification of the cluster's content.

The first word in the ranking identifies one of the four attractions: the *picture gallery*, *S. Giulia museum*, the *castle*, and the *temple*, most likely referring to the famous Roman temple at the archaeological site of the Roman Brixia.

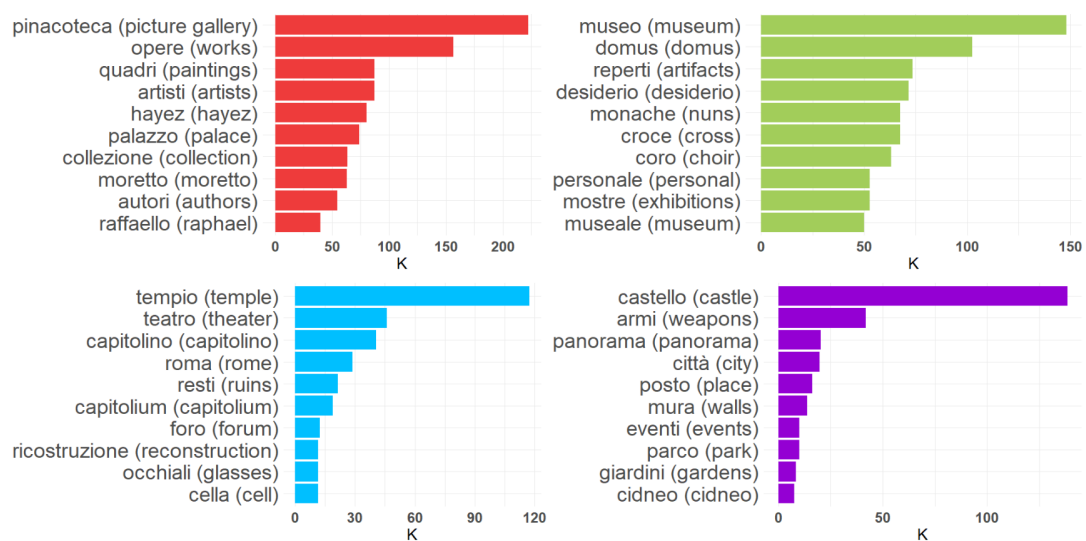
Regarding of the temple, among the words in the bottom-left cluster, there are also *capitolino* or *capitolium*, all names of the Roman temple whose *ruins* are at the roman *site*. It's interesting to note the term *glasses*, which perhaps refers to the ART-GLASS initiative, an augmented reality experience at the site offered by the Fondazione Brescia Musei since 2015.

In the *picture gallery* cluster in the top-left graph, names related to *paintings* stand out, such as the names of painters like *Hayez*, *Raphael*, and *Moretto*, the latter being an important Renaissance artist from the city of Brescia.

Many words evoke the Santa Giulia Museum in the cluster in the top-right part of the graph. The museum is housed in an ancient monastery, with treasures once kept by the *nuns*, including the gem-studded cross of King *Desiderio*. Furthermore, the museum offers a collection of archaeological *artifacts* and a visit to the Roman *domus*.

In the cluster related to the castle, there are very generic words like *place* and *city* that need further investigation. However, there are mainly words associated with the attraction. The castle is surrounded by medieval *walls*, within which there are *parks* and *gardens*, as well as a *weapons* museum dating back over a long period from the 15th to the 19th century. Furthermore, it should be mentioned that the castle is perched on the hill called *Cidneo*, from which a *panorama* of the entire city is visible.

**Figure 3.9 Top 10 by-cluster ranking of nouns based on the value of the *Keyness K* (English translations of nouns in parenthesis).**



<sup>13</sup> Available at <https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3131>

## 3.6 Conclusions and discussion

In this chapter, a research project is presented within the context of the "Data Science for Brescia - Arts and Cultural Places" Project, whose primary goal is to enhance our comprehension of how individuals engage with cultural sites.

The purpose of the research here is to build a language model on the Italian-written online reviews to classify them into four distinct semantic *areas*, defined by the main four attractions of the city of Brescia, in Italy.

The great utility of such a model stems from the fact that it can be used to identify the attraction in text documents, such as posts on social media platforms, forums, and other online spaces, when the attraction is not explicitly mentioned in the document metadata. Therefore, the proposed model can process and classify those texts, allowing for the expansion of the online discourse database concerning those cultural attractions.

The *Multilingual* BERT model has been fine-tuned on this multiclassification task, and it yielded very good performance results on the validation set, i.e. the set that was not used for training. Given the multilingual capabilities of the BERT model, future developments of this study will involve expanding the dataset with reviews written in languages other than Italian.

To further validate the model, clusters of reviews from the validation set have been detected, based on their vector representations produced by the model. The count-based method of the *Keyness*  $K$  has been employed to extract cluster-specific keywords, and they have proven to be highly consistent with the domain of application, i.e., with the characteristics and offerings of the four cultural attractions.

Regarding the value of  $K$  for a word, interesting research questions arise. For instance, one may wonder if the  $K$  value for the same word is comparable across clusters and what role the cluster size plays in this context. On the other hand, the comparability of  $K$  values for different words within clusters should be investigated, along with the role of word frequencies in this comparison. Chapter 4 will address these research questions through a simulation study that analyzes the ability of well-known count-based methods to capture the association between a group of documents and a word.



## Chapter 4

# Group-specific term estimators: a simulation study

### 4.1 Introduction

In Chapter 1 the notions of *exhaustivity* of a document description and *specificity* of a term in the vocabulary have been introduced. While the former is the degree to which a document description is able to cover the various concepts expressed by the terms occurring in a document, the latter indicates the level of detail to which a term describes a given concept [3] [4]. With a count-based approach, [5] and [6] proposed document-level and corpus-level statistics to quantify those factors. From that perspective, exhaustivity relates to counting the occurrences of a term in a document, while specificity involves counting the number of documents in which the term appears.

Here we introduce the *group-specificity* (GS) of a term. We consider a corpus of grouped text documents, and we assume that not all the concepts are evenly distributed across the groups, but rather there are concepts that groups do not share. Therefore, using a count-based approach, group-specificity involves counting the number of documents, for a certain group, in which the term appears.

The subject of study in this chapter is the various methods proposed in the literature to calculate the group-specificity of a term, which we refer to as *group-specificity* estimators.

These estimators serve multiple purposes. On one hand, in Vector Space Models (VSM), group-specificity proves useful for the development of term-weighting schemes that, when calculating the weight of a term for a document, also consider the document's affiliation to a group. This was discussed in Section 1.1.

On the other hand, defining an appropriate method for calculating the group-specificity of a term is essential in cases where, given grouped documents, one aims to extract keywords to explore the concepts that distinguish each group. In Chapter 2, the use of *Information Gain* and *Modified Relevance Frequency* demonstrated that words related to the domain of stock markets differentiate professionals from non-professional users when they self-describe on the StockTwits platform. In Chapter 3, *Keyness* has been employed to extract group-specific keywords from clusters of reviews about cultural attractions, revealing that each cluster discusses the cultural offering and features of a different attraction.

However, when employing these estimators, there may be a temptation to utilize them as a measure of the intensity of the association between a word and a group, to compare the association of one word with multiple groups, or to compare the association of multiple words with the same group. Through this study, we have explored whether the corpus size, group imbalance, and the probability of a word appearing

in a document play a role in determining the admissibility of using estimators in these manners.

Furthermore, this study considers the setting in which two document variables are available in a corpus: for instance, an *outcome* variable and a *group* variable. In this case, intriguing connections revolve around the pairwise relationships among words, group variable, and outcome variable. Therefore, another aspect addressed in this study regarding estimators is their ability to capture the outcome-word relationship if they are computed on the group-word interaction.

To the best of our knowledge, there is a lack of studies investigating the specific context in which both group and outcome variables are available as document variables, despite its common occurrence. As an example, consider the analysis of politicians' opinions on Twitter concerning a particular topic: typically, the outcome variable is represented by a sentiment label, such as hate/non-hate, and the politicians' political affiliation serves as the group variable.

Other examples come from the previous chapters of this thesis. With the data from Chapter 2, for instance, we could use users' experience as the group variable and the approach in trading as the outcome variable. In the case of the reviews discussed in Chapter 3, apart from employing the group variable that categorizes reviews by cultural attractions, we could utilize the review ratings to construct an outcome variable distinguishing positive reviews from negative ones.

In conclusion, grouped text documents with the additional information of an outcome variable are the object of this study; a review focuses on the *supervised* GS measures, i.e. measures that consider the group information in term-weighting, and based on boolean term-frequency, i.e. on the presence/absence of a term in a document. Moreover, the analysis is restricted to the *class-specific* GS measures, i.e. measures that assume two different values: one value for the documents belonging to a specific group (*target* group), and another one for the remaining documents (in the *reference* group).

Our first purpose is to frame this problem with a statistical approach. This allows us: (1) to interpret several GS measures as estimators of population quantities, (2) to define the simulation study proposed in this Chapter <sup>1</sup>.

This Chapter is structured as follows. Section 4.2 will illustrate the methodology; specifically, Section 4.2.1 will show how we chose to model the corpus of documents in order to draw samples of documents with certain characteristics, and Section 4.2.3 will follow that model to review and reframe well-known GS estimators. Section 4.3 will describe the design of the proposed study. Section 4.4 will provide some visualizations useful for exploring the simulations and will also highlight some examples of scenarios to demonstrate their interpretation. Section 4.5 will describe the features of a web application, created to enable free exploration of the simulation results.

## 4.2 Methods

A simulation study is proposed to assess the performance of GS estimators in various scenarios, e.g. a low-frequency term in a small corpus made of imbalanced groups, and comparing them against a ground truth, i.e. a measure of the true association between a term and a group of text documents. This entails making choices regarding how to model a corpus, by focusing solely on relevant information, and thus, how to generate corpus examples in specific scenarios.

---

<sup>1</sup>A preliminary version of this study is registered in the Proceedings of the 5th International Conference on Advanced Research Methods and Analytics, in Sevilla (Spain) [116].



The proposed methodology (1) models a corpus of grouped text documents as a sample coming from a *Multivariate Binomial* distribution, and (2) exploits the Gaussian copula method to simulate corpora with specific characteristics.

#### 4.2.1 Multivariate Bernoulli sampling with Gaussian copula

The class of GS measures we consider leverages only two pieces of information from a document: whether the document includes a term, and its outcome label; as mentioned, we also look at the group to which the document belongs.

We consider the following random variables:  $Y$  is the binary outcome variable,  $G$  is the binary variable for group membership, i.e. the group variable, and  $T$  indicates whether a document includes the term or not. Thus, we describe a document as a multivariate Bernoulli trial, whose result is a realization of the random variable  $(Y, G, T) \sim B_3(1, p_Y, p_G, p_T, \theta)$ , i.e. the variable distributed as a *Multivariate Bernoulli* distribution, where 1 is the number of trials (documents),  $p_Y, p_G$ , and  $p_T$  are the success probabilities of the marginals  $Y, G$ , and  $T$ , respectively, and  $\theta$  is the vector of the four joint probabilities  $p_{YG}, p_{YT}, p_{GT}$ , and  $p_{YGT}$ , where, for instance,  $p_{YG} = P(Y = 1, G = 1)$  [117].

To obtain samples from a Multivariate Bernoulli distribution we propose to use the Gaussian Copula Method [118].

Here we illustrate the method:

1. We start by considering the random vector  $\mathbf{X} = (X_Y, X_G, X_T)$  distributed as a Multivariate Normal distribution, where:

$$\mathbf{X} \sim N(0, \mathbf{P})$$

$$\mathbf{P} = \begin{bmatrix} 1 & \rho_{X_Y, X_G} & \rho_{X_Y, X_T} \\ \rho_{X_Y, X_G} & 1 & \rho_{X_G, X_T} \\ \rho_{X_Y, X_T} & \rho_{X_G, X_T} & 1 \end{bmatrix}$$

where for the linear correlation coefficients we use Pearson's  $\rho$ ;

2. We define three continuous Uniform random variables as cumulative distribution functions of the three aforementioned random variables, i.e.  $U_i := F_i(X_i)$ ,  $U_i \sim U(0, 1)$ ,  $i = Y, G, T$ . The *copula* function is defined as [119] [120]:

$$\begin{aligned} C(u_Y, u_G, u_T) &= P(U_Y \leq u_Y, U_G \leq u_G, U_T \leq u_T) = \\ &= P(X_Y \leq \Phi^{-1}(u_Y), X_G \leq \Phi^{-1}(u_G), X_T \leq \Phi^{-1}(u_T)) = \\ &= \Phi(\Phi^{-1}(u_Y), \Phi^{-1}(u_G), \Phi^{-1}(u_T); \mathbf{P}) \end{aligned}$$

where  $\Phi$  is the CDF of a standard normal distribution, and  $\Phi(\cdot; \mathbf{P})$  is the joint CDF of  $\mathbf{X}$ ;

3. We set the thresholds  $u_Y^*, u_G^*, u_T^*$ , to generate  $Y, G, T$ :

$$\begin{aligned} Y &= 1 \text{ if } U_Y \leq u_Y^* \\ G &= 1 \text{ if } U_G \leq u_G^* \\ T &= 1 \text{ if } U_T \leq u_T^* \end{aligned}$$

Note: the thresholds are exactly the marginal probabilities  $p_Y, p_G, p_T$  [121] [122].

If we denote with  $N$  the number of i.i.d. documents<sup>2</sup>, the corpus of documents can be described by means of a Multivariate Binomial distribution with parameters  $N, p_Y, p_G, p_T, \theta$  [123]. Each document can be considered as a random draw from the Multivariate Bernoulli distribution by means of the Gaussian copula.

In Figure 4.1, an example of this process is graphically shown: with  $(\rho_{X_Y, X_G}; \rho_{X_Y, X_T}; \rho_{X_G, X_T}) = (0.9; 0.6; 0.25)$ , 2000 observations from a Multivariate Normal distribution are drawn (Figure 4.1a) and the copula function let us obtain samples from a Multivariate Uniform distribution respecting the set correlation structure (Figure 4.1b); we set the thresholds  $(u_Y^*, u_G^*, u_T^*) = (p_Y; p_G; p_T) = (0.7; 0.5; 0.2)$  on  $U_Y, U_G$ , and  $U_T$  to obtain a sample of 2000 documents from the Multivariate Binomial distribution, illustrated as contingency tables, in the form of heatmaps, in Figure 4.1c. The contingency tables give us information about the three pairwise associations between the outcome variable, the group variable, and the presence of a term.

The Figure shows that the intensity of the association between two Bernoulli variables depends both on its ground truth  $\rho$  and on the marginal probabilities of the two variables. For example, the true value of the intensity of the association between  $Y$  and  $G$  is measured by  $\rho_{X_Y, X_G}$ , and it is set very high, equal to 0.9. On the contingency table, on the one hand, the two most frequent combinations are those on the diagonal for which  $Y = G$ , but, on the other hand, the observations are particularly concentrated on the combination  $Y = G = 1$ , due to the unbalance of  $Y$ , whose marginal probability has been set at 0.7.

As for the association between document group and term, we can look at the last line of Figure 4.1. The true value of the intensity of the association between  $G$  and  $T$  is  $\rho_{X_G, X_T} = 0.2$ , small but positive. In addition, the probability of occurrence of the term  $p_T$  is 0.2. Thus, from the contingency table, we see that most observations are on combinations with  $T = 0$ , but still the term has a small attraction with the target group (i.e.  $G = 1$ ) and a repulsion with the reference group (i.e.  $G = 0$ ).

## 4.2.2 Parameters of the simulation

To summarize, the simulation implies setting the values of the following parameters:

- $n$ , i.e. the number of text documents in the corpus;
- $p_Y, p_G, p_T$ , i.e. a measure of potential imbalance of both the outcome variable and the groups, and the probability of a word appearing in a document, respectively;
- $\rho_{X_Y, X_G}, \rho_{X_Y, X_T}, \rho_{X_G, X_T}$ , i.e. the ground truth for the pairwise associations between the outcome variable  $Y$ , the group variable  $G$ , and the term presence  $T$ .

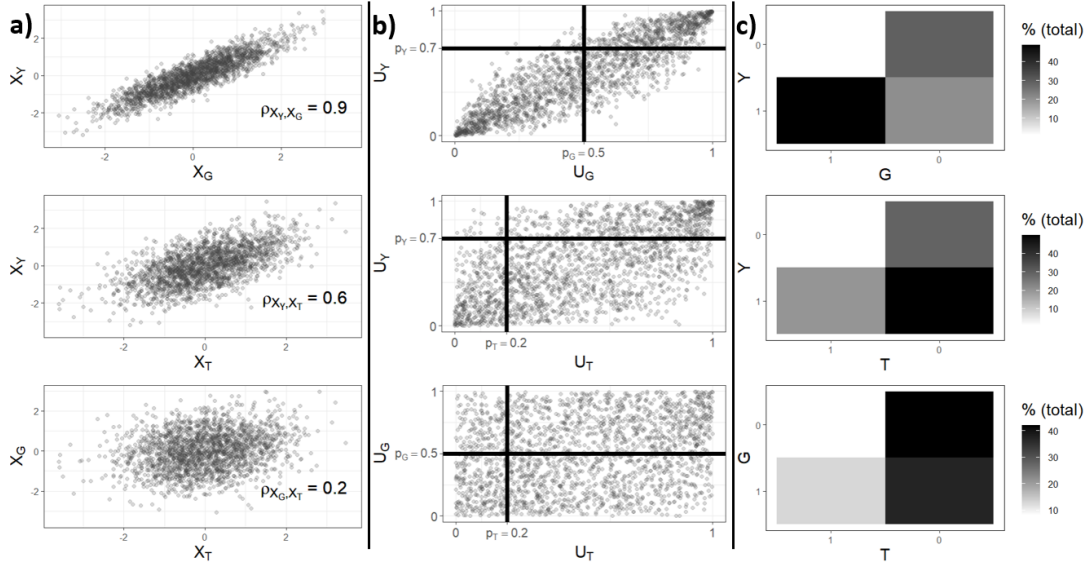
Lastly, each scenario is simulated for  $R$  repetitions, that is on  $R$  corpora, to obtain an estimate of the distribution of the GS estimators.

Although this sequence of steps does not precisely mirror the copula method, for the purpose of result analysis, a hierarchical approach is more convenient: the initial configuration of  $n, p_Y, p_G, p_T$  defines the scenario, while the correlation structure specifies the subscenario.

It should be mentioned that not all the combinations among the correlation coefficients are admissible. Indeed, if one sets two of the coefficients, e.g.  $\rho_{X_Y, X_G}$  and

<sup>2</sup>Assuming that documents are *independent* and *identically distributed* (i.i.d.) means assuming they are independently generated, and are all described as Bernoulli trials, with the same  $p_Y, p_G, p_T, \theta$ .

**Figure 4.1 Illustration of the Gaussian Copula Method to draw 2000 samples from a Multivariate Bernoulli distribution, with  $(\rho_{X_Y, X_G}; \rho_{X_Y, X_T}; \rho_{X_G, X_T}) = (0.9; 0.6; 0.25)$ , and  $(u_Y^*, u_G^*, u_T^*) = (p_Y; p_G; p_T) = (0.7; 0.5; 0.2)$ .**



$\rho_{X_Y, X_T}$ , then the value of  $\rho_{X_G, X_T}$  is constrained to satisfy the following<sup>3</sup>:

$$|\rho_{X_G, X_T} - \rho_{X_Y, X_G} * \rho_{X_Y, X_T}| \leq \sqrt{(1 - \rho_{X_Y, X_G}^2)(1 - \rho_{X_Y, X_T}^2)} \quad (4.1)$$

### 4.2.3 Analytical formulation of GS measures

Here we illustrate the considered supervised class-specific GS measures based on boolean term-frequency.

In each scenario, they are computed by starting from the joint distributions  $f_{YT}$  and  $f_{GT}$ , described by the following  $2 \times 2$  contingency tables:

	T = 1	T = 0	
G = 1	$\alpha$	$\beta$	$\alpha + \beta$
G = 0	$\gamma$	$\delta$	$\gamma + \delta$
	$\alpha + \gamma$	$\beta + \delta$	N

	T = 1	T = 0	
Y = 1	$a$	$b$	$a + b$
Y = 0	$c$	$d$	$c + d$
	$a + c$	$b + d$	N

The evaluation of the estimators consists in (1) comparing their value on  $f_{GT}$  with the ground truth  $\rho_{X_G, X_T}$ <sup>4</sup>, and (2) comparing their values on  $f_{GT}$  and  $f_{YT}$  by varying the ground truth  $\rho_{X_Y, X_G}$ .

Since the calculations are equivalent for  $f_{GT}$  and  $f_{YT}$ , and specular for  $G = 1$  (target group) and for  $G = 0$  (reference group), only the ones for  $f_{GT}$  and the target group are shown.

<sup>3</sup>The constraint ensures that the partial correlation between  $X_G$  and  $X_T$ , while controlling for  $X_Y$ , ranges from -1 to 1 [124].

<sup>4</sup>Notice that it is equivalent to comparing their value on  $f_{YT}$  with the ground truth  $\rho_{X_Y, X_T}$

The first GS measure we present is based on the the well-known *Inverse Document Frequency* (IDF) given by:

$$\text{IDF} = \log_2 [P(T = 1)]^{-1} = -\log_2 [P(T = 1)] = -\log_2 \left( \frac{\alpha + \gamma}{N} \right) \quad (4.2)$$

which, in computing the weight of the word for a document, sets a penalty based on how widespread a word is across documents; the higher probability for  $T = 1$ , the higher penalty [5].

[76] proposed a modified version of IDF, by limiting the penalty based on how frequent a word is in the reference group. For this rationale, we rename it *Inverse Document Frequency on the Reference group* IDFR; it ranges in  $[0, +\infty]$ , and is given by:

$$\begin{aligned} \text{IDFR} &= \log_2 [P(T = 1|G = 0)]^{-1} = -\log_2 [P(T = 1|G = 0)] \approx \\ &\approx -\log_2 \left( \frac{\gamma + 1}{\gamma + \delta + 1} \right) \end{aligned} \quad (4.3)$$

If there is no relation between  $G$  and  $T$ , IDFR equals the IDF and, with  $P(T)$  going towards 1, they both tend to 0.

The *Pointwise Mutual Information* PMI for  $G = 1$  and  $T = 1$  is defined as:

$$\text{PMI} = \log_2 \frac{P(G = 1, T = 1)}{P(G)P(T)} \approx \log_2 \frac{N(\alpha + 1)}{(\alpha + \beta)(\alpha + \gamma)} \quad (4.4)$$

and ranges from  $-\infty$  to  $+\infty$ . When  $G$  and  $T$  are independent, and so when  $P(T)$  tends to 1, it equals 0.

Note that PMI can also be written as:

$$\begin{aligned} \text{PMI} &= \log_2 \frac{P(G|T)}{P(G)} = \log_2 P(G|T) - \log_2 P(G) \\ \text{PMI} &= \log_2 \frac{P(T|G)}{P(T)} = \log_2 P(T|G) - \log_2 P(T) \end{aligned}$$

which suggests that PMI is highly influenced by the marginal probabilities [72]; therefore:

- For groups with an equal conditional probability, the word will have higher weight for smaller groups than for more numerous groups;
- For terms with an equal conditional probability, the word will have higher weight for rarer terms than for common terms.

In the context of our simulation, it means that, with equal ground truth  $\rho_{X_G, X_T}$ , for smaller groups or rarer terms PMI assumes higher values.

To illustrate the following measures, we first introduce the following odds:

$$\text{Odds}(G = 1|T = 1) = \frac{P(G = 1|T = 1)}{P(G = 0|T = 1)} = \frac{\alpha}{\gamma} \quad (4.5)$$

Starting from this object, we detect some existing measures related to it, e.g. the *Relevance Frequency* RF [75]:

$$\text{RF} = \log_2 \left( 2 + \frac{\alpha}{\max(1, \gamma)} \right) \quad (4.6)$$

ranging in the interval  $[1, +\infty]$ . To make it symmetric, i.e. to make its absolute value the same for both categories of  $G$ , we propose the following modification, which we call *Symmetric Relevance Frequency* SRF:

$$\text{SRF} = \log_2 \left( \frac{\alpha + 1}{\gamma + 1} \right) \quad (4.7)$$

with  $G$  and  $T$  independent,  $\text{Odds}(G = 1|T) = \text{Odds}(G)$  and then both and RF only depend on  $P(G)$ : the word assumes higher weight for documents in the larger group.

Additionally, as a modification of RF, [76] proposed the *Relevance Frequency Ratio* RFR:

$$\text{RFR} = \log_2 \left[ 2 + \frac{(\alpha + 1)(\gamma + \delta + 1)}{(\gamma + 1)(\alpha + \beta + 1)} \right] \quad (4.8)$$

and rearranging it results:

$$\text{RFR} = \log_2 \left( 2 + \frac{\alpha + 1}{\alpha + \beta + 1} \frac{\gamma + \delta + 1}{\gamma + 1} \right) \approx \log_2 \left( 2 + \frac{P(T = 1|G = 1)}{P(T = 1|G = 0)} \right)$$

RFR is equal to  $\log_2(3)$  when  $G$  and  $T$  are independent.

### Feature-specific measures as class-specific

Besides class-specific measures, there are single-value measures giving information about the ability of a term in discriminating between the target and the reference group. Since they assume only value for a *feature*, i.e. a term in the vocabulary, we may define them as *feature-specific*.

Inspired by the proposal of [114] for the  $\chi^2$  statistic, the *Keyness*  $K$  transformation converts feature-specific measures into class-specific measures. Specifically, we show here transformations of the  $\chi^2$  statistic, *Information Gain* with the *Gini* index, and the *Information Gain* with *Entropy* index [72] [22] [73] [74]. Let us first recall their original formulation.

If  $g$  and  $t$  are generic values assumed by  $G$  and  $T$ , respectively, then the *Chi-square statistic*  $\chi^2$  results:

$$\chi^2 = \sum_{g=0}^1 \sum_{t=0}^1 \frac{(n_{gt} - \hat{n}_{gt})^2}{\hat{n}_{gt}} \quad (4.9)$$

where  $n_{gt}$  is the absolute frequency for the combination  $G = g, T = t$ , and  $\hat{n}_{gt}$  is the expected frequency under the assumption of stochastic independence between  $G$  and  $T$ .

For a binary variable  $G$ , the formula for the *Information Gain*  $IG$  obtained by knowing the presence or the absence of a term is:

$$\text{IG} = H(G) - P(T = 1)H(G|T = 1) - P(T = 0)H(G|T = 0) \quad (4.10)$$

where if for  $H$  the *Entropy* index is used:

$$\begin{aligned}
H(G) &= - \sum_{g=0}^1 P(G = g) \log_2 P(G = g) \\
H(G|T = 1) &= - \sum_{g=0}^1 P(G = g|T = 1) \log_2 P(G = g|T = 1) \\
H(G|T = 0) &= - \sum_{g=0}^1 P(G = g|T = 0) \log_2 P(G = g|T = 0)
\end{aligned} \tag{4.11}$$

Conversely, if the *Gini* index is used, we have:

$$\begin{aligned}
H(G) &= 1 - \sum_{g=0}^1 P(G = g)^2 \\
H(G|T = 1) &= 1 - \sum_{g=0}^1 P(G = g|T = 1)^2 \\
H(G|T = 0) &= 1 - \sum_{g=0}^1 P(G = g|T = 0)^2
\end{aligned} \tag{4.12}$$

We can now define the *Keyness*. Let us assume to have one of those feature-specific measures, generically denoted as  $h(G, T)$ . Then the *Keyness*  $K$  of a term for a target group results:

$$K = \begin{cases} -h(G, T), & \text{if Odds}(G=1 | T=1) < 1 \\ 0, & \text{if Odds}(G=1 | T=1) = 1 \\ h(G, T), & \text{if Odds}(G=1 | T=1) > 1 \end{cases} \tag{4.13}$$

that is equal to:

$$K = \text{where} \begin{cases} -h(G, T), & \text{if } \alpha < \gamma \\ 0, & \text{if } \alpha = \gamma \\ h(G, T), & \text{if } \alpha > \gamma \end{cases} \tag{4.14}$$

### 4.3 Study design and research questions

In this section, we describe the design of the proposed study, showcasing the selected parameter values and subgroup of the aforementioned estimators, and illustrating the research questions.

We detected scenarios of interest by combining meaningful values for the simulation parameters. Table 4.1 summarizes the selected parameter values. For  $n$ , we chose 500 and 2000, to simulate small and large corpora; for the probability of a document to be in the target group,  $p_G$ , the value of 0.5 lets us compute the estimators when the groups of documents are balanced, whereas the cases of unbalancing with 0.25 or 0.75 let us simultaneously monitor the value of a GS measure on both the target and the reference group; the same holds for the probability of outcome to be 1,  $p_Y$ ; the set 0.05, 0.1, and 0.3 for the probability of a term being in a document,  $p_T$ , is motivated by the need of simulating scenarios from a rare word to a very frequent word, such as a stopword. Limiting on these parameters, their values imply 54 scenarios.

We covered the theoretical range of Pearson’s linear correlation, i.e.  $[-1, +1]$ , in steps of 0.2, for  $\rho_{X_Y X_T}$  and  $\rho_{X_G X_T}$ , and in steps of 0.25 for  $\rho_{X_Y X_G}$ . As mentioned, some combinations of correlation values are not valid; an admissible correlation structure identifies a subscenario. Considering valid correlations and parameter values yields 32,562 subscenarios.

The number of repetitions  $R$  is 1000, i.e. in each subscenario we drew 1000 random corpora of documents.

**Table 4.1 Selected parameter values for simulation.**

$R$ (# of corpora)	$n$ (# of documents)	$p_Y, p_G$	$p_T$
1000	(500;2000)	(0.25;0.50;0.75)	(0.05;0.10;0.30)
$\rho_{X_Y X_T}, \rho_{X_G X_T}$		$\rho_{X_Y X_G}$	
from -1 to 1 with step 0.2		from -1 to 1 with step 0.25	

We selected a subset of the illustrated GS estimators and used them to show how to evaluate their behaviour in certain scenarios. A web application has been created to enable free exploration of the simulation results for all the GS estimators on several dashboards, and it will be described in detail in Section 4.5.

We selected the subgroup by looking at the magnitude of the correlations between their mean values across all scenarios. Table 4.2 shows the results of this analysis, where all the correlations are positive.

The table reports strong correlations between the means of IDFR, RF, SRF, and RFR. Therefore, we have chosen to present the results for IDFR and RFR, because the former originates from the well-known IDF, whose declension in the case of grouped documents is worth exploring, whereas the latter represents the latest modification proposed for RF.

Furthermore, the table shows strong correlations between the estimators of Keynes. The estimator calculated on  $\chi^2$  is widespread in the literature on Text Categorization and Dimensionality Reduction, as well as the one on IG with Entropy. We chose to use the former because it is the one proposed in [114] for calculating Keynes.

To summarize, we now show the results of the simulation on IDFR, RFR, PMI, and K with  $\chi^2$ . On the web application, the results regarding all the measures illustrated in Section 4.2.3 are available.

**Table 4.2 Correlation matrix between average GS estimates across simulations.**

	PMI	RF	SRF	RFR	K ( $\chi^2$ )	K (IG Entropy)	K (IG Gini)
IDFR	0.17	0.78	0.64	0.82	0.41	0.49	0.47
PMI		0.32	0.63	0.35	0.25	0.34	0.31
RF			0.88	0.89	0.40	0.51	0.47
SRF				0.80	0.46	0.58	0.54
RFR					0.52	0.62	0.60
K ( $\chi^2$ )						0.85	0.85
K (IG Entropy)							0.99

The proposed method makes it possible to address the following research questions:

- About the general tendency, *Does the estimator value account for the intensity of the true group-term association?* (Section 4.4.1)
- If two terms have a different association with the group, *what is the difference in their group specificity estimation?* (Section 4.4.2)
- About the role of  $p_G$ , *does the group unbalancing impact the estimator value?* (Section 4.4.3)
- About the role of  $p_T$ , *does the rarity of the term impact the estimator value?* (Section 4.4.4)
- When using the estimator to capture both outcome-term and group-term relationships, *what's the percentage difference in the estimator's value between the two?* (Section 4.4.5).

These research questions will be addressed by analyzing the means of the  $R$  estimates, i.e. the means of the  $R$  estimates obtained in a subscenario. Furthermore, the simulation provides an estimate of the estimator's distribution in each subscenario, which will be used to assess the reliability of the mean value, as usual.

Except for the last question, all comments are made considering only the association between  $G$  and  $T$ , and thus the relative calculations are made on the contingency table between the two.

However, in the case where  $\rho_{X_Y, X_G} = 1$  and  $p_Y = p_G$ , the estimates obtained on that contingency table are equivalent to those obtained on the table crossing  $Y$  and  $T$ , i.e. to study the outcome-term relation.

## 4.4 Results

### 4.4.1 General tendency

Figure 4.2 shows the mean values of the estimates (in red) and their distribution, by means of boxplots, across multiple subscenarios defined by values of the correlation between  $X_G$  and  $X_T$ ,  $\rho_{X_G, X_T}$  ( $\text{corr}(X_G, X_T)$  in the graph), i.e. the ground truth for the group-term association. The other parameters have been set as follows:  $n = 2000$ ,  $p_G = 0.5$ ,  $p_T = 0.1$ .

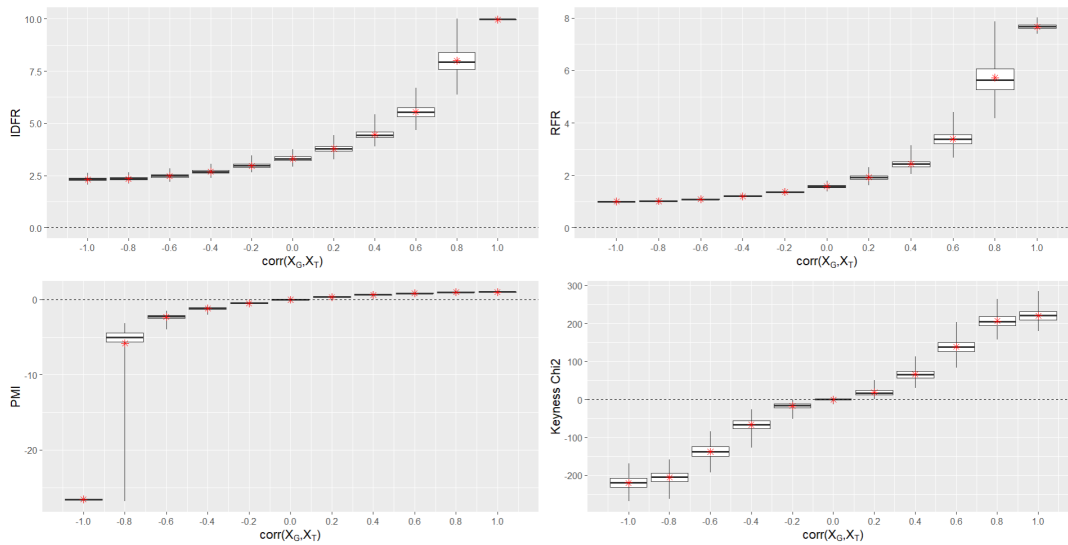
The tendency is that the estimation of the group-specificity of the term, as measured by the estimators, increases with the increase in the true group-term association for all estimators. Specifically, the estimation of the weight of the term for the target group has an increasing tendency.

When there is no association (i.e.  $\text{corr}(X_G, X_T) = 0$ ), both PMI and K are 0, while IDFR is  $-\log_2(p_T)$ , in this scenario, approximately  $-\log_2(0.1)$ , and RFR is  $\log_2(3) \approx 1.58$ . Hereafter we refer to the value of the estimator for which there is no group-term association as null value of the estimator.

The similar tendency in the plots of the estimates for IDFR and RFR confirms the correlation between the mean values of IDFR and RFR.



**Figure 4.2 Mean values (in red) and boxplots of the estimates, across multiple subsenarios; scenario parameter values:  $n = 2000$ ,  $p_G = 0.5$ ,  $p_T = 0.1$ .  $\text{corr}(X_G, X_T)$  stands for  $\rho^{X_G, X_T}$ .**



#### 4.4.2 Two terms with different group specificity

Figure 4.2 also enables us to compare the weights of two terms that only differ in their true association with a group. Let us consider two distinct cases: a slight difference and two opposite relationships.

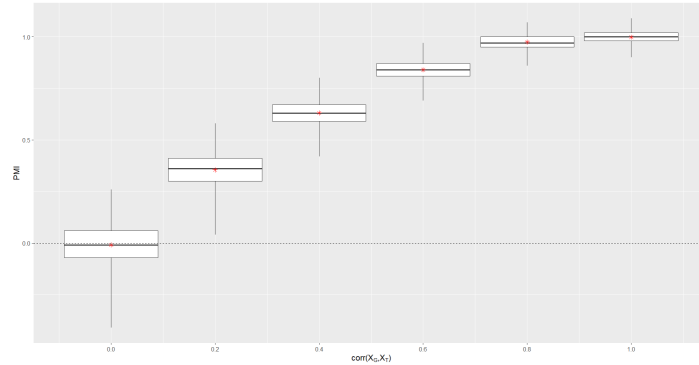
In the first case, we analyze the slope of the curve. In the scenario under consideration, let's suppose we have three terms, each with a relationship to the group measured by correlation coefficients of 0.4, 0.6, and 0.8, respectively. When we use PMI to estimate their weights, we observe that the mean values are quite similar to each other, consistently increasing as the strength of the association intensifies, as it can be more easily observed in Figure 4.3. However, when we employ IDFR or RFR estimators, the dynamics change: the difference between the mean estimate for the term with a correlation of 0.8 and the one with a correlation of 0.6 becomes more pronounced compared to the difference between the latter and the term with a correlation of 0.4.

One may further notice that the comments for PMI and K hold for those values of the correlations, but not for the entire correlation range.

The consequence of those comments for text mining practitioners is the following. In practice, the estimator is used to predict the weight of a term for the group. By comparing three terms in this scenario, one may comment that one term is particularly more relevant to the group than the other two. However, under the same conditions, these weights should only be used for ranking terms, not for commenting on the intensity of their relationships with a group of documents.

Furthermore, in this scenario, there is not much variability, but it increases for higher correlation values (IDFR and RFR) and extremes (K), whereas for positive values of PMI, it decreases for correlations close to zero, as seen in Figure 4.3. The difference in the reliability, albeit contained, of the means is another risk with commenting on the estimators as a measure of intensity and not only for ranking purposes.

**Figure 4.3** Mean values (in red) and boxplots of the estimates of PMI for positive values of correlations, across multiple subscenarios; scenario parameter values:  $n = 2000$ ,  $p_G = 0.5$ ,  $p_T = 0.1$ .



Secondly, we consider two terms with opposite group-term relationships, i.e. they share the same relationship intensity but with different groups.

The null value of PMI and K is 0. The null value of IDFR is  $-\log_2(p_T)$ , in this case, approximately  $-\log_2(0.1)$ ; the null value of RFR is  $\log_2(3) \approx 1.58$ . For all measures, above (below) the null value, we know that the word is more attracted to the target (reference) group. However,  $K$  is the only estimator that is symmetric around its null value; for  $K$ , in practice, if two words have the same value of  $K$ , but of opposite sign, we can say that they have the same association intensity but with different groups. This holds in this scenario, where variability for sub-scenarios is limited. Therefore, the means of the estimates suggest that there is the same intensity, with low variability, but with different groups. This argument does not apply to the other measures, which do not exhibit asymmetry around the null value.

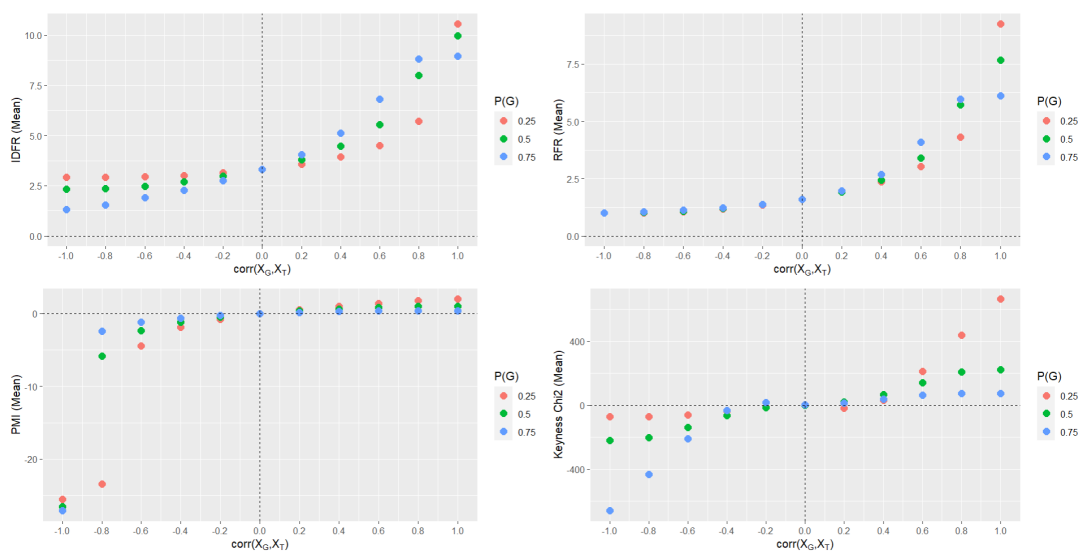
#### 4.4.3 Role of group unbalancing

In Figure 4.4, we explore the role of  $p_G$  in the estimation of term weights for groups within the scenario defined by  $n = 2000$ ,  $p_T = 0.1$ , and three values of  $p_G$ : 0.25, 0.5, and 0.75. Figure 4.6 (left) highlights the differences in estimated weights using PMI as  $p_G$  varies for positive values of the true group-term association, in the same scenario.

As previously mentioned, the selected values for  $p_G$  allow us to assess the estimated values for both the target and reference groups. Specifically, the combination  $(\rho_{X_G X_T}^*; p_G^*)$  provides the weight estimate for the target group on the curve, while the combination  $(-\rho_{X_G X_T}^*; 1 - p_G^*)$  provides the estimate for the reference group. Therefore, the curve for  $p_G = 0.5$  is sufficient for monitoring the weights of both the target and reference groups in the situation of balanced groups. In cases of group imbalance, this can be achieved by examining the  $p_G = 0.25$  and  $p_G = 0.75$  curves. For instance, if  $p_G^* = 0.25$  and  $\rho_{X_G X_T}^* = 0.8$ , we can find the weight estimate for the target group on the red curve, while the weight estimate of the term for the reference group can be found on the blue curve (where  $p_G = 0.75$ ) at  $-\rho_{X_G X_T}^* = -0.8$ .

Generally, weight estimation varies with changes in  $p_G$ , given the true intensity of the group-term relationship. For IDFR, as associations become stronger (both positive and negative), the effect of  $p_G$  becomes more pronounced. This also holds for  $K$  and PMI, but with reversed values of  $p_G$ . For example, in cases of attraction between the target group and the term, the larger the target group, the lower the

**Figure 4.4 Mean values of the estimates under the scenarios defined by  $n = 2000$ ,  $p_T = 0.1$  and three values of  $p_G$ , across multiple subscenarios.**



estimated weight by K and PMI, and conversely, the lower the estimated weight by IDFR.

When working with text data, this entails the following drawback. If the document variables  $G_1$  and  $G_2$  separately divide the corpus into groups, the weight estimates of the same word for the two target groups cannot be compared if the groups have different sizes. In other words, with the same associations, there are two different weights if  $p_{G_1} \neq p_{G_2}$ .

For RFR, there are two distinct cases. When there is repulsion between the target group and the term, there is no influence of the target group's probability (at least as long as the proportions of target and reference groups remain around the proposed values). RFR, on the other hand, resembles IDFR in this sense in cases of attraction between the target group and the term.

We compare the ratio between the target and reference weights as  $p_G$  varies.

For IDFR, K, and PMI, the relationship reverses beyond their null value. For IDFR, when the target group is less numerous, the difference between target and reference weights flattens (this is the interpretation of the lower slope). This also holds conversely for PMI and K. It means that the difference in weight estimates for the two groups is due to the group proportions. So, it confirms here as well: estimated weights can only be used for ranking, in the scenario considered.

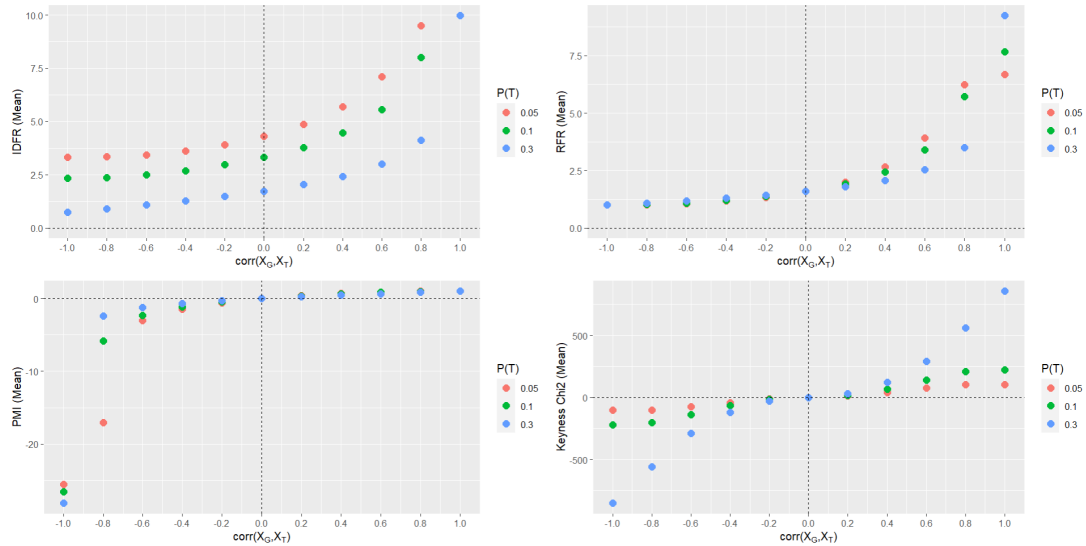
In contrast, for RFR in different cases, i.e., as  $p_G$  varies, if there is attraction, then there is a difference between target weights (for higher values of true association), but not between reference group weights. Note that it can be read in reverse: if there is repulsion, there is no difference between target weights as  $p_G$  varies, but there is for reference group weights.

#### 4.4.4 Role of term rarity

The following kind of evaluation is useful to answer the question: *can the weights of two words with different occurrence probability (e.g. such as a rare word and a stopword) be compared?*

Figure 4.5 allows us to explore the role of  $p_T$  in the estimation of term weights. Figure 4.6 (right) highlights the variation in estimated weights using PMI as  $p_T$  changes for positive values of the true group-term association.

**Figure 4.5 Mean values of the estimates under the scenarios defined by  $n = 2000$ ,  $p_G = 0.5$  and three values of  $p_T$ , across multiple subscenarios.**



The estimators exhibit limitations when applied to terms with different probabilities of appearing in a document. Indeed, in such a case, these measures cannot effectively serve as a basis for ranking. This limitation arises because, even when these terms share the same true association with the target group, their estimated weights differ due to their distinct probabilities of occurrence. This limitation becomes increasingly concerning as the difference in occurrence probability grows.

Moreover, the considered measures would also yield different rankings. Indeed, there are some differences between these four measures, in this perspective.

First, for IDFR, and RFR and PMI for positive correlations, a higher probability of term occurrence corresponds to a lower weight estimate. This means that terms with higher probabilities are given lower weights by these measures. The opposite is true for K.

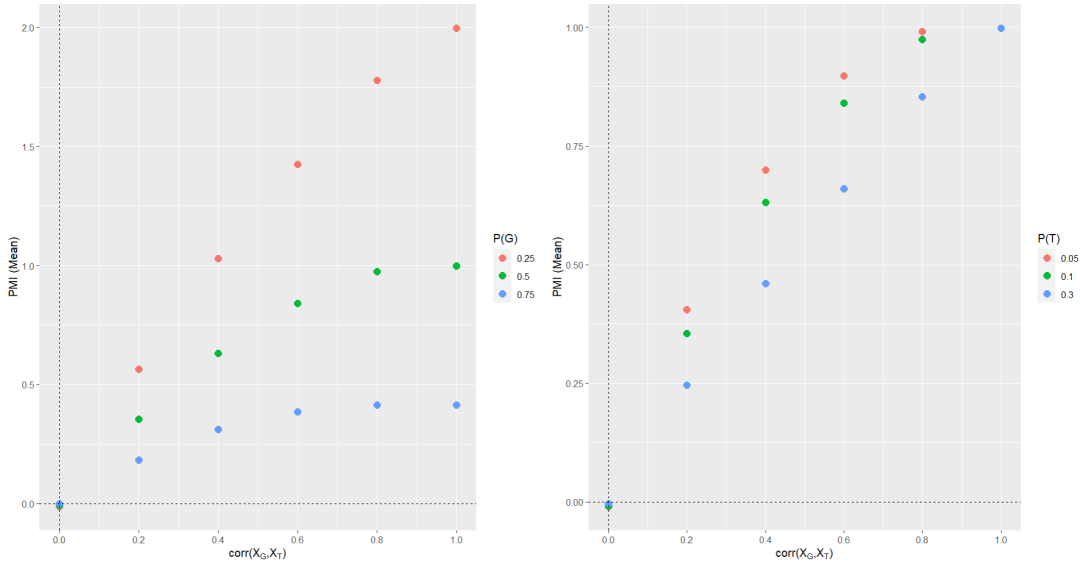
RFR demonstrates distinct characteristics. When there is repulsion between the group and the term, the term's probability of occurrence does not substantially influence its weight. However, in cases of attraction, term frequency does have some impact, albeit to a limited extent. Notably, the occurrence probability becomes a relevant factor only when a term is very typical of a group.

#### 4.4.5 Outcome-term vs group-term relationships

We now consider both the contingency table between  $G$  and  $T$  and that between  $Y$  and  $T$ , and thus compare the estimators on the two.

When using the estimator to capture both group-term and outcome-term relationships, *what's the percentage difference in the estimator's value between the two?* To answer this question we propose a graphical tool. For each subscenario, it shows the ratio between the mean estimate obtained on the contingency table between  $G$  and  $T$  and that obtained on the contingency table between  $Y$  and  $T$ .

**Figure 4.6 Mean values of the estimates of PMI for positive values of the correlations under the scenarios defined by  $n = 2000$ ,  $p_T = 0.1$  and three values of  $p_G$  (left) and scenarios defined by:  $n = 2000$ ,  $p_G = 0.5$  and three values of  $p_T$  (right), across multiple subscenarios.**



If  $S_{GT} = S(G, T)$  is the estimator computed on the contingency table between  $G$  and  $T$ , and  $S_{YT} = S(Y, T)$  is computed on the contingency table between  $Y$  and  $T$ , we investigate the following ratio under a certain subscenario:

$$V = \frac{\bar{S}_{GT}}{\bar{S}_{YT}} * 100 \quad (4.15)$$

where  $\bar{S}_{GT}$  and  $\bar{S}_{YT}$  are mean estimators over the  $R$  repetitions in a certain subscenario. The ratio tells us how much an estimator retains its value for the outcome-term relationship if computed on the group-term interaction.

To show the results of the simulation for several subscenarios, we propose the following graphical tool. In a scenario defined by  $n$ ,  $p_Y$ ,  $p_G$ , and  $p_T$ , the correlation matrix between  $X_Y$ ,  $X_G$ , and  $X_T$  defines the subscenarios. Given a value for  $\rho_{X_Y, X_G}$ , i.e. the ground truth for the association between the outcome and the group variable, a heatmap shows the estimate of  $V$  for each admissible combination of the selected values for the parameters  $\rho_{X_Y, X_T}$  and  $\rho_{X_G, X_T}$ .

It's worth considering that  $V$  is a ratio of means, and does not consider variability of estimates. Thus, a best practice to make these heatmaps interpretable is to (1) use the tools proposed by Section 4.4.1 to Section 4.4.4 to search for scenarios with low variability, and hence reliable estimates, and (2) to look at the heatmap to evaluate  $V$  in subscenarios of interest.

For the sake of understanding, we show examples below.

For RFR and IDFR, given  $n = 2000$  and  $p_T = 0.1$ , we found two scenarios with low variability, *inter alia*:

- $(p_Y, p_G) = (0.5, 0.5)$ ;
- $(p_Y, p_G) = (0.75, 0.25)$ .

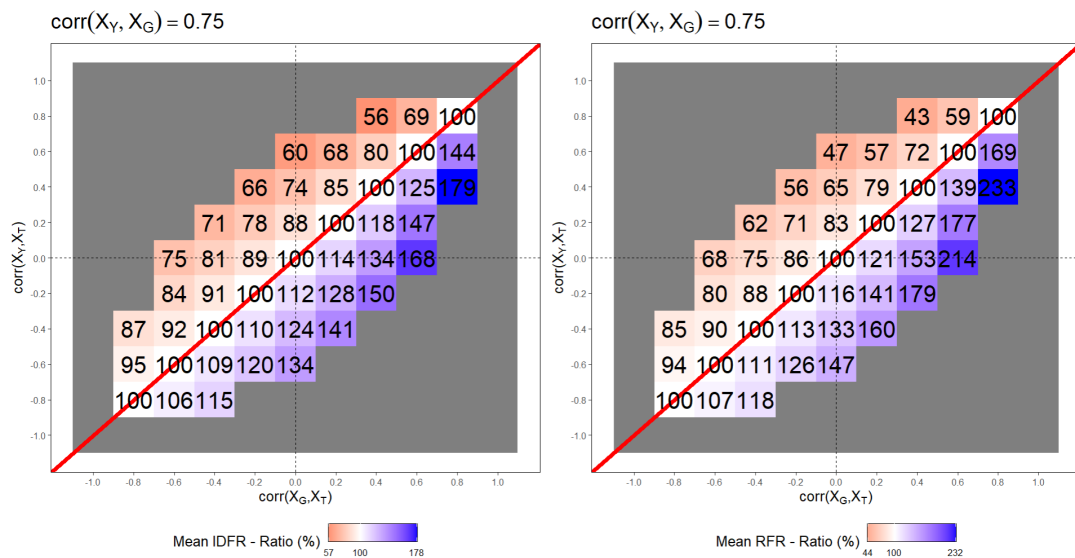
and for each scenario we look at  $V$  in the admissible combinations of the values of  $\rho_{X_Y, X_T}$  and  $\rho_{X_G, X_T}$  and  $\rho_{X_Y, X_G} = 0.75$ .

Figure 4.7 shows the estimates of  $V$  for IDFR (left) and RFR (right) in the first scenario, where  $p_Y = p_G = 0.5$ , i.e. when both the outcome and the group variables are balanced.

The behaviour of  $V$  for the two estimators is very similar in this scenario. First of all, the graph shows that in subscenarios where  $p_Y = p_G$  (red bisector) the estimate of the ratio  $V$  is always equal to 100. Which means that, as mentioned in Section 4.3, the outcome-term relation can be fully studied by analysing the group-term relation.

As expected, as one moves away from the bisector, the estimate of  $V$  moves away from 100, and thus always worse  $\bar{S}_{GT}$  captures  $\bar{S}_{YT}$ . The difference between  $\bar{S}_{YT}$  and  $\bar{S}_{GT}$  is greater as  $\rho_{X_G, X_T}$  increases, since, as seen in the graph in Figure 4.4 (top-left and top-right), the mean estimates of IDFR and RFR are not a linear function of  $\rho_{X_G, X_T}$ , but approximately exponential, and thus not symmetric around their null value.

**Figure 4.7 Estimates of the ratio  $V$  for IDFR (left) and RFR (right), across multiple subscenarios; scenario parameter values:  $n = 2000$ ,  $p_Y = 0.5$ ,  $p_G = 0.5$ ,  $p_T = 0.1$ .**



The behaviour of  $V$  for the two estimators is instead different in the second scenario considered here, where  $p_Y = 0.75$  and  $p_G = 0.25$ , shown in Figure 4.8. To understand the dynamics underlying the estimation of  $V$ , it is useful to recall the mean estimates of IDFR and RFR in Figures 4.4 in the top-left and top-right graphs, respectively. There we saw that two curves, e.g. the red and the blue, can be seen as relating to the mean estimates of GS for two perfectly correlated, but differently unbalanced document variables; in the case of IDFR they are not comparable for any level of association of the variable with the term, while in the case of RFR they are not comparable for high values of this association. This dynamic is exactly what we see on the red bisector of the graphs in Figure 4.8. We then illustrate what happens for the two estimators.

For IDFR (Figure 4.8, left), on the bisector, we see that as both the value of the two correlations increase and decrease, the ability of  $\bar{S}_{GT}$  to capture  $\bar{S}_{YT}$  decreases. Beyond the bisector, the graph shows the ratio  $V$  when  $\rho_{X_Y, X_T} \neq \rho_{X_G, X_T}$ . As  $\rho_{X_Y, X_T}$  and  $\rho_{X_G, X_T}$  tend towards negative values, the colour gradient tends towards blue:



described in Section 4.2.3. What has been shown here has served, on the one hand, to highlight the methodological contribution of this work, and, on the other, to suggest an interpretation of visualisation tools from a Text Mining perspective.

In this section, we describe the features of a web application<sup>5</sup>, created to enable free exploration of all the simulation results. Although the platform may be subject to future developments, it is already accessible, and its main features will not change over time.

The application has been realized using the *Shiny* package of the *R* software, i.e. a well-known open source tool for building web applications.

The platform is structured in 5 panels.

The **Methods** panel summarises the methodological details of the study described in Section 4.2.

The **One simulation** panel allows the user to exploit the Gaussian Copula method to simulate a corpus of documents, described by the outcome, group, and term variables, by independently defining the values of the parameters involved. The simulation returns contingency tables both between the group variable and the term variable and between the target variable and the term variable, as well as sample estimates of all proposed estimators calculated on the two tables.

The **Mean Estimates** and **Distribution of Estimates** panels allow the simulation results described here to be explored in the manner suggested by Sections 4.4.1 to 4.4.4 but for all estimators illustrated in Section 4.2.3 and all scenarios described in Section 4.3. The former panel shows in one dashboard an overview of the mean estimates for all estimators in a user-defined scenario, while the latter shows the boxplots of the estimates by estimator.

Finally, in the last panel, **Group vs Outcome**, the analyses of the relationships between group-term and outcome-term interactions, of the type described in Section 4.4.5, are shown, allowing the user to choose the scenario and the estimator.

## 4.6 Conclusions and discussion

This study considered the setting of grouped text documents in which two document variables are available in a corpus: an *outcome* variable and a *group* variable.

The contribution of this work is to frame this context with a statistical approach, by modeling the corpus of documents with a Multivariate Binomial distribution. The advantage of this solution is two-fold: it allows (1) to review, in a statistical framework, some weighting measures for grouped text documents used in the literature, and (2) to simulate corpora with predefined characteristics by means of the Gaussian Copula method.

This simulation is useful to analyze the behavior of the existing measures from several perspectives. First, it can be used to assess their ability to differently weight the group specificity of two terms that share all characteristics except the true value of their specificity. Then, the impact of group imbalance and term rarity on the value of the measures can be examined. Finally, the simulation allows for the investigation of the ability of the existing measures, computed on the group-word interaction, to capture both the group-word relationship itself and the outcome-word association.

---

<sup>5</sup>The URL to access the platform is <https://unknownauthor.shinyapps.io/shiny/>. Since the study illustrated in this chapter, including the web application, will be submitted for publication in a scientific journal, the author's name has been temporarily masked with the nickname "*unknownauthor*" to ensure blindness in the peer review process. In the future, the platform will be accessible at: <https://riccardoricciardi.shinyapps.io/shiny/>.



Results from the simulation study show interesting relationships that can be exploited by nice visualization tools, and all are made available on a web application. The application is also valuable since the results presented here are only related to example scenarios, which serve to demonstrate the utility of the simulation and provide guidance for interpreting the visualizations offered.

These examples have been useful in alerting text mining practitioners to some critical issues. Studying the case of terms that differ only in their ground truth relationship with the group, it was noted that some measures can only be used to rank terms based on their group specificity, but not as a measure of its intensity. In addition, in some scenarios, these cannot even be used in this way if the occurrence probabilities of the words are different.

Regarding the role of proportions between groups, it was noted that the estimated specificity is affected by the imbalance of groups. In addition, if there are two group variables with different marginal probabilities, the weights of the same term for the two variables are not comparable.

It is worth mentioning that within the proposed statistical approach one may frame the problem in different ways. For example, assuming that a word is very rare in a very large corpus makes the variable of term presence distribute as a Poisson distribution, and, considering additional term variables, co-occurrences and thus semantic relationships between words can be modeled and then explored.



# Conclusions

This thesis addressed the statistical analysis of textual data, with a focus on the common case in which text samples can be grouped. It aimed to contribute to the field with both applications and methodological proposals, and it has been conceived as a collection of papers, arranged in chronological order based on their presentation sequence.

In Chapter 2, we focused on *categorical* and *textual* self-descriptions to study how different groups of people present themselves on social media, by exploring the case of the StockTwits platform. Specifically, StockTwits users write a short bio, and specify whether they are either technical or fundamental traders, about their approach, either short-term or long-term investors, about their primary holding period, and either professionals or non-professionals, about their experience in trading.

The methodology proposed consisted of training a language model on a sample of text documents. We worked on a sample with balanced categorical characteristics, to prevent the model from overfitting specific linguistic signals. Then, by using a list of both domain-specific and statistically relevant words as a guide, similarities between word and document representations were explored to analyze group differences in self-describing. Eventually, a bootstrap procedure was leveraged to assess the validity of the results.

This study suggested that the words related to trading characteristics expressly proposed by the platform are used by users to self-describe, and therefore they are useful to distinguish one group of traders from another, for each trading feature considered separately. Generally, bios reflect the relationship between the approach and the primary holding, while are transversely influenced by the experience in trading. Particularly, technical traders have very different semantics from fundamental traders, often referring to the tools they use to assess the profitability of their current and potential investments, like charting, predictive and automated models, etc. In addition, what distinguishes the professionals from the non-professionals is the use of domain-specific words, but still very specific words.

About the bootstrap procedure, it was useful to assess the validity of what was suggested by the feature selection measures: some words were initially detected as able to discriminate between groups of users, but the proposed methodology shed light on the high variability of some of those results.

In conclusion, the proposed procedure can prove beneficial in other contexts as well, that is whenever one wants to explore the heterogeneity of a language across grouped text documents.

This study aimed to contribute to the analysis of the relationships between labels that summarize characteristics and the texts related to them. Apart from checking whether the two correspond, it would be interesting to understand how individuals understand a given characteristic, and to study the heterogeneity of the language of people belonging to a certain group, as well as to compare the heterogeneity of groups. For instance, it can be insightful to understand how politicians with different political alignments use the term *democratic* and whether the term itself takes on varying nuances within different political groups.

From a methodological point of view, the purpose of further research is to improve the sample selection step, to obtain a balanced sample. When working with text data, it is worth considering that statistical units, i.e. words, are organized in a three-level hierarchy, where if usually documents belong to only one group, as in the context of this paper, words instead may appear in more than one document. In this case, specific techniques of multilevel stratification should be used to create homogeneous groups of statistical units that share the same characteristics at multiple levels.

Within the context of the "Data Science for Brescia - Arts and Cultural Places" Project, Chapter 3 showed the results of the training of a language model on the Italian-written online reviews to classify them into four distinct semantic *areas*, defined by the main four attractions of the city of Brescia, in Italy.

The great utility of such a model stems from the fact that it can be used to verify whether a text document, such as a post on social media platforms, forums, and other online spaces, refers to an attraction (for example, a given museum) when the attraction is not explicitly mentioned in the document metadata. Therefore, the proposed model can process and classify those texts, allowing for the expansion of the online discourse database concerning those cultural attractions.

The *Multilingual* BERT model has been fine-tuned on this multiclassification task, and it yielded very good performance results on the validation set, i.e. the set that was not used for training. Given the multilingual capabilities of the BERT model, future developments of this study will involve expanding the dataset with reviews written in languages other than Italian.

To further validate the model, clusters of reviews from the validation set have been detected, based on their vector representations produced by the model. The count-based method of the *Keyness*  $K$  has been employed to extract cluster-specific keywords, and they have proven to be highly consistent with the domain of application, i.e., with the characteristics and offerings of the four cultural attractions.

Chapter 4 showed to a methodological study about the notion of *group-specificity* of a term, since it analyzed *group-specificity estimators*. These are functions of word statistics, proposed in the literature in various contexts, and which are collected here under the definition of group-specificity estimators, which is introduced here for the first time.

The study considered the setting of grouped text documents in which two document variables are available in a corpus: an *outcome* variable and a *group* variable.

The contribution of this work is to frame this context with a statistical approach, by modeling the corpus of documents with a Multivariate Binomial distribution. The advantage of this solution is two-fold: it allows (1) to review, in a statistical framework, some term-weighting measures used in the literature, and (2) to simulate corpora with predefined characteristics by means of the Gaussian Copula method. This simulation is useful for investigating the behavior of the existing measures from several perspectives. First, it can be used to assess their ability to differently weight the group specificity of two terms that share all characteristics except the true value of their specificity. Then, it provides an opportunity to explore the influence of group imbalance and term rarity on the value of the measures. Lastly, the simulation enables an exploration of the capability of the existing measures, computed on the group-word interaction, to capture both the group-word relationship itself and the outcome-word association.

The web application is also valuable since the results presented in Chapter 4 were only related to example scenarios, which serve to demonstrate the utility of the simulation and provide guidance for interpreting the visualizations offered.

It is worth mentioning that within the proposed statistical approach one may frame the problem in different ways. For example, assuming that a word is very rare in a very large corpus makes the variable of term presence distribute as a Poisson distribution, and, considering additional term variables, co-occurrences and thus semantic relationships between words can be modeled and then explored.



# Bibliography

- [1] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.
- [2] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.
- [3] C. Cleverdon, J. Mills, and M. Keen, "Factors determining the performance of indexing systems volume 1. design," 1966.
- [4] F. W. Lancaster, "Information retrieval systems; characteristics, testing, and evaluation," 1968.
- [5] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.
- [6] G. Salton and C.-S. Yang, "On the specification of term values in automatic indexing," *Journal of documentation*, vol. 29, no. 4, pp. 351–372, 1973.
- [7] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.
- [8] K. Chen, Z. Zhang, J. Long, and H. Zhang, "Turning from TF-IDF to TF-IGM for term weighting in text classification," *Expert Systems with Applications*, vol. 66, pp. 245–260, 2016.
- [9] Z.-H. Deng, K.-H. Luo, and H.-L. Yu, "A study of supervised term weighting scheme for sentiment analysis," *Expert Systems with Applications*, vol. 41, no. 7, pp. 3506–3513, 2014.
- [10] A. E. Aladağ, S. Muderrisoglu, N. B. Akbas, O. Zahmacioglu, and H. O. Bingol, "Detecting suicidal ideation on forums: Proof-of-concept study," *Journal of medical Internet research*, vol. 20, no. 6, e9840, 2018.
- [11] L. Wang, X.-k. Wang, J.-j. Peng, and J.-q. Wang, "The differences in hotel selection among various types of travellers: A comparative analysis with a useful bounded rationality behavioural decision support model," *Tourism management*, vol. 76, p. 103961, 2020.
- [12] F. Rustam, M. Khalid, W. Aslam, V. Rupapara, A. Mehmood, and G. S. Choi, "A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis," *Plos one*, vol. 16, no. 2, e0245909, 2021.
- [13] M. Misuraca and M. Spano, "Unsupervised analytic strategies to explore large document collections," in *Iezzi, D.F., Mayaffre, D., Misuraca, M. (eds) Text Analytics. JADT 2018. Studies in Classification, Data Analysis, and Knowledge Organization*, Springer, Cham, 2020.
- [14] M. Misuraca, G. Scepi, and M. Spano, "Network-based dimensionality reduction for textual datasets," in *Brentari, E., Chiodi, M., Wit, E.J.C. (eds) Models for Data Analysis. SIS 2018. Springer Proceedings in Mathematics & Statistics, vol 402*, Springer, Cham, 2023.

- [15] G. Salton, "Automatic text processing: The transformation, analysis, and retrieval of," *Reading: Addison-Wesley*, vol. 169, 1989.
- [16] K. Kageura and B. Umino, "Methods of automatic term recognition: A review," *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, vol. 3, no. 2, pp. 259–289, 1996.
- [17] C. Kit and X. Liu, "Measuring mono-word termhood by rank difference via corpus comparison," *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, vol. 14, no. 2, pp. 204–229, 2008.
- [18] T. G. Harwood and T. Garry, "An overview of content analysis," *The marketing review*, vol. 3, no. 4, pp. 479–498, 2003.
- [19] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Icml*, Nashville, TN, USA, vol. 97, 1997, p. 35.
- [20] F. Debole and F. Sebastiani, "Supervised term weighting for automated text categorization," in *Proceedings of the 2003 ACM symposium on Applied computing*, 2003, pp. 784–788.
- [21] M. Lan, C. L. Tan, J. Su, and Y. Lu, "Supervised and traditional term weighting methods for automatic text categorization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 4, pp. 721–735, 2008.
- [22] M. F. Caropreso, S. Matwin, and F. Sebastiani, "A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization," *Text databases and document management: Theory and practice*, vol. 5478, no. 4, pp. 78–102, 2001.
- [23] S. S. Samant, N. B. Murthy, and A. Malapati, "Improving term weighting schemes for short text classification in vector space model," *IEEE Access*, vol. 7, pp. 166 578–166 592, 2019.
- [24] R. J. Firth, "A synopsis of linguistic theory," *Studies in Linguistic Analysis. Special Volume of the Philological Society*, pp. 1–31, 1957.
- [25] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20 150 202, 2016.
- [26] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 2013, pp. 746–751.
- [27] O. Levy and Y. Goldberg, "Linguistic regularities in sparse and explicit word representations," in *Proceedings of the eighteenth conference on computational natural language learning*, 2014, pp. 171–180.
- [28] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.
- [29] G. W. Stewart, "On the early history of the singular value decomposition," *SIAM review*, vol. 35, no. 4, pp. 551–566, 1993.
- [30] Y. Goldberg, *Neural network methods for natural language processing*. Springer Nature, 2022.
- [31] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *ArXiv*, 2013. DOI: <https://doi.org/10.48550/arXiv.1301.3781>.



- [32] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on Machine Learning*, E. P. Xing and T. Jebara, Eds., ser. Proceedings of Machine Learning Research, vol. 32, Beijing, China: PMLR, 2014, pp. 1188–1196.
- [33] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162.
- [34] D. Hovy and S. Prabhume, "Five sources of bias in natural language processing," *Language and Linguistics Compass*, vol. 15, no. 8, e12432, 2021.
- [35] D. Bamman, C. Dyer, and N. A. Smith, "Distributed Representations of Geographically Situated Language," in *ACL*, 2014. DOI: 10.3115/v1/P14-2134.
- [36] D. Hovy, "Demographic factors improve classification performance," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 752–762. DOI: 10.3115/v1/P15-1073.
- [37] C. Welch, J. K. Kummerfeld, V. Pérez-Rosas, and R. Mihalcea, "Compositional demographic word embeddings," *arXiv preprint arXiv:2010.02986*, 2020.
- [38] L. Rheault and C. Cochrane, "Word embeddings for the analysis of ideological placement in parliamentary corpora," *Political Analysis*, vol. 28, no. 1, pp. 112–133, 2020.
- [39] F. Petroni, V. Plachouras, T. Nugent, and J. L. Leidner, "Attr2vec: Jointly learning word and contextual attribute embeddings with factorization machines," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 453–462.
- [40] K. Tian, T. Zhang, and J. Zou, "Cover: Learning covariate-specific vector representations with tensor decompositions," in *International Conference on Machine Learning*, PMLR, 2018, pp. 4926–4935.
- [41] N. A. Smith, "Contextual word representations: A contextual introduction," *arXiv preprint arXiv:1902.06006*, 2019.
- [42] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [44] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," *arXiv preprint arXiv:1908.08345*, 2019.
- [45] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification?" In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, Springer, 2019, pp. 194–206.
- [46] X. Zhang, Y. Zhang, Q. Zhang, *et al.*, "Extracting comprehensive clinical information for breast cancer using deep learning methods," *International journal of medical informatics*, vol. 132, p. 103985, 2019.

- [47] J.-S. Lee and J. Hsiang, "Patent classification by fine-tuning BERT language model," *World Patent Information*, vol. 61, p. 101965, 2020.
- [48] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," *arXiv preprint arXiv:1903.10676*, 2019.
- [49] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.
- [50] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.
- [51] T. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [52] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.
- [53] R. Ricciardi, "What does your self-description reveal about you? A pipeline to analyse StockTwits users," in *A. Balzanella, M. Bini, C. Cavicchia, R. Verde (eds.), Book of the Short Papers of the 51st Scientific Meeting of the Italian Statistical Society*, Caserta, Italy: Pearson, 2022, pp. 1809–1814.
- [54] D. Yates and S. Paquette, "Emergency knowledge management and social media technologies: A case study of the 2010 haitian earthquake," *International journal of information management*, vol. 31, no. 1, pp. 6–13, 2011.
- [55] B. Liu, "Sentiment Analysis and Opinion Mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, May 23, 2012, ISSN: 1947-4040. DOI: 10.2200/S00416ED1V01Y201204HLT016.
- [56] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos, "Community detection in social media: Performance and application considerations," *Data mining and knowledge discovery*, vol. 24, pp. 515–554, 2012.
- [57] T. H. Nguyen, K. Shirai, and J. Velcin, "Sentiment analysis on social media for stock movement prediction," *Expert Systems with Applications*, vol. 42, no. 24, pp. 9603–9611, 2015.
- [58] M. Thelwall, S. Thelwall, and R. Fairclough, "Male, female, and nonbinary differences in uk twitter self-descriptions: A fine-grained systematic exploration," *Journal of Data and Information Science*, vol. 6, no. 2, pp. 1–27, 2021.
- [59] X. Chen, Y. Wang, E. Agichtein, and F. Wang, "A comparative study of demographic attribute inference in twitter," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 9, 2015, pp. 590–593.
- [60] F. Luo, G. Cao, K. Mulligan, and X. Li, "Explore spatiotemporal and demographic characteristics of human mobility via twitter: A case study of chicago," *Applied Geography*, vol. 70, pp. 11–25, 2016.
- [61] E. Gilbert, K. Karahalios, and C. Sandvig, "The network in the garden: Designing social media for rural life," *American Behavioral Scientist*, vol. 53, no. 9, pp. 1367–1388, 2010.
- [62] J. Li, A. Ritter, and E. Hovy, "Weakly supervised user profile extraction from twitter," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, pp. 165–174.

- [63] L. Tang and H. Liu, "Leveraging social media networks for classification," *Data Mining and Knowledge Discovery*, vol. 23, pp. 447–478, 2011.
- [64] M. Pennacchiotti and A.-M. Popescu, "Democrats, republicans and starbucks aficionados: User classification in twitter," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 430–438.
- [65] F. Audrino, F. Sigris, and D. Ballinari, "The impact of sentiment and attention measures on stock market volatility," *International Journal of Forecasting*, vol. 36, no. 2, pp. 334–357, 2020.
- [66] N. Oliveira, P. Cortez, and N. Areal, "On the predictability of stock market behavior using stocktwits sentiment and posting volume," in *Portuguese Conference on Artificial Intelligence*, Springer, 2013, pp. 355–365.
- [67] S. Sohangir, D. Wang, A. Pomeranets, and T. M. Khoshgoftaar, "Big Data: Deep Learning for financial sentiment analysis," *Journal of Big Data*, vol. 5, no. 1, pp. 1–25, 2018.
- [68] N. Oliveira, P. Cortez, and N. Areal, "Stock market sentiment lexicon acquisition using microblogging data and statistical measures," *Decision Support Systems*, vol. 85, pp. 62–73, 2016.
- [69] T. Renault, "Intraday online investor sentiment and return patterns in the US stock market," *Journal of Banking & Finance*, vol. 84, pp. 25–40, 2017.
- [70] B. Lu, M. Ott, C. Cardie, and B. K. Tsou, "Multi-aspect sentiment analysis with topic models," in *2011 IEEE 11th international conference on data mining workshops*, IEEE, 2011, pp. 81–88.
- [71] T. Hoang, H. Le, and T. Quan, "Towards autoencoding variational inference for aspect-based opinion summary," *Applied Artificial Intelligence*, vol. 33, no. 9, pp. 796–816, 2019.
- [72] Y. Yang and J. O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," presented at the International Conference on Machine Learning, Jul. 8, 1997.
- [73] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, Mar. 2002.
- [74] H. Altınçay and Z. Erenel, "Analytical evaluation of term weighting schemes for text categorization," *Pattern Recognition Letters*, vol. 31, no. 11, pp. 1310–1323, Aug. 2010.
- [75] M. Lan, C. L. Tan, J. Su, and Y. Lu, "Supervised and traditional term weighting methods for automatic text categorization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 4, pp. 721–735, Apr. 2009.
- [76] S. S. Samant, N. L. Bhanu Murthy, and A. Malapati, "Improving Term Weighting Schemes for Short Text Classification in Vector Space Model," *IEEE Access*, vol. 7, pp. 166 578–166 592, 2019.
- [77] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Boston: John Wiley & Sons Inc., 2005.
- [78] L. Chen, Y. Li, X. Deng, Z. Liu, M. Lv, and T. He, "Semantic-aware network embedding via optimized random walk and paragraph2vec," *Journal of Computational Science*, vol. 63, p. 101 825, 2022.

- [79] M. Çetinkaya Rundel and J. Hardin, *Introduction to Modern Statistics*. 2021, ISBN: 1943450145.
- [80] J. Ooms, *Cld2: Google's compact language detector 2*, <https://docs.ropensci.org/cld2/>, 2022.
- [81] J. Ooms, *Cld3: Google's compact language detector 3*, <https://docs.ropensci.org/cld3/>, 2023.
- [82] J. J. Murphy, *Study Guide to Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications*. New York: Penguin, 1999.
- [83] B. Graham and L. Daniels, *The Intelligent Investor Rev Ed*. New York: Harper-Collins, 2015.
- [84] T. R. Price, *A successful investment philosophy based on the growth stock theory of investing*, 1973.
- [85] M. Ventura and R. Ricciardi, "Correspondence analysis and hierarchical clustering to analyse stocktwits users," *Statistics & Applications*, vol. 19, 2022.
- [86] R. P. Schumaker and H. Chen, "A quantitative stock prediction system based on financial news," *Information Processing & Management*, vol. 45, no. 5, pp. 571–583, 2009.
- [87] R. Elliott, *The wave principle*. New York: Snowball Publishing, 2012.
- [88] S. E. Bibri and J. Krogstie, "Smart sustainable cities of the future: An extensive interdisciplinary literature review," *Sustainable cities and society*, vol. 31, pp. 183–212, 2017.
- [89] M. Manisera and P. Zuccolotto, "A mixture model for ordinal variables measured on semantic differential scales," *Econometrics and Statistics*, vol. 22, pp. 98–123, 2022.
- [90] D. Piccolo and R. Simone, "The class of CUB models: Statistical foundations, inferential issues and empirical evidence," *Statistical Methods & Applications*, vol. 28, pp. 389–435, 2019.
- [91] K. Kim, O.-j. Park, S. Yun, and H. Yun, "What makes tourists feel negatively about tourism destinations? application of hybrid text mining methodology to smart destination management," *Technological Forecasting and Social Change*, vol. 123, pp. 362–369, 2017.
- [92] R. Socher, A. Perelygin, J. Wu, *et al.*, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1631–1642.
- [93] F. Mehraliyev, A. P. Kirilenko, and Y. Choi, "From measurement scale to sentiment scale: Examining the effect of sensory experiences on online review rating behavior," *Tourism Management*, vol. 79, p. 104096, 2020.
- [94] A. Frazier, "Developing a vocabulary of the senses," *Elementary English*, vol. 47, no. 2, pp. 176–184, 1970.
- [95] M. Thelwall, "The heart and soul of the web? sentiment strength detection in the social web with sentistrength," *Cyberemotions: Collective emotions in cyberspace*, pp. 119–134, 2017.
- [96] L. Serrano, A. Ariza-Montes, M. Nader, A. Sianes, and R. Law, "Exploring preferences and sustainable attitudes of airbnb green users in the review comments and ratings: A text mining approach," *Journal of Sustainable Tourism*, vol. 29, no. 7, pp. 1134–1152, 2021.

- [97] B. Ray, A. Garain, and R. Sarkar, "An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews," *Applied Soft Computing*, vol. 98, 2021.
- [98] Y.-C. Chang, C.-H. Ku, and D.-D. Le Nguyen, "Predicting aspect-based sentiment using deep learning and information visualization: The impact of covid-19 on the airline industry," *Information & Management*, vol. 59, no. 2, 2022.
- [99] C. De Lucia, P. Paziienza, P. Balena, and D. Caporale, "Exploring local knowledge and socio-economic factors for touristic attractiveness and sustainability," *International Journal of Tourism Research*, vol. 22, no. 1, pp. 81–99, 2020.
- [100] Y. Chen and I. P. Tussyadiah, "Service failure in peer-to-peer accommodation," *Annals of Tourism Research*, vol. 88, 2021.
- [101] C. Chantrapornchai and A. Tunsakul, "Information extraction on tourism domain using SpaCy and BERT," *ECTI Transactions on Computer and Information Technology*, vol. 15, no. 1, pp. 108–122, 2021.
- [102] S. Qi, C. U. I. Wong, N. Chen, J. Rong, and J. Du, "Profiling Macau cultural tourists by using user-generated content from online social media," *Information Technology & Tourism*, vol. 20, pp. 217–236, 2018.
- [103] D. Agostino, M. Brambilla, S. Pavanetto, and P. Riva, "The contribution of online reviews for quality evaluation of cultural tourism offers: The experience of italian museums," *Sustainability*, vol. 13, no. 23, p. 13 340, 2021.
- [104] M. Phuong and M. Hutter, "Formal algorithms for transformers," *arXiv preprint arXiv:2207.09238*, 2022.
- [105] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013, vol. 112.
- [106] K. P. Murphy, "Machine learning: A probabilistic perspective (adaptive computation and machine learning series)," *The MIT Press: London, UK*, 2018.
- [107] A. Mao, M. Mohri, and Y. Zhong, "Cross-entropy loss functions: Theoretical analysis and applications," *arXiv preprint arXiv:2304.07288*, 2023.
- [108] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.
- [109] D. L. Olson and D. Delen, *Advanced data mining techniques*. Springer Science & Business Media, 2008.
- [110] A. Adhikari, A. Ram, R. Tang, and J. Lin, "DocBERT: BERT for document classification," *arXiv preprint arXiv:1904.08398*, 2019.
- [111] S. Lamprinidis, F. Bianchi, D. Hardt, and D. Hovy, "Universal joy a data set and results for classifying emotions across languages," in *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2021, pp. 62–75.
- [112] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia conference on knowledge discovery and data mining*, Springer, 2013, pp. 160–172.
- [113] L. McInnes, J. Healy, and S. Astels, "Hdbscan: Hierarchical density based clustering.," *Journal Open Source Software*, vol. 2, no. 11, p. 205, 2017.
- [114] K. Benoit, K. Watanabe, H. Wang, *et al.*, "Quanteda: An R package for the quantitative analysis of textual data," *Journal of Open Source Software*, vol. 3, no. 30, p. 774, Oct. 6, 2018.

- [115] M. Straka and J. Straková, *Universal dependencies 2.5 models for UDPipe (2019-12-06)*, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, 2019.
- [116] R. Ricciardi and M. Manisera, "Evaluation of term-weighting measures for grouped text documents with a target variable: A simulation study," in *Proceedings of 5th International Conference on Advanced Research Methods and Analytics, Sevilla, Spain*, Editorial Universitat Politècnica de València, 2023, pp. 97–97.
- [117] B. Dai, S. Ding, and G. Wahba, "Multivariate Bernoulli distribution," *Bernoulli*, vol. 19, no. 4, pp. 1465–1483, Sep. 2013.
- [118] P. Xue-Kun Song, "Multivariate dispersion models generated from gaussian copula," *Scandinavian Journal of Statistics*, vol. 27, no. 2, pp. 305–320, 2000.
- [119] A. Sklar, "Fonctions de répartition à n dimensions et leurs marges," *Publications de l'Institut Statistique de l'Université de Paris*, vol. 8, pp. 229–47, 1959.
- [120] F. Durante, J. Fernandez-Sanchez, and C. Sempi, "A topological proof of Sklar's theorem," *Applied Mathematics Letters*, vol. 26, no. 9, pp. 945–948, 2013.
- [121] C. Genest and J. MacKay, "The joy of copulas: Bivariate distributions with uniform marginals," *The American Statistician*, vol. 40, no. 4, pp. 280–283, 1986.
- [122] A. Lee, "Generating random binary deviates having fixed marginal distributions and specified degrees of association," *The American Statistician*, vol. 47, no. 3, pp. 209–215, 1993.
- [123] W. N. Hudson, H. G. Tucker, and J. A. Veeh, "Limit theorems for the multivariate binomial distribution," *Journal of Multivariate Analysis*, vol. 18, no. 1, pp. 32–45, Feb. 1986.
- [124] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken, *Applied multiple regression / correlation analysis for the behavioral sciences*. Routledge, 2013.