mathematics

MDPI

Article

A Unified Formal Framework for Factorial and Probabilistic Topic Modelling

Karina Gibert * and Yaroslav Hernandez-Potiomkin

Intelligent Data Science and Artificial Intelligence Research Group, Universitat Politècnica de Catalunya, 08034 Barcelona, Spain; yaroslav.hernandez@upc.edu
* Correspondence: karina.gibert@upc.edu

Abstract: Topic modelling has become a highly popular technique for extracting knowledge from texts. It encompasses various method families, including Factorial methods, Probabilistic methods, and Natural Language Processing methods. This paper introduces a unified conceptual framework for Factorial and Probabilistic methods by identifying shared elements and representing them using a homogeneous notation. The paper presents 12 different methods within this framework, enabling easy comparative analysis to assess the flexibility and how realistic the assumptions of each approach are. This establishes the initial stage of a broader analysis aimed at relating all method families to this common framework, comprehensively understanding their strengths and weaknesses, and establishing general application guidelines. Also, an experimental setup reinforces the convenience of having harmonized notational schema. The paper concludes with a discussion on the presented methods and outlines future research directions.

Keywords: multivariate methods; topic modelling; probabilistic methods

MSC: 62H22; 62H25

Citation: Gibert , K.; Hernandez-Potiomkin, Y. A Unified Formal Framework for Factorial and Probabilistic Topic Modelling. *Mathematics* **2023**, *11*, 4375. https://doi.org/10.3390/math11204375

Academic Editor: Carmen Patino-Alonso

Received: 27 July 2023
Revised: 19 October 2023
Accepted: 19 October 2023
Published: 21 October 2023

1. Introduction

The analysis of textual data has become one of the hottest topics recently. It is widely used for (1) semantic document processing through topic extraction and (2) text summarization by associating principal concepts to it. Topic modelling finds many applications, e.g., enriched document clustering [1], trend identification in topics [2], classification in high dimensions [3], and dimensionality reduction problems [4].

Despite the amount of available techniques, there are few challenges that need to be addressed: the semantic structures of texts, i.e., synonymy and polysemy, orthography, and outlying artifacts, as well as the subjectivity of interpretation.

Topic modelling methods belong to three main families, (1) Factorial, (2) Probabilistic, and (3) Natural Language Processing (NLP)-based methods. The former consist of a decomposition over the multivariate so-called design matrix, where a given objective function is optimized with a set of constraints [5]. These types of techniques are usually supported by powerful geometrical interpretation capabilities as the original data are projectedover the factorial space, which is expected to properly summarize relevant patterns in the data. They also benefit from a wide range of algebraic properties, resulting in theoretically well-based and robust methods. Some examples of such techniques are Principal Components Analysis (PCA) [6,7], Latent Semantic Analysis (LSA) [8], Non-negative Matrix Factorization (NMF) [9], Canonical Correlation Analysis (CCA), Multiple Correspondence Analysis (MCA) [10], Correspondence Analysis (CA) [7,10], Archetypal Analysis (AA) [5], the Nonlinear Iterative Partial Least Squares algorithm (NIPALS) [6], BERTopic models successfully combined with Kernel-PCA [11], and others.

Probabilistic methods are based on a statistical model definition, built from a probability model and a parameter space definition. Hence, the parameter estimation is tackled through the frequentist approach by making use of the Maximum Likelihood function, or

Bayesian framework, in which the definition of prior distribution and the application of Bayes theorem to obtain the posterior is needed [12]. Probabilistic methods are valuable due to their generative nature, and they also provide clear interpretation, flexibility, and extensibility. Furthermore, additional levels of hierarchy can be introduced to become Bayesian hierarchical models [13,14]. Probabilistic topic modelling is successfully applied to multi-document summarization [15,16], text classification [17], automatic extraction of topics [18] and document topic classification [19]. Despite Bayesian approaches having gained popularity, factorial methods are still more common as an application, as well as research-wise.

Additional consideration is made for methods that make use of M-estimators (e.g., likelihood functions) for robust estimation of the parameters to prevent the effects of outliers as well as multicollinearity [20,21].

NLP methods combine language analysis and statistical methods [22], and were powerful in inferring the meaning from text. Linguistic annotations, such as Treebanks [23–26], were widely used for unsupervised training stages, and were especially useful in part-of-speech (POS) tagging, morphological analysis, word sense disambiguation [25,27], and syntactic parsing. Hence, the topic modelling field can benefit from NLP, especially in the pre-processing stage, which will be shown in the experimental setup.

The goal of this work is to introduce a unified and homogeneous notation over different techniques, building a bridge between different families of topic modelling methods, especially Factorial and Bayesian approaches. Moreover, we provide an experimental setup that highlights the usefulness of this harmonization of the notation, showing that the comparison, in particular qualitative comparison, of different methods becomes an easier task. Moreover, both harmonized notation and experimental setups reinforce the need for analysing different assumptions of the methods to efficiently derive the right conclusions from extracted patterns.

The paper is structured as follows. First, a common notation is presented. Then, the two families analysed in the paper, i.e., factorial and probabilistic methods, follow.

LSA is introduced in Section 3, and its applications and extensions are presented (Section 3.1), as well as alternative factorization models (Sections 3.1.1 and 3.1.2) in the topic modelling field. Then, the Principal Component Analysis method and the mixture variation MPPCA (Section 4.1) is presented as a midpoint between factorial and probabilistic families. Then, Bayesian mixture models are briefly introduced in Section 4.2, and Latent Dirichlet Allocation is presented in Section 4.2.5.

Then, the experimental setup and results are presented in Section 5. And, finally, Sections 6 and 7 are devoted to the discussion, conclusions, and future work.

2. A Standardized Notation for Textual Data Analysis

This section introduces the notation associated with elements of the methods presented in the following sections.

2.1. Corpus Numerization: Common Notation

This is one of the first and most basic steps in most of the methods (from Factorial and Probabilistic families). It consists of representing a set of documents through numeric matrices, where each row corresponds to the words distribution in the document.

It takes into account that:

- A set of documents is represented by $\mathcal{D}$ and is of size $n_{\mathcal{D}}$;
- A set of terms is represented by $\mathcal{T}$ and is of size $n_{\mathcal{T}}$.

A document is composed of a sequence of words, such that for each document $d_j \in \mathcal{D}$ with $j = 1 \dots n_{\mathcal{D}}$, $d_j = (w_1, w_2, \dots, w_{n_{d_j}})$, with n_{d_j} being the number of words in the document d_j and $w_{\not i} \in \mathcal{T}$ having $\not i = 1 \dots n_{d_j}$.

The numerization of the corpus consists of producing a matrix X of dimensionality $n_{\mathcal{T}} \times n_{\mathcal{D}}$ (see below), with the rows corresponding to terms $t \in \mathcal{T}$ (the vocabulary of the

corpus $\mathcal{D}$), and columns corresponding to documents $d \in \mathcal{D}$. The number of rows of X is the length of the vocabulary ($n_{\mathcal{T}}$), and the number of columns is $n_{\mathcal{D}}$. Each cell (i, j) contains n_{ij}, i.e., the number of occurrences of the term t_i in the document d_j.

$$X = \begin{array}{c} \\ 1 \\ \vdots \\ n_{\mathcal{T}} \end{array} \begin{array}{c} \begin{array}{ccc} 1 & \dots & n_{\mathcal{D}} \end{array} \\ \begin{bmatrix} & \vdots & \\ \cdots & n_{ij} & \cdots \\ & \vdots & \end{bmatrix} \end{array} \tag{1}$$

where X is the term–document matrix (TDM). Later, other generalizations of it will be provided.

The vector $\boldsymbol{x}^i = (n_{i1}, \dots, n_{in_{\mathcal{D}}})$ describes the profile of a given term in a corpus, i.e., the distribution of the occurrences of term t_i in a document belonging to $\mathcal{D}$.

The information retrieval field [28] successfully deployed the TDM structure in search engines, as well as the tf-idf computation [29].

2.2. Binarization of Documents: Common Notation

The textual corpus can also be represented through the Binarization of documents. In it, each document $d_j \in \mathcal{D}$ corresponds to a binary matrix, $d^{(j)}$, the terms' distribution in the document.

The matrix $d^{(j)}$ has the dimensionality $n_{d_j} \times n_{\mathcal{T}}$, and has $n_{\mathcal{T}}$ terms in columns and n_{d_j} rows, with each of those representing the positions of terms inside the document. $\not{i} = 1 \dots n_{d_j}$ is the index (position) of the terms in the document d_j.

$$d^{(j)} = \begin{array}{c} \\ 1 \\ \vdots \\ n_{d_j} \end{array} \begin{array}{c} \begin{array}{ccc} 1 & \dots & n_{\mathcal{T}} \end{array} \\ \begin{bmatrix} & \vdots & \\ \cdots & \boldsymbol{d}^{(j)}_{\not{i}i} & \cdots \\ & \vdots & \end{bmatrix} \end{array} \tag{2}$$

$$\begin{bmatrix} n_{1j} & \dots & n_{\mathcal{T}j} \end{bmatrix}$$

The definition of the cell, $\boldsymbol{d}^{(j)}_{\not{i}i}$ of the binary matrix, $d^{(j)}$, is as follows:

$$\boldsymbol{d}^{(j)}_{\not{i}i} = \begin{cases} 1, & \text{if term } t_i \text{ appears at position } \not{i} \text{ of the document } d_j \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

The i-th column vector of the matrix $d^{(j)}$ has a 1 at each position of d_j that contains term $t_i \in \mathcal{T}$. And the marginal of the i-th column of matrix $d^{(j)}$ corresponds to $n_{ij} = \sum_{\not{i}=1\dots n_{d_j}} \boldsymbol{d}^{(j)}_{\not{i}i}$, and coincides with cell (i, j) of matrix X.

The row marginal of the $d^{(j)}$ matrix, $[n_{1j} \dots n_{\mathcal{T}j}]$, coincides with the j-th column of matrix X, and represents the profile of document d_j by indicating the distribution of terms in the document.

The set of these binary matrices is denoted by $\boldsymbol{\mathcal{D}} = \{d^{(1)}, d^{(2)}, \dots, denotedd^{(n_{\mathcal{D}})}\}$, and $\boldsymbol{\mathcal{D}}$ is the binary representation of a set of documents $\mathcal{D}$ in a corpus.

3. Latent Semantic Analysis and Extensions

LSA [8] infers the semantic structures of the terms in documents, and serves as the basis for the extraction of relevant textual patterns.

LSA is based on two-way factorial analysis, which uses Singular Value Decomposition (SVD). The term 'two' is due to the fact that terms and documents can be represented in the same latent (factorial) space. This allows us to analyse the relationships between them.

Given a set of documents $\mathcal{D}$, the first step is the numerization of the corpus, as described in Section 2.1 (i.e., building the X matrix).

Then, using SVD, matrix X can be decomposed as the product of three matrices:

$$X = \mathcal{V}_{(n_\mathcal{T} \times \mathcal{K})} \Lambda^{\frac{1}{2}}_{(\mathcal{K} \times \mathcal{K})} (\mathcal{U}')_{(\mathcal{K} \times n_\mathcal{D})} \tag{4}$$

where:

- $\mathcal{V}_{(n_\mathcal{T} \times \mathcal{K})}$ is the eigenvectors' matrix of XX';
- $\mathcal{U}_{(n_\mathcal{D} \times \mathcal{K})}$ is the eigenvectors' matrix of $X'X$;
- $\Lambda_{(\mathcal{K} \times \mathcal{K})}$ is the diagonal eigenvalues' matrix;
- $\mathcal{K} = \min\{n_\mathcal{T}, n_\mathcal{D}\}$ is the rank of X.

The eigenvectors of $\mathcal{U}$ identify $\mathcal{K}$ rotation directions over original documents that constitute the factorial space of terms. Each eigenvector is a linear combination of the original set of "document-variables" (i.e., the columns of X). For the particular case of TDM, these new artificial factors can be thought of as concepts or topics.

Let $\boldsymbol{u}_\alpha$ be one of the eigenvectors of the matrix $\mathcal{U}_{(n_\mathcal{D} \times \mathcal{K})}$, i.e., $\boldsymbol{u}_\alpha = \mathcal{U}_\alpha$. And, the projection of X on $\boldsymbol{u}_\alpha$, $\Psi_\alpha = X\boldsymbol{u}_\alpha$, is the α-th principal component. The corresponding eigenvalue λ_{kk} measures the information retained by Ψ_α from the total information in X [30]. In practice, SVD is often used for dimensionality reduction and visualization, such that only a subset of principal components is retained. Therefore, we will use K as the number of retained dimensions and $K \leq \mathcal{K}$.

There exists a rule of thumb to consider K as the number of components that keep an 80% of original information from X. However, there is no global consensus about the method to be used to determine the parameter K and, currently, it appears to be one of the research topics.

Taking only the K columns of $\mathcal{U}_{(n_\mathcal{D} \times \mathcal{K})}$ leads to $U_{(n_\mathcal{D} \times K)}$. And

$$X_K = V_{(n_\mathcal{T} \times K)} \Lambda^{\frac{1}{2}}_{(K \times K)} (U')_{(K \times n_\mathcal{D})} \tag{5}$$

is a lower dimensional representation of original data. $\Psi_K = XU_{(n_\mathcal{D} \times K)}$ is the projection of the original cloud of points X to the K-dimensional factorial subspace. The quantity of information retained in the K-dimensional subspace can be quantified by the sum of the corresponding K eigenvalues (stored in the diagonal matrix $\Lambda_{(K \times K)}$).

Both entities (terms and documents) admit joint representation onto a factorial space based on the transition relations between $\mathcal{V}$ and $\mathcal{U}$ [31] and the dual analysis of the columns of X and the rows. Joint representation becomes possible through the rescaling factor or a biplot representation [7,32].

To determine the meaning of factorial components, the contribution of terms to each factorial component identifies the subset of terms relevant in that factorial direction. As a consequence, this elicits the topic associated with the axis. The contribution of each term to the axis measures the degree of relationship between the term itself and the topic corresponding to the axis.

Moreover, to ensure that large documents do not distort the analysis, a previous normalization of X may be helpful. However, the quantity of information on a term in a document is inverse to the occurrence of the term in the corpus [30]. For this reason, a relevant improvement was to perform the LSA over a transformation function of X that takes this fact into consideration. Therefore, the term frequency–inverse document frequency, or tf-idf for short [33], can be applied to the corpus first, and then the LSA methodology can be applied.

In the textual data analysis, there are three leading issues. The first one is associated with synonymy. In this way, only a fraction of terms are held by the document, and those that are searched for by the group of users may not appear in it. Nevertheless, the document

remains relevant for that query, and should be listed as relevant. The second issue is about polysemy, which is classically treated with vocabularies and term coordination. The third obstacle is rare event detection, which means that whenever a pair of words appear together in very specific situations, their detection is much more effective than when they appear in almost all other possible scenarios.

The tests on the LSA method report that it is successful in handling synonymy scenarios (recall), but not so good with polysemy problems (precision). The third issue is not covered by LSA method, but is treated in a very elegant way in another factorial method, called Correspondence Analysis [7,10,34], by using a χ^2 metric.

LSA can be applied to different use cases. For instance, in [35], the authors apply LSA to word sense discrimination. Another example of an LSA application is related to the characterization of meaning similarities among words and entire passages, which is the synonymy problem [36].

One of the drawbacks of LSA is the need to periodically update, adding new terms and documents. Also, the difficulties that the method experiences for polysemy problems is due to the fact that the word is represented as a single point in the factorial space, leading to the weighted average of the different meanings it may have, as reported in [8] and in the LSA synonym test results' interpretation in [36]. Therefore, adding stemming as a pre-processing step could be not too meaningful. In addition, it does not consider any word order dependence and, therefore, both morphosyntactic as well as grammatical relationships are neglected.

3.1. Extensions of Latent Semantic Analysis

In this section, two extensions of Latent Semantic Analysis are presented, which overcome several limitations of LSA. The first method introduces the context of the words in the formulation of original LSA, and the second model takes into account the similarity among sentences.

3.1.1. Distributional Semantic Model

In order to overcome the context-awareness limitation of LSA, in [37], the authors present a Distributional Semantic Model. They extend the Vector Space Model representation by introducing the co-occurrence of the terms matrix, $C_{(n_\mathcal{T} \times n_f)}$, in which $n_\mathcal{T}$ is the number of terms in a document. Here, a more general view is adopted, and the terms can be reduced to the following entities, going from less to more elevated semantic or morphologic elements:

- Word;
- n-gram (a contiguous sequence of n items of text);
- Stem (part of the word to which affixes may be added);
- Lemma (canonical form of the word);
- Compound (lexeme consisting of more than one stem).

$\mathcal{T}_f = \{t_1, \ldots, t_{n_f}\} \subseteq \mathcal{T}$, which is an a priori chosen subset of terms used to evaluate the co-occurrences with document terms. The co-occurrence matrix C is as follows:

$$C = \begin{array}{c} \\ 1 \\ \vdots \\ n_\mathcal{T} \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & & \ldots & & n_f \end{array} \\ \begin{bmatrix} & & \vdots & & \\ & \cdots & c_{ij} & \cdots & \\ & & \vdots & & \end{bmatrix} \end{array} \tag{6}$$

where c_{ij} is the number of co-occurrences of terms t_i and t_j from $\mathcal{T}_f$, measured as the frequency of term t_i in the context of term t_j. c_{ij} is the number of occurrences of t_i in context of t_j. The context of t_j is a window of a certain number of positions around the term t_j, or

the whole sentence where t_j appears. The co-occurrence of t_i and t_j considers the context of t_j along all of the corpus (all sentences containing t_j or all windows around all occurrences of t_j).

The set of indexing features, $\mathcal{T}_f$, can be chosen by evaluating the discriminative power of the terms in the document collection. Then, the corpus matrix, over which the LSA is applied, is built as $X'C$. Matrix X is the usual TDM matrix as defined in Section 2.1 (in this context, it represents the lexical profiles provided, which are that terms must be more general than words), and the matrix C is the co-occurrences matrix described above in (6). The following equation is the formulation of the Distributional Semantic Model, where defined the objective function and a set of constraints are:

$$\max_{\boldsymbol{u}_\alpha} \boldsymbol{u}'_\alpha (X'C)'(X'C)\boldsymbol{u}_\alpha \text{ s.t. } \boldsymbol{u}'_\alpha \boldsymbol{u}_\alpha = 1 \tag{7}$$

3.1.2. Archetypal Analysis

In the domain of the multi-document summarization problem, in [5], the authors present a framework based on a content–graph joint model. It is based on Archetypal Analysis (AA) [38], which belongs to an optimization problem family similar to LSA/PCA, k-means, or NMF.

The matrix $X^{(\mathcal{S})}$ is defined as a term–sentence matrix (TSM) over a set $\mathcal{S}$ of $n_\mathcal{S}$ sentences (all sentences from the corpus). The dimensionality of the matrix is $n_\mathcal{T} \times n_\mathcal{S}$, and it is composed of rows, which correspond to terms, and columns, which correspond to sentences:

$$X^{(\mathcal{S})} = \begin{array}{c} \\ 1 \\ \vdots \\ n_\mathcal{T} \end{array} \begin{array}{c} \begin{array}{ccc} 1 & \dots & n_\mathcal{S} \end{array} \\ \begin{bmatrix} & \vdots & \\ \cdots & n_{ij}^{(\mathcal{S})} & \cdots \\ & \vdots & \end{bmatrix} \end{array} \tag{8}$$

where $n_{ij}^{(\mathcal{S})}$ is the frequency of term t_i in the j-th sentence of the whole corpus. The TSM matrix $X^{(\mathcal{S})}$ uses the term frequency–inverse sentence frequency (tf-isf) [33] weighting scheme (which is the same as tf-idf, but for sentences).

Additionally, the sentence similarity matrix, $A_{(n_\mathcal{S} \times n_\mathcal{S})}$, is computed for each pair of sentences using the cosine similarity between sentences. Let $X_j^{(\mathcal{S})}$ be the column of $X^{(\mathcal{S})}$ corresponding to sentence j. It is a vector containing the distribution of terms in the sentence. Each cell in matrix A is computed as follows:

$$a_{j,j'} = \cos(X_j^{(\mathcal{S})}, X_{j'}^{(\mathcal{S})}) = \frac{\sum_{i=1}^{n_\mathcal{T}} n_{ij}^{(\mathcal{S})} \cdot n_{ij'}^{(\mathcal{S})}}{\sqrt{\sum_{i=1}^{n_\mathcal{T}} \left(n_{ij}^{(\mathcal{S})}\right)^2} \sqrt{\sum_{i=1}^{n_\mathcal{T}} \left(n_{ij'}^{(\mathcal{S})}\right)^2}} \tag{9}$$

Therefore, the content–graph joint model consists of using both matrices, the term–sentence matrix, $X^{(\mathcal{S})}$, and the sentence similarity matrix, A, to provide the design matrix $J_{(n_\mathcal{T} \times n_\mathcal{S})} = X^{(\mathcal{S})}A$. The idea is to decompose the matrix J in $HW'J$ by using the AA technique. Equation (10) is the formulation of the AA as an optimization problem, where the objective function and a set of constraints are defined as:

$$\begin{aligned} &\min_{H,W} \quad \|J - H_{(n_\mathcal{T} \times K)} W' J\|^2 \\ &s.t. \sum_{k}^{K} h_k^i = 1, h_k^i \geq 0, \forall i \in \{1 \dots n_\mathcal{T}\} \\ &s.t. \sum_{i}^{n_\mathcal{T}} w_k^i = 1, h_k^i \geq 0, \forall k \in \{1 \dots K\} \end{aligned} \tag{10}$$

where $Y_{(n_S \times K)} = J'W$ is the matrix of archetypes (K columns). Those archetypes are built as convex combinations of data points or observations. $W_{(n_T \times K)}$ is the convex combination definition of J in a way that the columns of Y are located on the convex hull of the data point J [38]. Moreover, convex combinations of archetypes are used to approximate the observations. From $J \approx HY'$, it can be seen that the weighting matrix H approximates the archetypal space into the matrix J. In contrast to NMF [9], AA performs a decomposition of the matrix J into sparser stochastic matrices. The archetypes (the columns of Y) can be interpreted as topics. Indeed, each column of Y, $\boldsymbol{y}_k$, is:

$$\boldsymbol{y}_k = \sum_{i=1}^{n_T} (\boldsymbol{j}^i)' w_k^i \tag{11}$$

where $\boldsymbol{j}^i$ is the row profile of the design matrix, J, and hence $(\boldsymbol{j}^i)'$ is the corresponding vector. The above expression can also be regarded in matrix notation as follows:

$$\boldsymbol{y}_k = J' \boldsymbol{w}_k \tag{12}$$

According to [38], under certain conditions, a Y that minimizes the expression in (10) is a Y that maximizes $Y'J'JY$. Then, under those conditions, Y can be written as $Y = J'V\Lambda^{-\frac{1}{2}}$, and the columns are the principal directions of the $J'J$. Note that in those circumstances, Y would be U, and J would be the data matrix X in Section 3.

3.1.3. NMF Topic Modelling

The NMF decomposition method [9] imposes the non-negativity constraint of basis vectors, whereas SVD imposes an orthogonality constraint. Different optimization schemes can be used to derive the factorial space, such as the minimization of the least squares or the Kullback–Leibler divergence. In [39], the authors compare Latent Dirichlet Allocation (LDA), presented further in this work, and NMF, along with a k-means algorithm, to identify email threads. And for that particular use-case, the NMF method showed better performance, which suggests this technique occupies a relevant position among leading topic modelling methods.

3.1.4. Explicit Semantic Analysis

In [40], the authors present a novel approach, Explicit Semantic Analysis (ESA), to relate different fragments of text with a set of pre-defined Wikipedia-based concepts. As opposed to LSA, ESA maps each word to a set of pre-defined concepts (Wikipedia articles) and, by using the tf-idf numerical statistic from [33], builds the semantic interpretation vector, which provides a measure of relevance of the set of pre-defined concepts to the given text fragment.

4. Probabilistic Methods for Topic Modelling

4.1. Probabilistic PCA

In this section, a description of the Probabilistic Principal Component Analysis (PPCA) is presented. It is the probabilistic version of the Latent Semantic Analysis presented in Section 3.

PCA aims to maximize the projection of the original data space X onto the latent (factorial) space Ψ. Nevertheless, in the probabilistic setting, the link from factorial space Ψ to original space X is first established, and then the reverse mapping is derived by using the posterior distribution, which is achieved with Bayes theorem. PPCA is considered a linear Gaussian *latent variable* model [41–43].

The term profile, $\boldsymbol{x}^i$, is defined in ([41]) and it corresponds to the stochastic linear combination of its projection in the factorial space (see Section 3), namely $\boldsymbol{\psi}^i$, which is the i-th row of matrix Ψ, and a noise term

$$\boldsymbol{x}^i | \boldsymbol{\psi}^i = \boldsymbol{\mu_i} + \boldsymbol{\varepsilon}^i, \qquad \boldsymbol{\varepsilon}^i \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 I), \tag{13}$$

where

- Vector $\boldsymbol{\psi}^i$ is of dimensionality $\mathcal{K}$, usually $n_{\mathcal{D}}$;
- $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in_{\mathcal{D}}})$, is the expectation of $\boldsymbol{x}^i|\boldsymbol{\psi}^i$, and it is assumed to be a linear combination of the latent factors $\boldsymbol{\mu}_i = \boldsymbol{\mu} + W\boldsymbol{\psi}^i$, where:
  - $\boldsymbol{\mu}$ is the mean of $X_{n_{\mathcal{T}} \times n_{\mathcal{D}}}$, i.e., the centre of gravity of the original data cloud;
  - W is the parameter matrix, and it contains the *factor loadings*. The columns of W span the principal subspace of data, X, when the log-likelihood is maximized. The matrix W is defined as follows:

$$W_{(n_{\mathcal{D}} \times n_{\mathcal{D}})} = \mathcal{U}_{(n_{\mathcal{D}} \times n_{\mathcal{D}})}(\Lambda_{(n_{\mathcal{D}} \times n_{\mathcal{D}})} - \sigma^2 I_{(n_{\mathcal{D}} \times n_{\mathcal{D}})})^{\frac{1}{2})} R \tag{14}$$

    with $\mathcal{U}$ being the matrix of eigenvectors of $X'X$, as defined in (4).
    The diagonal matrix $\Lambda_{n_{\mathcal{D}} \times n_{\mathcal{D}}}$, containing the corresponding eigenvalues λ_k, and R is an arbitrary orthogonal matrix, which can be eventually found using the following:

$$W'W = R'(\Lambda - \sigma^2 I)R \tag{15}$$

    where the columns of R' are the singular vectors of the matrix $(W'W)$.
    In practice, matrix W is estimated through a computing method in an iterative way, which does not involve matrix R.

- $\sigma^2 I_{(n_{\mathcal{D}} \times n_{\mathcal{D}})}$ represents the matrix of variance and covariances of the noise, and it is the diagonal matrix $(n_{\mathcal{D}} \times n_{\mathcal{D}})$ with the variance of $\boldsymbol{x}^i|\boldsymbol{\psi}^i$ in each document of the corpus, which is assumed to be constant and equal to $\sigma^2 = V(\boldsymbol{x}^i|\boldsymbol{\psi}^i), \forall d \in 1 \ldots n_{\mathcal{D}}$ for all documents in the corpus, thus implying a homoschedasticity hypothesis. Also, the diagonal form of this matrix implies an assumption of independence, which means that the model assumes independence in the distribution of the word in all documents of the corpus given the latent variable.

$$\sigma^2 I_{(n_{\mathcal{D}} \times n_{\mathcal{D}})} = \begin{array}{c} \\ 1 \\ \\ \\ n_{\mathcal{D}} \end{array} \begin{array}{c} \begin{array}{cccc} 1 & & \ldots & n_{\mathcal{D}} \end{array} \\ \begin{bmatrix} \sigma^2 & & \mathbb{O} & \\ & \ddots & & \\ & & \sigma^2 & \\ \mathbb{O} & & & \ddots \\ & & & \sigma^2 \end{bmatrix} \end{array} \tag{16}$$

In the general case, it might happen that $\text{rank}(X) < n_{\mathcal{D}}$. Let us name $\mathcal{K} = \text{rank}(X)$, and

$$W_{n_{\mathcal{D}} \times \mathcal{K}} = \mathcal{U}_{n_{\mathcal{D}} \times \mathcal{K}}(\Lambda_{(\mathcal{K} \times \mathcal{K})} - \sigma^2 I_{\mathcal{K} \times \mathcal{K}})^{\frac{1}{2}} R \tag{17}$$

where $\mathcal{K} = \min\{n_{\mathcal{T}}, n_{\mathcal{D}}\}$, as defined in expression (4).

In practice, only $K < \mathcal{K}$ principal components are retained, and therefore matrix W is approximated accordingly:

$$W^*_{n_{\mathcal{D}} \times K} = U_{n_{\mathcal{D}} \times K}(\Lambda_{K \times K} - \sigma^2 I_{K \times K})^{\frac{1}{2}} R \tag{18}$$

where matrix W^* is of dimensionality $n_{\mathcal{D}} \times K$ with $K < \mathcal{K} = \text{rank}(X)$, and the diagonal matrix $\Lambda_{K \times K}$ is first composed of K eigenvalues λ_k, and $U_{n_{\mathcal{D}} \times K}$ also contains only the first K eigen vectors.

Moreover, it can be proven that whenever $K < \mathcal{K}$, then the residual variance $\sigma^2 > 0$ and the model covariance matrix $C = \sigma^2 I + W^* W^{*\prime}$ is nonsingular (which will be explained later). W^* is used to estimate $\hat{\boldsymbol{\mu}}^i$, and $\boldsymbol{\psi}^{i*} = (\boldsymbol{\psi}^i_1, \ldots, \boldsymbol{\psi}^i_K)$ is used instead.

Under the above conceptual frame, the prior distribution for $\boldsymbol{\psi}^i$ is defined as pure white noise:

$$\boldsymbol{\psi}^i \sim \mathcal{N}(\mathbf{0}, I) \tag{19}$$

According to the authors of [41–43], other choices of prior distribution would lead to probabilistic models equivalent to (21).

The predictive distribution of the data, $p(\boldsymbol{x}^i)$, is used to compute the parameters by maximum likelihood in a closed-form solution. It is defined using the Conditional Probability Law, and under the previous distributional assumptions, it follows a normal distribution:

$$p(\boldsymbol{x}^i \cap \boldsymbol{\psi}^i) = p(\boldsymbol{x}^i|\boldsymbol{\psi}^i)p(\boldsymbol{\psi}^i) \tag{20}$$

marginalizing $\boldsymbol{x}^i$

$$p(\boldsymbol{x}^i) = \int p(\boldsymbol{x}^i|\boldsymbol{\psi}^i)p(\boldsymbol{\psi}^i)d\boldsymbol{\psi}^i = \mathcal{N}(\boldsymbol{\mu}, C) \tag{21}$$

with $C_{n_\mathcal{D} \times n_\mathcal{D}}$ being a model covariance matrix. Using the fact that, for any random variables X and Y, this matrix can be derived from the following:

$$\begin{aligned} \mathrm{Cov}[\boldsymbol{x}^i|\boldsymbol{\psi}^i, \boldsymbol{x}^l|\boldsymbol{\psi}^i] &= \mathbb{E}[((\boldsymbol{\mu} + W\boldsymbol{\psi}^i + \boldsymbol{\varepsilon}^i) - \mathbb{E}[\boldsymbol{x}^i|\boldsymbol{\psi}^i])((\boldsymbol{\mu} + W\boldsymbol{\psi}^i + \boldsymbol{\varepsilon}^l) - \mathbb{E}[\boldsymbol{x}^l|\boldsymbol{\psi}^i])] \\ &= \mathbb{E}[(W\boldsymbol{\psi}^i + \boldsymbol{\varepsilon}^i)(W\boldsymbol{\psi}^i + \boldsymbol{\varepsilon}^l)'] \\ &= \mathbb{E}[W\boldsymbol{\psi}^i\boldsymbol{\psi}^{i\prime}W'] + \mathbb{E}[\boldsymbol{\varepsilon}^i\boldsymbol{\varepsilon}^{l\prime}] \\ &= WW' + \sigma^2 I \end{aligned} \tag{22}$$

where an assumption has been made that $\boldsymbol{\psi}^i$, as well as $\boldsymbol{\varepsilon}^i$, are independent random variables (hence are uncorrelated). Then, it follows that $C_{n_\mathcal{D} \times n_\mathcal{D}} = \sigma^2 I + W^* W^{*\prime}$, since $\boldsymbol{\psi}^i$ is white noise and does not add variance to the result.

By using the Bayes Law, the posterior distribution of the latent variables ($\boldsymbol{\psi}^i$) is derived as follows:

$$p(\boldsymbol{\psi}^i|\boldsymbol{x}^i) = \mathcal{N}\left(M^{-1}W'(\boldsymbol{x}^i - \boldsymbol{\mu}),\ \sigma^2 M^{-1}\right) \tag{23}$$

where the matrix $M_{K \times K} = \sigma^2 I + W'W$.

And the marginal log-likelihood of X is as follows:

$$\mathcal{L}(\boldsymbol{\mu}, \sigma^2, W) = \sum_{i=1}^{n_\mathcal{T}} \log\left\{p(\boldsymbol{x}^i)\right\} = -\frac{n_\mathcal{T}}{2}\left\{n_\mathcal{D}\log(2\pi) + \log|C| + \mathrm{tr}(C^{-1}S)\right\} \tag{24}$$

where S is the empirical covariance matrix of $X_{n_\mathcal{T} \times n_\mathcal{D}}$ and $\pi = 3.1415\ldots$

The SVD of $\hat{W}'\hat{W}$ leads to

$$\hat{U} = \hat{W}R'(\hat{\Lambda} - \hat{\sigma}^2 I)^{-\frac{1}{2}} \tag{25}$$

Now, the estimates of $\hat{U}$ and $\hat{\Lambda}$ can be used for the projection of PPCA model. Furthermore, this formulation produces an approximation to the same axes obtained with LSA or PCA.

Furthermore, there is a mixture component version of PPCA, namely the Mixture Probabilistic Principal Component Analysis (MPPCA) [42]. This latter method, according to our experiments with it in the text processing domain, provides a higher degree of flexibility, as the Normality assumption still holds, although for each component individually. Therefore, it becomes a very interesting approach as a topic modelling technique.

4.2. Bayesian Mixture Models

In this section, the main probabilistic approaches in textual data analysis will be discussed. The main difference from factorial techniques is that the probabilistic models are based on the probabilistic framework for parameter estimation and provide a complete distributional output, whereas the former provide an optimal solution in terms of maximizing a given objective function and, depending on this objective function and a set of constraints, different types of solutions are derived.

First, Bayesian mixture models will be presented. Then, LDA is presented as an extension of the mixture models. Finally, applications and variations of the previous models from the existing literature are discussed.

First of all, a brief notational framework for Bayesian models will be presented. Essentially, there are a few concepts that have to be represented: a document, a word, and a topic.

Let Z be a qualitative r.v. that indicates which topic is observed for a document or other textual unit. Z is a discrete r.v. with values in $\mathcal{Z} = \{z_1, \ldots, z_K\}$, and its probability space is:

$$\langle \mathcal{Z}, \mathscr{P}(\mathcal{Z}), P_{\mathcal{Z}} \rangle \tag{26}$$

where $\mathcal{Z}$ is the sample space (i.e., the set of topics), $\mathscr{P}(\mathcal{Z})$ is defined as all subsets of $\mathcal{Z}$, and $P_{\mathcal{Z}}$ appears to be the probability function of $\mathscr{P}(\mathcal{Z})$. Then, $P_{\mathcal{Z}}$ is built on top of $p_{\mathcal{Z}} = P(Z = z_k)$ for $k = 1 \ldots K$, given that $\mathscr{P}(\mathcal{Z})$ is a σ-algebra. Thus, $p_{\mathcal{Z}}$ is the prior probability distribution of topics (not to be confused with Bayesian *prior* parameter distribution).

As said before, a document $d \in \mathcal{D}$ is defined as a sequence of words, $(w_1, \ldots, w_{n_d}) \in \mathcal{T}^{n_d}$, where $\mathcal{T}$ is the set of possible terms (i.e., the vocabulary), and n_d is the length of document d. The set of terms that can occur in the position $\not{p} \in 1 \ldots n_d$ of a document d is defined as $\mathcal{T}_{\not{p}}$, and it holds that $\mathcal{T}_{\not{p}} = \mathcal{T}$, $\forall \not{p} \in 1 \ldots n_d$. Let $D_{\not{p}}$ be an r.v. that indicates which term is observed in a position $\not{p}$ of the document d. $D_{\not{p}}$ is a discrete r.v. with values in $\mathcal{T}_{\not{p}}$ $(= \mathcal{T})$, and its probability space is:

$$\langle \mathcal{T}, \mathscr{P}(\mathcal{T}), P_{\mathcal{T}} \rangle \tag{27}$$

where $\mathcal{T}$ is the sample space (i.e., the possible terms that can appear at any position $\not{p}$ of a document), $\mathscr{P}(\mathcal{T})$ is the σ-algebra of events set (all subsets of $\mathcal{T}$), and $P_{\mathcal{T}}$ is the associated probability function of $\mathscr{P}(\mathcal{T})$ for observing the different words.

Given that a document $d \in \mathcal{D}$ of length n_d is considered as a sequence of words, it can now be defined as a random vector $D = (D_1, \ldots, D_{\not{p}}, \ldots, D_{n_d})$ that considers all the combinations of words that can appear along an entire document d of length n_d, and it is associated with the following probability space:

$$\langle \mathcal{T}^{n_d}, \mathscr{P}(\mathcal{T}^{n_d}), P_{\mathcal{T}^{n_d}} \rangle \tag{28}$$

where $\mathcal{T}^{n_d}$ is the Cartesian product of $\mathcal{T}$, and $\mathscr{P}(\mathcal{T}^{n_d})$ is the σ-algebra of parts of $\mathcal{T}^{n_d}$. $P_{\mathcal{T}^{n_d}}$ is built on top of $p_{\mathcal{T}^{n_d}} = P(D = d)$ with $d \in \mathcal{D}$, and it is the probability of observing a certain document d:

$$P(D = d) = P(D = (w_1, \ldots, w_{n_d})) \tag{29}$$

Many authors [44] have developed this joint probability in the form:

$$P(D = d) = P(\bigwedge_{\not{p}=1:n_d} D_{\not{p}} = w_{\not{p}}) \tag{30}$$

In the probabilistic mixture models [43,45], the data are generated by one of the mixture components. For instance, it is proven that a mixture of Gaussians can approximate any type of continuous distributions, in particular multimodal ones [46]. In this work, each mixture component will be considered to correspond to a topic.

The probability of a document can be also written in terms of a set of possible topics. Applying the law of total probabilities and considering a given set of topics $\mathcal{Z}$, the probability of document can be expressed in terms of conditional probabilities with regard to topics $\mathcal{Z}$ as follows:

$$P(D = d) = \sum_{k=1}^{K} P(D = d | Z = z_k) P(Z = z_k) \tag{31}$$

Bayesian topic models are generative by nature, and the documents are generated by a mixture model parameterized by a set of parameters θ. Hence, the probability model of the document formulated in (31) can be expressed in terms of parameters θ:

$$P(D = d|\theta) = \sum_{k=1}^{K} P(D = d|Z = z_k \wedge \theta)P(Z = z_k|\theta) \tag{32}$$

Now, the term $P(D = d|Z = z_k \wedge \theta)$ from the expression (32), which corresponds to the probability of observing a document d given the topic $Z = z_k$ and a set of parameters θ, can be defined on the basis that the document d can be represented as a sequence of n_d words $d = (w_1, \ldots, w_{n_d})$, by using expression (30):

$$P(D = d|Z = z_k \wedge \theta) = P(\bigwedge_{\not{}=1:n_d} D_{\not{}} = w_{\not{}}|Z = z_k \wedge \theta) \tag{33}$$

where each discrete random variable $D_{\not{}}$ indicates which word occurs in position $\not{}$ of the document, and it has its corresponding probability space defined in (27).

Now, taking into account the whole corpus of documents $\mathcal{D}$, and assuming that documents behave as an iid sample, the likelihood function of θ in the corpus $\mathcal{D}$ is derived as follows:

$$\mathcal{L}(\theta) = P(\mathcal{D}|\theta) = \prod_{j=1}^{n_\mathcal{D}} P(D = d_j|\theta) \tag{34}$$

Now, substituting the expression (33) into (32), and using it in (34), the likelihood function becomes:

$$\mathcal{L}(\theta) = \prod_{j=1}^{n_\mathcal{D}} \sum_{k=1}^{K} P(\bigwedge_{\not{}=1:n_{d_j}} D_{\not{}} = w_{\not{}}|Z = z_k \wedge \theta)P(Z = z_k|\theta) \tag{35}$$

4.2.1. Considering the Document Size

In this formulation, the fact that the length of the documents can be modelled as a random variable as well has been omitted. In [47], the authors consider this scenario and claim the importance of taking the length of the documents as a random variable into account, and also conditioning it to the topic.

Let N be a random variable indicating the length of a document $d \in \mathcal{D}$, with the values of N being in $[1, \infty)$, and let us name n_d the length of a specific document d, $\forall\, d \in \mathcal{D}$. In this way, expression (29) would develop into a different expression:

$$P(D = d) = P(D = (w_1, \ldots, w_{n_d}) \wedge N = n_d) \tag{36}$$

and, consequently, expression (33) becomes:

$$P(D = d|Z = z_k \wedge \theta) = P\Big((\bigwedge_{\not{}=1:n_d} D_{\not{}} = w_{\not{}}) \wedge N = n_d|Z = z_k \wedge \theta\Big) \tag{37}$$

In principle, it can be assumed that the size of the document and the sequence of words are independent, so (37) can be written as:

$$\begin{aligned} P(D = d|Z = z_k \wedge \theta) = {} & P(N = n_d|Z = z_k \wedge \theta) \\ & \cdot P(\bigwedge_{\not{}=1:n_d} D_{\not{}} = w_{\not{}}|Z = z_k \wedge \theta) \end{aligned} \tag{38}$$

Finally, incorporating all the assumptions and notations, the expression of the generic likelihood function can be rewritten from (35) as follows:

$$\mathcal{L}(\theta) = \prod_{j=1}^{n_{\mathcal{D}}} \sum_{k=1}^{K} \Big(P(N = n_{d_j} | Z = z_k \wedge \theta) \cdot P(\bigwedge_{\not=1:n_{d_j}} D_{\not} = w_{\not} | Z = z_k \wedge \theta) \cdot P(Z = z_k | \theta) \Big) \tag{39}$$

4.2.2. Generative Model

In [19], the authors develop expression (39) under the assumption of the independence of word occurrences in several positions of a document. This means to assume that the probability of the occurrence of a word in a document is constant with regard to the position in the document. This type of model corresponds to a family of n-gram models described in [44], specifically to the 1-gram model. The joint probability term from the expression (39) for a given document $d \in \mathcal{D}$ has the following form:

$$P(\bigwedge_{\not=1:n_d} D_{\not} = w_{\not} | Z = z_k \wedge \theta) = \prod_{\not=1}^{n_d} P(D_{\not} = w_{\not} | Z = z_k \wedge \theta) \tag{40}$$

where $D_{\not}$ is the discrete random variable defined earlier, with its corresponding probability space (27). Variable $D_{\not}$ follows a Categorical distribution:

$$D_{\not} \sim \text{Cat}(\pi_{1d\not}, \dots \pi_{n_{\mathcal{T}} d\not}) \tag{41}$$

where $\pi_{id\not}$ is the probability that term $t_i \in \mathcal{T}$ appears in position $\not \in [1 \dots n_d]$ of document $d \in \mathcal{D}$. When the topic is known:

$$D_{\not} | Z = z_k \sim \text{Cat}(\pi_{1d\not k}, \dots \pi_{n_{\mathcal{T}} d\not k}) \tag{42}$$

Thus, expression (39) becomes

$$\mathcal{L}(\theta) = \prod_{j=1}^{n_{\mathcal{D}}} \sum_{k=1}^{K} \Big(P(N = n_{d_j} | Z = z_k \wedge \theta) \cdot \prod_{\not=1:n_{d_j}} P(D_{\not} = w_{\not} | Z = z_k \wedge \theta) \cdot P(Z = z_k | \theta) \Big) \tag{43}$$

$\pi_{id\not k}$ is the probability of a term t_i appearing in position $\not$ of document d, given topic $Z = z_k$

$$\mathcal{L}(\theta) = \prod_{j=1}^{n_{\mathcal{D}}} \sum_{k=1}^{K} \Big(P(N = n_{d_j} | Z = z_k \wedge \theta) \cdot \Big(\prod_{\not=1:n_{d_j}} \pi_{id\not k} \Big) \cdot P(Z = z_k | \theta) \Big) \tag{44}$$

The authors assume that the probability of a word stays constant for all documents, and is also independent of the words at other positions of the document. Also, it is independent of the position where the word is observed (conditioned on the topic and parameters [19]). Hence,

$$\pi_{id\not k} = \pi_{ik} \qquad \forall d, \not$$

where the subindices d and $\not{}$ have been removed by conditioning the variable $D_{\not{}}$ only on topic $Z = z_k$ and θ. So, definition (42) takes the following form:

$$D_{\not{}}|Z = z_k \wedge \theta \sim \text{Cat}(\pi_{1k}, \dots \pi_{n_{\mathcal{T}}k}) \tag{45}$$

where π_{ik} is the probability of the occurrence of term $t_i \in \mathcal{T}$ (given the topic $Z = z_k$). Likewise, the likelihood function in (39) can be rewritten as:

$$\begin{aligned}\mathcal{L}(\theta) = \prod_{j=1}^{n_{\mathcal{D}}} \sum_{k=1}^{K} \Bigg(& P(N = n_{d_j}|Z = z_k \wedge \theta) \\ & \cdot \left(\prod_{i=1}^{n_{\mathcal{T}}} \pi_{ik}^{n_{ij}} \right) \cdot P(Z = z_k|\theta) \Bigg)\end{aligned} \tag{46}$$

where n_{ij} is the number of occurrences of term t_i in document d_j. Therefore, the probability of a certain sequence of terms is the product of their corresponding probabilities (π_{ik}), and terms repeat for all positions $\not{}$ containing the same term, so that they can be factorized in term $\pi_{ik}^{n_{ij}}$, and the product moves to iterate over the vocabulary, instead of iterating over the positions of the document.

4.2.3. Multinomial Model

Finally, in [47], the authors consider another possibility based on the Multinomial model, which assumes that the words in the document follow a Multinomial distribution, and it can be formalized as follows.

This model takes into account the number of times the term t_i appears in the document, still disregarding the positions where the term t_i appears, as it is assumed in the Generative model presented in Section 4.2.2. For each term $t_i \in \mathcal{T}$, a random variable Q_i is defined as the number of occurrences of term t_i in document d. The realization of this variable is related to the matrix defined in Section 2.2 as a possible representation of a document through a binary matrix, and $Q_i = \sum_{\not{}=1}^{n_d} \boldsymbol{d}_{\not{}i} = n_i$, as introduced in Section 2.1. In fact, by construction, $Q_i \sim \text{Bin}(n_d, \pi_{id})$, where π_{id} is the probability of occurrence of term t_i in the document d. Conditioning the variable Q_i to the topic $Z = z_k$ and a set of parameters θ, its distribution changes to $Q_i|Z = z_k \wedge \theta \sim \text{Bin}(n_d, \pi_{idk})$. One of the assumptions is that the probability π_{idk} stays constant and independent of the words that occur on other positions (and documents), given that the variable Q_i is conditioned on topic $Z = z_k$ and parameter θ [19]. This leads to a redefinition of Q_i, such that $Q_i|Z = z_k \wedge \theta \sim \text{Bin}(n_d, \pi_{ik})$. A spectral representation of the document takes the occurrences of all possible words into account. Hence, the random vector $Q = (Q_1, Q_2, \dots, Q_{n_{\mathcal{T}}})$ describes the distribution of words of a given document, and $Q|Z = z_k \wedge \theta \sim \text{Mult}(n_d, \pi_{1k}, \dots \pi_{n_{\mathcal{T}}k})$ follows a Multinomial distribution. Therefore, the term $P(D = d|Z = z_k \wedge \theta)$ can be formulated in terms of Q, and expression (38) becomes:

$$\begin{aligned}P(D = d|Z = z_k \wedge \theta) = & P(N = n_d|Z = z_k \wedge \theta) \\ & \cdot P(Q = (n_{1d}, \dots, n_{\mathcal{T}d})|Z = z_k \wedge \theta)\end{aligned} \tag{47}$$

According to the Multinomial distribution, the term $P(Q = (n_{1d}, \dots, n_{\mathcal{T}d})|Z = z_k \wedge \theta)$ from expression (47) can be rewritten as

$$P(Q = (n_{1d}, \dots, n_{\mathcal{T}d})|Z = z_k \wedge \theta) = \frac{n_d!}{\prod_{i=1}^{n_{\mathcal{T}}} n_{id}!} \prod_{i=1}^{n_{\mathcal{T}}} P(Q_i = n_{id}|Z = z_k \wedge \theta) \tag{48}$$

Now, using the assumptions and definition of Multinomial distribution, the expression (48) is transformed into:

$$P(Q = (n_{1d}, \dots, n_{\mathcal{T}d})|Z = z_k \wedge \theta) = \frac{n_d!}{\prod_{i=1}^{n_{\mathcal{T}}} n_{id}!} \prod_{i=1}^{n_{\mathcal{T}}} \pi_{ik}^{n_{id}} \tag{49}$$

One of the drawbacks of this model is that sentences with same words in different sequences account for the same spectral vector. For example, the sentence ***white cat*** *in the car* would be the same event as *cat in the* ***white car***.

Finally, considering all the assumptions presented above, and also using (35) and (49), the likelihood of the Multinomial model would be as follows:

$$\mathcal{L}(\theta) = \prod_{j=1}^{n_{\mathcal{D}}} \sum_{k=1}^{K} \Bigg[P(N = n_{d_j} | Z = z_k \wedge \theta) \cdot \frac{n_{d_j}!}{\prod_{i=1}^{n_{\mathcal{T}}} n_{ij}!} \prod_{i=1}^{n_{\mathcal{T}}} \pi_{ik}^{n_{ij}} \cdot P(Z = z_k | \theta) \Bigg] \tag{50}$$

4.2.4. Multivariate Bernoulli Model

The Multivariate Bernoulli model [47] only considers whether the term t_i actually appears in the document, without taking into account the number of occurrences. Therefore, this model does not incorporate the length of the document into its formulation. The Bernoulli random variable $X_d^i \sim \text{Bern}(\pi_{id})$ indicates whether term t_i appears in a document d. $x_j^i \in \{0, 1\}$ is the realization of the variable X_j^i for a given document d_j, and it states whether term t_i is present or not in the document d_j.

The likelihood function of the Multivariate Bernoulli model is:

$$\mathcal{L}(\theta) = \prod_{j=1}^{n_{\mathcal{D}}} \sum_{k=1}^{K} \left[\left(\prod_{i=1}^{n_{\mathcal{T}}} \left(x_j^i \pi_{ik} + (1 - x_j^i)(1 - \pi_{ik}) \right) \right) P(Z = z_k | \theta) \right] \tag{51}$$

4.2.5. Latent Dirichlet Allocation

LDA is a Bayesian hierarchical model [13] based on the *bag-of-words* assumption or, more formally, the *exchangeability* assumption [48]. Exchangeability means that the order of words can be neglected, i.e., they can be permuted inside the document. Therefore, it means that the random variables representing the words in a document ($D_{\not}$) are assumed to be conditionally (with respect to an underlying latent parameter) independent and identically distributed, and are supposed to be exchangeable. This implies that the conditional distribution, conditioned on some latent parameter, can be factorized easily; however, the marginal distribution over this latent parameter may be quite involved. This model has been proven to provide very appealing results. For instance, it was applied to the customer care/call-centre use case and compared to other methods, such as NMF, Neural-LDA and Contextualized Topic Modelling [49], and also in the fields of language analysis/therapy [50] and financial markets [51].

Unlike the models presented in Section 4.2, where it was assumed that the document is associated with one topic at a time (i.e., the probability distribution of the document is conditioned on topic $Z = z_k$), in LDA, the authors do not make this assumption and provide additional flexibility to the model by allowing the documents to be associated with multiple topics simultaneously [13]. In (26), the probability space related with the topic of a document was introduced as follows:

$$\langle \mathcal{Z}, \mathscr{P}(\mathcal{Z}), P_{\mathcal{Z}} \rangle \tag{52}$$

Following the approach presented by [13], a different topic can be observed on each position of a given document.

Also, all previous approaches presented in this work assume a traditional categorical probability distribution for $\mathcal{Z}$. In the LDA approach, the topics are considered random variables distributed as $z \sim \text{Cat}(\zeta)$, and ζ is considered a random variable in turn, following a Dirichlet distribution of parameter α (in the original work, the authors use θ, which we transform into ζ in this work, to avoid notation conflicts with the rest of the paper).

In the following, we will try to refer the LDA approach to the general notation used in our work.

Let $Z_{\not}$ be a discrete r.v. that indicates which topic is observed on position $\not$ of document d. In the LDA approach, it is assumed that a different probability space $Z_{\not}$ is associated with each position $\not$ of the document, provided that the probability space of all $Z_{\not}$ is a copy of the one defined in expression (26).

Now, for a given document $d \in \mathcal{D}$ of length n_d, a random vector

$$\mathbb{Z} = (Z_1, \ldots, Z_{\not}, \ldots, Z_{n_d})$$

can be associated with the document, considering all the possible sequences of topics that can occur along the sequence of words of an entire document d of length n_d. The probability space of the multiple topics observed in a document can be modelled in the following probability space:

$$\langle \mathcal{Z}^{n_d}, \mathscr{P}(\mathcal{Z}^{n_d}), P_{\mathcal{Z}^{n_d}} \rangle \tag{53}$$

where $\mathcal{Z}^{n_d}$ is the Cartesian product of $\mathcal{Z}$, and $\mathscr{P}(\mathcal{Z}^{n_d})$ is the σ-algebra of parts of $\mathcal{Z}^{n_d}$. $P_{\mathcal{Z}^{n_d}}$ is built on top of $p_{\mathbb{Z}} = P(\mathbb{Z} = (z_1, z_2, ...z_{n_d}))$, which is the probability of observing a certain sequence of topics, of length n_d, for a given document $d \in \mathcal{D}$.

It is worth mentioning that the authors consider the length of a document, the variable $N = n_d$ but, in general, they ignore its randomness in their developments [13].

The model is presented as a joint probability of the joint occurrence of the random vector D (the sequence of words in the document), and the random vector $\mathbb{Z}$ (defined in (53)), the sequence of topics of the document:

$$P(D = d \wedge \mathbb{Z} = (z_1, z_2, \ldots, z_{n_d})) = P(\bigwedge_{\not=1:n_d} (D_{\not} = w_{\not} \wedge Z_{\not} = z_{\not})) \tag{54}$$

In expression (54), the authors make an independence assumption that the pairs, word, and topic are independent from the words and topics observed in other positions. Then, expression (54) is transformed as follows:

$$P(D = d \wedge \mathbb{Z} = (z_1, z_2, \ldots, z_{n_d})) = \prod_{\not=1}^{n_d} P(D_{\not} = w_{\not} \wedge Z_{\not} = z_{\not}) \tag{55}$$

The value $z_{\not}$ corresponds to the topic observed in position $\not$ of the document. This topic is one of the possible elements of $Z_{\not}$. Since all $Z_{\not}$ are copies of $\mathcal{Z}$, all of them take values of $z_1, \ldots, z_K$. Thus, the probability of observing a concrete word $w_{\not}$ in a certain position $\not$ is

$$P(D_{\not} = w_{\not}) = P(D_{\not} = w_{\not} \wedge (\bigvee_{k=1}^{K} Z_{\not} = z_k)) = P(\bigvee_{k=1}^{K} (D_{\not} = w_{\not} \wedge Z_{\not} = z_k)) \tag{56}$$

Assuming that topics do not interact among them,

$$P(D_{\not} = w_{\not}) = P(\bigvee_{k=1}^{K} (D_{\not} = w_{\not} \wedge Z_{\not} = z_k)) = \sum_{k=1}^{K} P(D_{\not} = w_{\not} \wedge Z_{\not} = z_k) \tag{57}$$

Substituting expression (57) into (55), and marginalizing with respect to topics:

$$P(D = d) = \prod_{\not=1}^{n_d} \sum_{k=1}^{K} P(D_{\not} = w_{\not} \wedge Z_{\not} = z_k) \tag{58}$$

which corresponds to marginalization with respect to all possible topics that can occur in a position of the document.

Applying the properties of conditional probability (55) becomes:

$$P(D = d) = \prod_{\not{n}=1}^{n_d} \sum_{k=1}^{K} P(D_{\not{n}} = w_{\not{n}} | Z_{\not{n}} = z_k) P(Z_{\not{n}} = z_k) \tag{59}$$

where $P(D_{\not{n}} = w_{\not{n}} | Z_{\not{n}} = z_k)$ is a Categorical probability distribution for the word in position $\not{n}$, considering the topics observable in position $\not{n}$, although the authors refer to it as a Multinomial distribution [13].

In this approach, the authors of [13] assume that the topics are characterized by a distribution over words. Each topic $Z_{\not{n}}$ follows a Categorical distribution with parameter ζ. The parameter ζ, in turn, is considered a random variable that follows a Dirichlet distribution of parameter α, ($\zeta \sim \text{Dir}(\alpha)$), the so-called *hyper*-parameters in the Bayesian setting (because they are the parameters of parameters).

The joint probability of observing the document $D = d$ and a certain mixture of topics determined by parameters ζ can be written as follows:

$$P(D = d \wedge \zeta) = P(\zeta) \prod_{\not{n}=1}^{n_d} \sum_{k=1}^{K} P(D_{\not{n}} = w_{\not{n}} | Z_{\not{n}} = z_k \wedge \zeta) P(Z_{\not{n}} = z_k | \zeta) \tag{60}$$

ζ can be omitted from first multiplication term in the summation, provided that the probabilities of words are not affected by them. Moreover, conditioning the previous expression (60) to the distributional parameters (α and θ) gives:

$$P(D = d \wedge \zeta | \alpha \wedge \theta) = P(\zeta | \alpha) \prod_{\not{n}=1}^{n_d} \sum_{k=1}^{K} P(D_{\not{n}} = w_{\not{n}} | Z_{\not{n}} = z_k \wedge \theta) P(Z_{\not{n}} = z_k | \zeta) \tag{61}$$

The marginal distribution of a document can be obtained by integrating expression (61) over ζ gives:

$$\begin{aligned} &P(D = d | \alpha \wedge \theta) = \\ &\int P(\zeta | \alpha) \left(\prod_{\not{n}=1}^{n_d} \sum_{k=1}^{K} P(D_{\not{n}} = w_{\not{n}} | Z_{\not{n}} = z_k \wedge \theta) P(Z_{\not{n}} = z_k | \zeta) \right) d\zeta \end{aligned} \tag{62}$$

Assuming that the documents are an iid sample, the probability of a whole corpus $\mathcal{D}$, i.e., the likelihood of the parameters α and θ in the corpus $\mathcal{D}$, can written as follows:

$$\begin{aligned} \mathcal{L}(\alpha, \theta) &= \prod_{j=1}^{n_{\mathcal{D}}} P(D = d_j | \alpha \wedge \theta) \\ &= \prod_{j=1}^{n_{\mathcal{D}}} \int P(\zeta_j | \alpha) \left(\prod_{\not{n}=1}^{n_{d_j}} \sum_{k=1}^{K} P(D_{\not{n}} = w_{\not{n}} | Z_{\not{n}} = z_k \wedge \theta) P(Z_{\not{n}} = z_k | \zeta_j) \right) d\zeta_j \end{aligned} \tag{63}$$

So, (63) is very difficult to compute [13] and, for this reason, the approach adopted in [13] is to use the variational inference method. It consists of finding the lower bound of the log-likelihood function (logarithm of expression (63)), and then this lower bound is maximized with respect to parameters α and θ. This is done iteratively, using the Expectation Maximization method. This procedure is called the empirical Bayes parameter estimation.

4.3. Robust Estimators

For the methods presented above, in particular those that involve M-estimator (i.e., the likelihood function in this case, or its equivalent), it is convenient to consider the need to efficiently and robustly estimate the parameters. This is due to the fact that the breakdown point of some methods is low, which leads to poor generalization, given that with a small amount of outlying observations, the estimated parameters can be completely invalid.

For instance, in [20], the authors propose an optimization algorithm with the LASSO penalization schema [52] and with the least trimmed squares model [53]. They apply their method to medical data with promising results in terms of MSE. In [21], the authors go one step beyond and propose two mixed-integer nonlinear optimization models to deal with outliers, as well as multicollinearity at the same time.

5. Experiments

In order to effectively compare the two families of methods, factorial and probabilistic, in this work, the experimental setup is composed of two steps. First, data preprocessing, and second, model fitting and the gathering of the results to later extract insights and compare different methods' families.

The pre-processing consists of the following steps:

1. Morphosyntactic analysis and lemmatization;
2. Word Sense Disambiguation (WSD) with Part-Of-Speech (POS) tagging;
3. Filtering by words composed of alphanumeric characters and at least length 3;
4. Stopwords filtering;
5. Filtering of word categories: Nouns, Verbs and Adjectives;
6. Filtering of terms that at least appear 4 times in the set of documents.

Steps 1, 2, and 5 were accomplished with the help of the Freeling tool [54]. Moreover, by introducing the first two steps in the pre-processing stage, the model fitting process showed much more consistent and robust results, leading the different methods to effectively extract meaningful patterns; e.g., in the factorial methods, the inertia captured by the factorial axes was observed to be much more uniform due to elimination of undesired artefacts (highly frequent irrelevant words) or polysemy effects, which were removed with WSD treatment.

Regarding the data, the experiments are based on the Reuters-21578 R8 database [55]. It consists of 7674 news documents of eight different classes. The data are split into two files, one for the training set and the other for the testing set. Each file contains documents in rows, with the first string being the class of the document, and the rest of the line corresponding to the document itself. The lines are of variable length, and the mean length of the article is 102 words with standard deviation of 118, given that 75% of documents do not surpass 113 words and the maximum is 964.

This dataset is very interesting due to its widely accepted benchmark status, lexical variety, complete sentences, grammatical richness and, hence, its suitability for the application of Natural Language Processing elements, the availability of exhaustive documentation and sample size. Additionally, the documents are already labelled, which may be very useful even in exploratory tasks.

In this first experimentation phase, 12 documents of the 3 classes have been randomly selected, which are crude, money-fx, and trade. This is because at this point the goal is purely exploratory, and to compare different methods, not benchmarking.

In order to extract meaningful results and simplicity in interpretation, only a few terms would have been selected by Equation (64).

$$U = X'V\Lambda^{-\frac{1}{2}} \tag{64}$$

where $X = V\Lambda^{\frac{1}{2}}U'$ is an SVD of matrix X, $U \in \mathbb{R}^d$, and $U'U = I$. Therefore, each column j of matrix $V\Lambda^{-\frac{1}{2}}$ is the contribution of terms on principal direction $\boldsymbol{u}_j$.

Similarly, for probabilistic methods, terms' distribution for each topic is used to select most relevant terms.

Results

In this section, will analyse the patterns that have been found by different methods from different families, which will also be compared in both quantitative as well as qualitative ways.

In Figure 1, the projection of the synsets (terms) onto the LSA latent space is represented. The fact that we were working with synsets instead of words by leveraging WSD

allowed us to capture more meaningful statistical information from the data, as many words were mapped to only one synset taking the canonical form of the word (lemmatization). Also, the polysemy effect was removed, which helped LSA to derive the latent space more accurately (see Section 3).

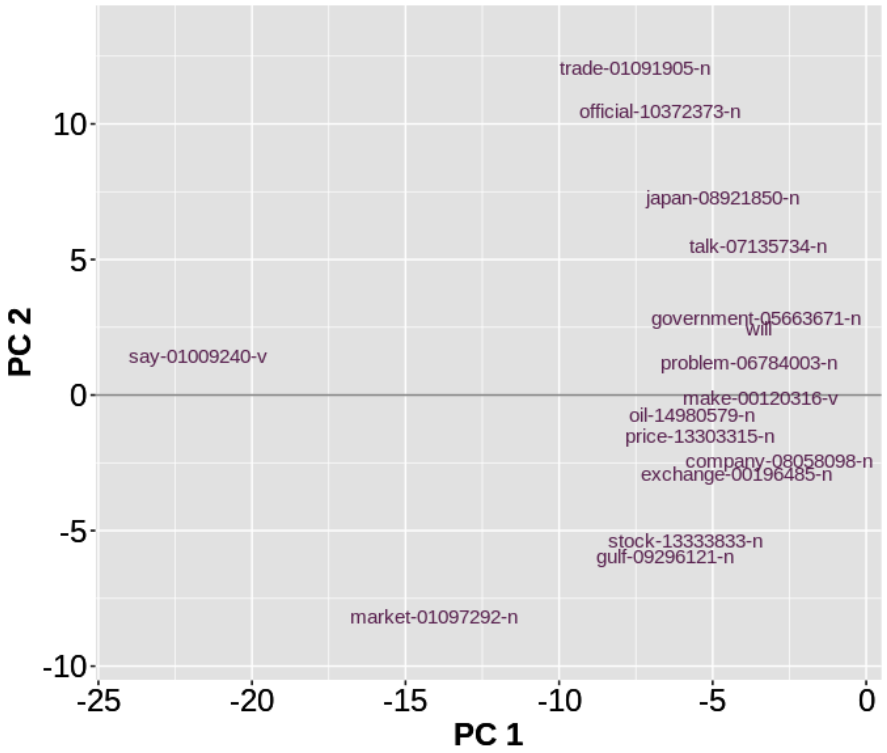
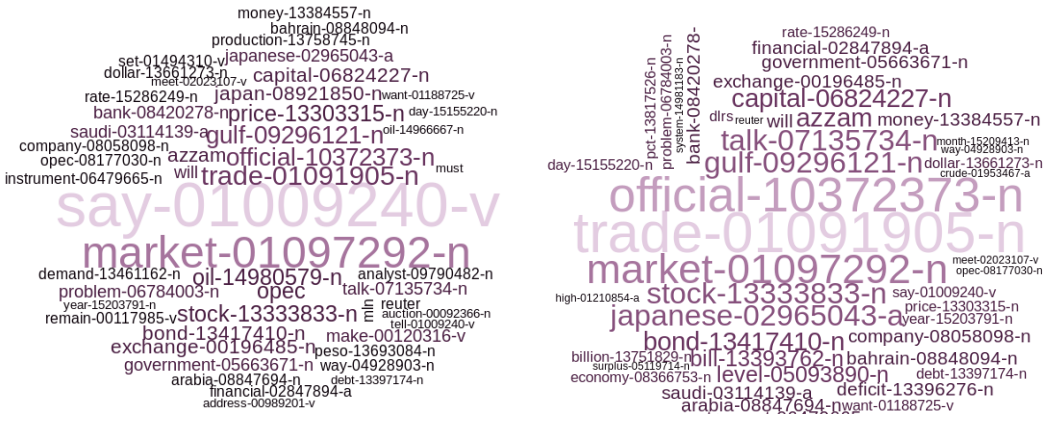


Figure 1. Scores of the first two principal components of the LSA model. Some of the synsets were omitted due to overlap.

Now, in order to more clearly see the terms projected over first two principal axes, in Figure 2, we show wordclouds for each axis. In this way, it is easy to see what terms are relevant for each component. In this case, we used the term "contribution" instead of the term "frequency". This actually connects very well with the qualitative analysis below.



(**a**) (**b**)

Figure 2. Wordclouds of the terms by their contribution to the first two principal components of LSA model. (**a**) Wordcloud of the first principal component. (**b**) Wordcloud of the second principal component.

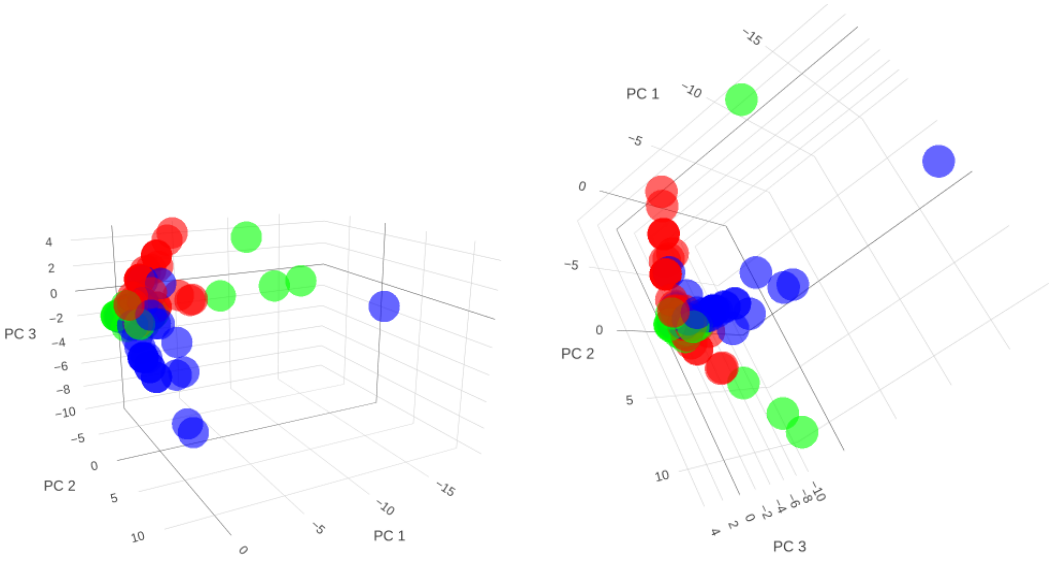
In Table 1, we can see the quantitative comparison between MPPCA and LSA. It is clear that since MPPCA uses mixture components to better capture the shape of the cloud

of points, the weighted average inertia of the first two principal components is higher than for the LSA method by 30 percentual points. Each of the mixture components is capable of capturing the inertia in a very efficient way.

Table 1. Evaluation and comparison of the performance between LSA and MPPCA.

		PC 1		PC 2		**Total** (%)
Model	**EM comp.**	λ_1	λ_1 (%)	λ_2	λ_2 (%)	
MPPCA	1	4.96	42.89	3.57	30.83	73.72
	2	7.28	31.34	6.88	29.59	60.93
	3	7.08	41.89	5.49	32.46	74.35
	W. Avg.	-	**38.70**	-	**30.96**	**69.66**
LSA	-		22.38		16.08	38.46

In Figure 3, it becomes clear that the Normality assumption made by PPCA method does not hold, as the data points have a conic shape around zero, showing the shape of Zipf's law distribution (most of the words with very low frequency and long tail). Nevertheless, it is also worthy noticing that thanks to the introduction of mixture component modelling in the PPCA (i.e., becoming MPPCA), the Normality assumption is leveraged by the flexibility that brings mixture modelling.



(**a**) (**b**)

Figure 3. Representation of mixture components found by MPPCA model over the 3-D LSA factorial space. (**a**) Projected terms in factorial space where each colour represents a mixture component. (**b**) Zoomed-in and rotated near the vertex.

For the LDA model, each topic is described by the terms' distribution. In Figure 4, the terms are sorted in decreasing order by word–topic probability.

Finally, a qualitative comparison has been accomplished by representing the outcomes of the different methods in Table 2. In order to see which methods provide similar outcomes and to realize how accurately those estimations represent the data, we show the most relevant documents and terms to each topic (or the principal component for factorial methods), and also provide summarized texts of those documents (Table 3) with the help of ChatGPT [56].

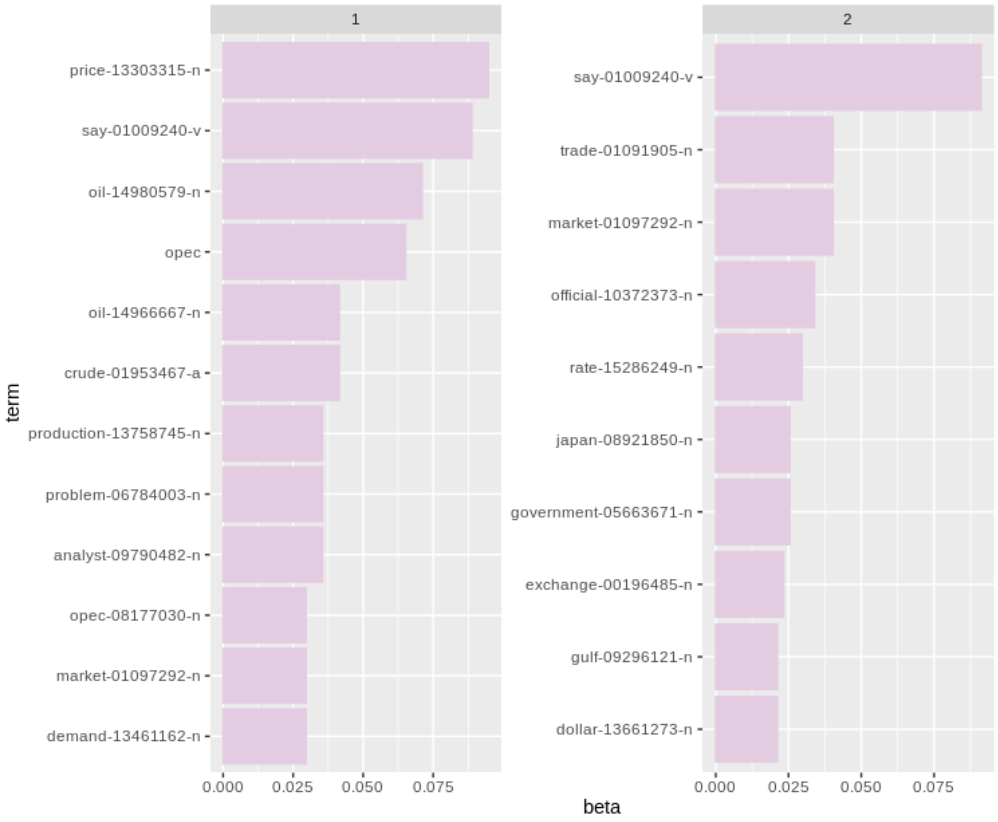


Figure 4. Terms' distribution for each of the first two topics in LDA model.

Table 2. Qualitative comparison between MPPCA, LSA, LDA, and AA models.

		Topic 1		Topic 2	
Model	**EM comp.**	**doc**	**Top-4 Terms**	**doc**	**Top-4 Terms**
MPPCA	1	9	gulf, stock, government, exchange	9	gulf, stock, talk, japanese
	2	9	market, trade, official, rate	4	trade, official, market, japan
	3	9	say, oil, price, opec	3	will, company, problem, last
LSA	-	9	say, market, trade, official	4	trade, official, market, japan
LDA	-	6	price, say, oil, opec	9	say, trade, market, official
AA	-	9	market, say, gulf, stock	8	say, will, price, crude

Table 3. Documents summarized with help of ChatGPT [56].

Document	**Summary**
3	Sweden's ruling Social Democratic Party has given the government the power to impose unilateral trade sanctions on South Africa, prioritizing the fight against apartheid over traditional UN-backed sanctions. The details of the trade boycott will be decided later by the government.
4	Japan and the U.S. are entering trade talks amidst mutual frustration. The U.S. wants Japan to reduce its trade surplus, while Japan faces domestic pressure to boost its economy. The discussions will address economic issues, including access to Japan's supercomputer market. Japan is working to address U.S. concerns despite objections to parts of the trade bill.
6	OPEC may need to hold an emergency meeting before June to address falling oil prices caused by excess supply. Initial optimism about their production control has waned, with doubts about the effectiveness of any emergency meeting. Demand is expected to rise in the next two months, but some believe OPEC may have already exceeded its agreed-upon production quota due to increased demand in the first quarter. The situation remains challenging for OPEC as it tries to stabilize prices and production.

Table 3. *Cont.*

Document	Summary
8	Texaco Canada has decreased the price it pays for crude oil to Canadian CTS per barrel, reducing the posted price for the benchmark grade to Canadian dollars per barrel. This change is effective immediately, following the last adjustment made by the company in February.
9	Gulf money markets have grown in the past decade, but bond and stock markets in the region are underdeveloped and fragmented, according to Gulf International Bank (GIB). Challenges include a recession from falling oil prices, family firms avoiding going public, and limited financial awareness among investors. GIB calls for better financial sophistication, diversified capital market instruments, and improved disclosure requirements for company accounts. Progress is slow in establishing formal stock exchanges in Qatar, Oman, and the UAE, despite some improvements in Bahrain and Saudi Arabia.

For instance, the two purely factorial methods and MPPCA coincide in assigning document 9 as the representative of the first topic (principal component), which talks about stock markets, oil prices, recession, and the Gulf Bank. This is also reflected in the Top-4 most contributing terms.

A similar situation happens for the second topic (or principal component), except for the Archetypal Analysis method, which picks document 8 as the most representative, which talks about an oil price decrease by Texaco Canada, whereas document 4 (the common one between MPPCA mixture component 2 and LSA) talks about talks between Japan and the US, and Japan's difficulties on economic matters. Also, the Top-4 most relevant terms somehow reflect those scenarios.

The LDA method is actually pretty much aligned with factorial methods and MPPCA, as the probability of topic 2 is 0.67. Therefore, if we sort the topics by decreasing probability, we would see similar outcome to the rest of methods. The only difference would be document 6 as representative for the less probable topic number 1, which talks about OPEC intervention in the decrease of oil prices.

From this, it is important to realize that in order to compare factorial methods to LDA, one needs to be careful of the order of the topics of the latter method, as in factorial techniques, the principal components are sorted by their inertia by construction.

6. Discussion

Topic modelling is one of the hot research topics nowadays, and there are a number of approaches in the literature based on different principles. In this paper, 12 different approaches are presented. An important effort has been made to refer to all of them in a common notation so that it became suitable to make comparisons and to understand what the commonalities and particularities of the different techniques are.

Multivariate techniques apply to scenarios where no distributional assumptions are made (see the Distribution hypotheses column of Table 4), and they are based on factorial methods applied to TDM; thus, they use projection over the factorial space as a way to identify the topics. These methods are the first three in Table 5, and they are based on finding the most conservative (in terms of projected information) projection of the original TDM (the numerization of the corpus), except in the case if AA, which uses TSM (which means that it applies the same kind of approach on a higher granularity representation of the text, where numerization is computed inside sentences instead of documents). Thus, they provide a clear geometric and algebraic interpretation of the results and inferred topics, which can be characterized by the set of terms as well as associated documents. Multivariate techniques provide the projected coordinates of both terms and documents (sentences, n-grams, etc.), so global factorial maps can show the relationships among all of them. In standard implementations, these methods use TDM (or TSM) as input data.

In Table 5, the functions optimized by each method are listed, and it is very evident that those techniques belong to two different families: multivariate techniques on the one side, and probabilistic models on the other.

Table 4. Comparison of all topic models presented in this work.

	Functional Features						Distribution Hypotheses	Assumptions				
Model Name	**Type of Func.**	**Word Context**	**Exact Solut.**	**Input**	**Output**	**Doc. Size Par.**	**Distrib.**	**Word. Iid**	**Doc Iid**	**Pos. Iid**	**Pre-Fixed K**	**Topics/Doc**
Latent Semantic Analysis	Max var.	no	yes	TDM	Λ, U, V	no	free	no	no	yes	no	$1 \dots \mathcal{K}$
Distributional Semantic Model	Max var.	yes	yes	TDM, C	Λ, U, V	no	free	no	no	yes	no	$1 \dots \mathcal{K}$
Archetypal Analysis	Min SSE	no	no	TSM, A	H, W	no	free	no	no	yes	yes	$1 \dots K$
Probabilistic PCA	Max $\mathcal{L}$	no	no	TDM	$W, \boldsymbol{\mu}, \sigma$	no	Gaussian	no	yes	yes	no	$1 \dots \mathcal{K}$
Probabilistic Topic Modelling	Max $\mathcal{L}$	no	no	$\mathcal{D}$	$\boldsymbol{\theta}$	no	Parameterized	no	yes	yes	yes	1
Bernoulli Model	Max $\mathcal{L}$	no	no	$\mathcal{D}$	$\boldsymbol{\theta}$	no	Bernoulli	no	yes	yes	yes	1
Probabilistic Topic Modelling (doc. size)	Max $\mathcal{L}$	no	no	$\mathcal{D}$	$\boldsymbol{\theta}$	yes	Parameterized	no	yes	yes	yes	1
Generative Model	Max $\mathcal{L}$	no	no	$\mathcal{D}$	$\boldsymbol{\theta}$	yes	Categorical	no	yes	yes	yes	1
Multinomial Model	Max $\mathcal{L}$	no	no	$\mathcal{D}$	$\boldsymbol{\theta}$	yes	Multinomial	no	yes	yes	yes	1
Latent Dirichlet Allocation	Max $\mathcal{L}$	no	no	$\mathcal{D}$	$\boldsymbol{\theta}$	no	Categorical Dirichlet	no	yes	yes	yes	$1 \dots K$

Table 5. Formulations of the presented methods.

Model Name	Function to Optimize
Latent Semantic Analysis	$\max_{\boldsymbol{u}_\alpha} \boldsymbol{u}'_\alpha X'X\boldsymbol{u}_\alpha$ s.t. $\boldsymbol{u}'_\alpha\boldsymbol{u}_\alpha = 1$
Distributional Semantic Model	$\max_{\boldsymbol{u}_\alpha} \boldsymbol{u}'_\alpha (X'C)'(X'C)\boldsymbol{u}_\alpha$ s.t. $\boldsymbol{u}'_\alpha\boldsymbol{u}_\alpha = 1$
Archetypal Analysis	$\min_{H,W} \|J - H_{(n_\mathcal{T}\times K)}W'J\|^2$ with convexity constraints
Probabilistic PCA	$\mathcal{L}(\boldsymbol{\mu},\sigma^2,W) = \sum_{i=1}^{n_\mathcal{T}} \log\{p(\boldsymbol{x}^i)\} = -\frac{n_\mathcal{T}}{2}\{n_\mathcal{D}\log(2\pi) + \log\lvert C\rvert + \mathrm{tr}(C^{-1}S)\}$
Probabilistic Topic Modelling	$\mathcal{L}(\theta) = \prod_{j=1}^{n_\mathcal{D}} \sum_{k=1}^{K} P(\bigwedge_{\not=1:n_{d_j}} D_\not = w_\not \lvert Z = z_k \wedge \theta) P(Z = z_k\lvert\theta)$
Probabilistic Topic Modelling (considering document size)	$\mathcal{L}(\theta) = \prod_{j=1}^{n_\mathcal{D}} \sum_{k=1}^{K} \left(P(N = n_{d_j}\lvert Z = z_k \wedge \theta) P(\bigwedge_{\not=1:n_{d_j}} D_\not = w_\not \lvert Z = z_k \wedge \theta) P(Z = z_k\lvert\theta)\right)$
Probabilistic Topic Modelling: Generative Model	$\mathcal{L}(\theta) = \prod_{j=1}^{n_\mathcal{D}} \sum_{k=1}^{K} \left(P(N = n_{d_j}\lvert Z = z_k \wedge \theta) \cdot \left(\prod_{i=1}^{n_\mathcal{T}} \pi_{ik}^{n_{ij}}\right) \cdot P(Z = z_k\lvert\theta)\right)$
Probabilistic Topic Modelling: Multinomial Model	$\mathcal{L}(\theta) = \prod_{j=1}^{n_\mathcal{D}} \sum_{k=1}^{K} \left[P(N = n_{d_j}\lvert Z = z_k \wedge \theta) \cdot \frac{n_{d_j}!}{\prod_{i=1}^{n_\mathcal{T}} n_{ij}!} \prod_{i=1}^{n_\mathcal{T}} \pi_{ik}^{n_{ij}} \cdot P(Z = z_k\lvert\theta)\right]$
Probabilistic Topic Modelling: Bernoulli Model	$\mathcal{L}(\theta) = \prod_{j=1}^{n_\mathcal{D}} \sum_{k=1}^{K} \left[\left(\prod_{i=1}^{n_\mathcal{T}} \left(x_j^i \pi_{ik} + (1 - x_j^i)(1 - \pi_{ik})\right)\right) P(Z = z_k\lvert\theta)\right]$
Probabilistic Topic Modelling: Latent Dirichlet Allocation	$\mathcal{L}(\alpha,\theta) = \prod_{j=1}^{n_\mathcal{D}} \int P(\zeta_j\lvert\alpha) \left(\prod_{\not=1}^{n_{d_j}} \sum_{k=1}^{K} P(D_\not = w_\not \lvert Z_\not = z_k \wedge \theta) P(Z_\not = z_k\lvert\zeta_j)\right) d\zeta_j$

Factorial methods aim to optimize a function relative to the amount of information of the original data that is retained in the latent space; for instance, AA minimizes the residual sum of squares, whereas PCA maximizes the projected variance. In general, the solution (i.e., the optimal projection directions) is derived through diagonalization techniques applied to the TDM which, in turn, will depend on the method. The results are expressed in terms of rotation directions (which can be eigen vectors, depending on the method), the quantity of information retained in the factorial component (sometimes eigen values), and the coordinates of the original terms of documents in the projected space can be computed as linear combinations of the columns of TDM weighted by the obtained eigenvectors components. The Distributional Semantic Model takes the context of the words into account. It calculates co-occurrences of the terms inside a pre-defined window around some predetermined words. Explicit Semantic Analysis strongly relies on a pre-defined set of concepts.

On the other hand, the methods that come from the probabilistic modelling theories strongly rely on distributional assumptions. Nevertheless, they provide a solid basis to capture uncertainty due to their probabilistic nature.

Unlike multivariate methods, they use the corpus data with the basic unit being the term as the standard input, except for PPCA, which uses TDM data as input data. However, the frequencies of the terms in sentences or documents are often used inside the algorithms as well.

In order to reduce and simplify the estimation of the set of distributional parameters, probabilistic methods make different kinds of probabilistic assumptions. The words or terms depend on the topic, but not on the other words or terms that appear in the same document. Additionally, when conditioning on the topic, the words or terms are identically distributed across sentences, documents, and the whole corpus. The documents are assumed to be independent of each other and also identically distributed. In fact, all seven methods analysed in this paper make the strong assumption of independence, meaning that the occurrence of a term is independent of the document in the corpus, the position in the document, and the context. This is, in fact, a quite unrealistic assumption, although it is more than frequency in probabilistic and Bayesian statistics in general, as unless this can be assumed, the computation of the joint probability distribution of a document or that of a corpus becomes unfeasible. Assuming independence allows joint probability distributions as direct products of their components, which makes the model suitable from the theoretical point of view.

All probabilistic methods presented in this paper provide a probabilistic model for the likelihood of the model parameters, given the observed corpus, as usual in Bayesian approaches. And, all of them assume a predefined number of topics. Often, setting this parameter beforehand it is not easy.

Depending of the method, a more or less complex likelihood function, or its logarithm (see column 2 of Table 4) is maximized to estimate the model parameters. This allows us to get values for the probabilistic model assumed for the corpus (which can change from one method to another), so that the probability of a certain topic given a document can be computed, among others. In most methods, except LDA, it is assumed that a document belongs to a single topic (the most probable one).

So, the output of probabilistic models is that the entire probabilistic distribution of the corpus, given different topics among others, has a very different nature than the kind of outputs provided in factorial methods.

Additional simplifications and assumptions appear in some specific methods, leading to a different formulations and, therefore, different models arise, as shown in Table 5. For example, Probabilistic Topic Modelling under the Bernoulli approach can be considered the most simple method in this family. It only matters whether a specific term occurs or not in a document, while all six other approaches work with the number of occurrences of the term in a document.

All other probabilistic methods introduce the size of the document in the model. The Categorical (generative) model considers the same Categorical distribution of all words in each position of the document, with an additional assumption of equiprobability of terms along all positions in document. The Multinomial model introduces the number of occurrences of terms in a document, and proposes a model based on the multinomial distribution of these occurrences. It makes the same independence assumptions as the Categorial model. Both Categorical and Multinomial arrive to the same parameter estimation, but starting from different distributional assumptions.

LDA seems to be the most complex probabilistic method, as it assumes that the document can belong to several topics simultaneously. Although LDA is formulated considering the document length, in the original paper, the authors omit it for the sake of simplicity [13].

Then, PPCA is in the mid way between probabilistic and multivariate methods. It provides a probabilistic version of LSA. While PPCA formulates the likelihood to be estimated as a function related to the quantity of information kept in the projected space, it can be formulated in terms of a Likelihood function. It requires a distributional assumption for the counts of the words in the corpus, and it is formulated assuming Gaussian distribution which, in fact, seems not to fit properly for textual data (see Table 4). Nevertheless, the MPPCA variant seems to leverage this assumption by introducing mixture ,components leading to promising results and elegant interpretation.

Finally, from the experimental section and the analyses performed so far, it can be seen that all methods show interesting and similar behaviour amongst them, at least in this particular experimental setup. The similar outcome of all the compared methods could be due to the highly elaborate pre-processing step, which perhaps helped all the methods to find those signals or patterns and efficiently disregard the noise.

On the other hand, thanks to the common notation framework, it was easier to perform the analysis, especially for the qualitative comparison, as the calculated contributions and correlations spoke the same language, having the same or similar notation for concepts that are closely related.

In particular, this bridge was more notorious between factorial methods and MPPCA, allowing us to explore the mid point between factorial and probabilistic approaches.

7. Conclusions

This work represents a pioneering effort in unifying the notation for two distinct families of topic modelling methods: multivariate and probabilistic topic modelling. Despite their conceptual differences, the authors demonstrate that employing a shared notation enables a detailed analysis of commonalities and distinctions between these approaches.

In the multivariate setting, associations between topics, documents, and terms are interpreted and visually represented with clarity, facilitating a comprehensive overview of their interactions.

Conversely, probabilistic methods lack a geometric representation, but offer greater flexibility in capturing associations among topics, documents, and terms.

Another fundamental difference lies in the assumption made by probabilistic approaches, where the number of topics is predetermined from the outset. On the other hand, multivariate methods allow the determination of relevant topics as an output, achieved by assessing the quantity of information preserved in each topic and retaining the significant ones.

Also, all probabilistic approaches, except LDA, assume a single topic per document, which is a more limited approach, whereas the multivariate approaches show the relationship between each particular document and all the significant topics; so, it is possible to identify the more realistic situation of documents involved with several topics simultaneously. This is also the case of LDA, which provides the probabilities of a document to belong to all the predetermined set of topics. The fact that two consecutive words may belong to two completely different topics in LDA does not seem a very realistic assumption.

The PPCA appears to be an appealing approach, combining both probabilistic and multivariate methods. However, the Normality assumption does not align with the actual distribution of terms within documents.

On the other hand, multivariate models, despite being simple linear models, present a more conservative modelling approach, as they do not impose any distributional assumptions, making the interpretation of results straightforward.

Nevertheless, this analysis highlights a common characteristic among all the proposed methods: they offer means to characterize topics based on documents or words that are representative of the topics. However, the true essence of these topics still relies on the interpretational abilities of the analyst or domain expert. This indicates a missing final step in the field of topic modelling research, which is to provide a concept or label for each of the discovered topics. During the experimental stage, this issue was confirmed becoming very challenging to describe the topics discovered by different methods. Therefore, this task becomes arbitrary and has high degree of manual effort.

For this reason, currently, the research is focused on enriching the methodology with inductive reasoning and ontologies of terms in order to obtain final concepts or labels for the discovered topics in an automatic way, providing a clear interpretation for the last stage of the topic modelling task for any type and family of topic modelling technique.

Author Contributions: Conceptualization, K.G. and Y.H.-P.; Methodology, K.G. and Y.H.-P.; Software, Y.H.-P.; Validation, Y.H.-P.; Formal analysis, K.G. and Y.H.-P.; Investigation, K.G. and Y.H.-P.; Resources, Y.H.-P.; Data curation, Y.H.-P.; Writing—original draft, Y.H.-P.; Writing—review & editing, K.G.; Visualization, Y.H.-P.; Supervision, K.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All the data used in the experimental section can be found in the data collection [55].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ahmadi, P.; Gholampour, I.; Tabandeh, M. Cluster-based sparse topical coding for topic mining and document clustering. *Adv. Data Anal. Classif.* **2018**, *12*, 537–558. [CrossRef]
2. Gaul, W.; Vincent, D. Evaluation of the evolution of relationships between topics over time. *Adv. Data Anal. Classif.* **2017**, *11*, 159–178. [CrossRef]
3. Tadesse, D.G.; Carpenter, M. A method for selecting the relevant dimensions for high-dimensional classification in singular vector spaces. *Adv. Data Anal. Classif.* **2018**, *13*, 405–426. [CrossRef]
4. Iovleff, S. Probabilistic auto-associative models and semi-linear PCA. *Adv. Data An. Classif.* **2015**, *9*, 267–286. [CrossRef]

5. Canhasi, E.; Kononenko, I. Multi-document summarization via Archetypal Analysis of the content-graph joint model. *Knowl. Inf. Syst.* **2014**, *41*, 821–842. [CrossRef]
6. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [CrossRef]
7. Jolliffe, I. *Principal Component Analysis*; Springer Series in Statistics; Springer: Berlin/Heidelberg, Germany, 2002.
8. Deerwester, S.; Dumais, S.T.; Furnas, G.W.; Landauer, T.K.; Harshman, R. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 391–407. [CrossRef]
9. Lee, D.D.; Seung, H.S. Algorithms for Non-negative Matrix Factorization. In Proceedings of the 13th International Conference on Neural Information Processing Systems, NIPS'00, Cambridge, MA, USA, 27 November–2 December 2000; pp. 535–541.
10. Greenacre, M.; Nenadic, O. *Computation of Multiple Correspondence Analysis, with Code in R*; Economics Working Papers; Department of Economics and Business, Universitat Pompeu Fabra: Barcelona, Spain, 2005.
11. Ogunleye, B.; Maswera, T.; Hirsch, L.; Gaudoin, J.; Brunsdon, T. Comparison of Topic Modelling Approaches in the Banking Context. *Appl. Sci.* **2023**, *13*, 797. [CrossRef]
12. Font, M.; Puig, X.; Ginebra, J. Bayesian Analysis of the Heterogeneity of Literary Style. *Rev. Colomb. EstadíStica* **2016**, *39*, 205–227.
13. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
14. Hoffman, M.; Bach, F.R.; Blei, D.M. Online learning for latent dirichlet allocation. In *Proceedings of the Advances in Neural Information Processing Systems;* MIT Press: Cambridge, MA, USA, 2010; pp. 856–864.
15. Wang, D.; Zhu, S.; Li, T.; Gong, Y. Multi-document Summarization Using Sentence-based Topic Models. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, ACLShort '09, Stroudsburg, PA, USA, 4 August 2009; pp. 297–300.
16. Arora, R.; Ravindran, B. Latent Dirichlet Allocation and Singular Value Decomposition Based Multi-document Summarization. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; pp. 713–718. [CrossRef]
17. Zhang, Y.; Xu, H. SLTM: A Sentence Level Topic Model for Analysis of Online Reviews. In Proceedings of the SEKE, San Francisco, CA, USA, 1–3 July 2016; pp. 449–453.
18. Hofmann, T. Probabilistic Latent Semantic Indexing. In Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '99, New York, NY, USA, 15–19 August 1999; pp. 50–57. [CrossRef]
19. Nigam, K.; Mccallum, A.K.; Thrun, S.; Mitchell, T. Text Classification from Labeled and Unlabeled Documents using EM. *Mach. Learn.* **2000**, *39*, 103–134. :1007692713085. [CrossRef]
20. Roozbeh, M.; Maanavi, M.; Babaie-Kafaki, S. Robust high-dimensional semiparametric regression using optimized differencing method applied to the vitamin B2 production data. *Iran. J. Health Sci.* **2020**. [CrossRef]
21. Roozbeh, M.; Babaie–Kafaki, S.; Aminifard, Z. Two penalized mixed–integer nonlinear programming approaches to tackle multicollinearity and outliers effects in linear regression models. *J. Ind. Manag. Optim.* **2021**, *17*, 3475–3491. [CrossRef]
22. Indurkhya, N.; Damerau, F.J. *Handbook of Natural Language Processing*; CRC Press: Boca Raton, FL, USA, 2010; Volume 2.
23. Marcus, M.P.; Marcinkiewicz, M.A.; Santorini, B. Building a Large Annotated Corpus of English: The Penn Treebank. *Comput. Linguist.* **1993**, *19*, 313–330.
24. Francis, W.N.; Kucera, H. *Brown Corpus Manual;* Technical Report; Department of Linguistics, Brown University: Providence, RI, USA, 1979.
25. Fellbaum, C. *WordNet: An Electronic Lexical Database*; Bradford Books: St Bradford, PA, USA, 1998.
26. Johansson, S. *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers*; ICAME Collection of English Language Corpora; University, Department of English: Salt Lake City, UT, USA, 1978.
27. Manning, C.D.; Manning, C.D.; Schütze, H. *Foundations of Statistical Natural Language Processing*; MIT Press: Cambridge, MA, USA, 1999.
28. Baeza-Yates, R.; Ribeiro-Neto, B. *Modern Information Retrieval*; ACM Press: New York, NY, USA, 1999; Volume 463.
29. Salton, G.; McGill, M. *Introduction to Modern Information Retrieval*; McGraw-Hill Computer Science Series; McGraw-Hill: New York, NY, USA, 1983.
30. Benzécri, J.P.; Birou, A.; Blumenthal, S. *L'analyse des Donnees, Tome II. L'analyse des Correspondances (The Analysis of Data, Volume II. The Analysis of Correspondence)*; Dunod Press: Paris, France, 1973.
31. Devroye, L.; Lebart, L.; Morineau, A.; Fenelon, J.P. Tratement des Donnees Statistiques: Methodes et Programmes. *J. Am. Stat. Assoc.* **1980**, *75*, 1040. [CrossRef]
32. Greenacre, M. *Biplots in Practice*; Fundación BBVA: Bilbao, Spain, 2010.
33. Schütze, H.; Manning, C.D.; Raghavan, P. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, MA, USA, 2008; Volume 39.
34. Greenacre, M.J. Correspondence analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 613–619. [CrossRef]
35. Pino, J.; Eskenazi, M. An application of latent semantic analysis to word sense discrimination for words with related and unrelated meanings. In Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications, Boulder, CO, USA, 5 June 2009; pp. 43–46.
36. Landauer, T.K.; Foltz, P.W.; Laham, D. An introduction to latent semantic analysis. *Discourse Process.* **1998**, *25*, 259–284. [CrossRef]
37. Rajman, M.; Besançon, R. Stochastic Distributional Models for Textual Information Retrieval. In Procedings of the 9th International Symposium on Applied Stochastic Models and Data Analysis (ASMDA-99), Lisbon, Portugal, 14–17 June 1999; pp. 80–85.
38. Cutler, A.; Breiman, L. Archetypal Analysis. *Technometrics* **1994**, *36*, 338–347. [CrossRef]

39. Sharaff, A.; Nagwani, N.K. Email thread identification using latent Dirichlet allocation and non-negative matrix factorization based clustering techniques. *J. Inf. Sci.* **2016**, *42*, 200–212. [CrossRef]
40. Gabrilovich, E.; Markovitch, S. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In Proceedings of the International Joint Conference on Artificial Intelligence, Hyderabad, India, 6–12 January 2007; Volume 7, pp. 1606–1611.
41. Tipping, M.E.; Bishop, C. Probabilistic Principal Component Analysis. *J. R. Stat. Soc. Ser. B* **1999**, *61*, 611–622. [CrossRef]
42. Tipping, M.E.; Bishop, C.M. Mixtures of probabilistic principal component analyzers. *Neural Comput.* **1999**, *11*, 443–482. [CrossRef] [PubMed]
43. Bishop, C.M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*; Springer: Berlin/Heidelberg, Germany, 2006.
44. Jurafsky, D.; Martin, J.H. *Speech and Language Processing*, 2nd ed.; Prentice-Hall, Inc.: Upper Saddle River, NJ, USA, 2009.
45. Peña, D. *Análisis de Datos Multivariantes*; Mc Graw Hill: New York, NY, USA, 2002.
46. Carreira-Perpinan, M.A. Mode-finding for mixtures of Gaussian distributions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1318–1323. [CrossRef]
47. Mccallum, A.; Nigam, K. A Comparison of Event Models for Naive Bayes Text Classification. *Work. Learn. Text Categ.* **2001**, *752*, 41–48.
48. Heath, D.; Sudderth, W. De Finetti's Theorem on Exchangeable Variables. *Am. Stat.* **1976**, *30*, 188–189. [CrossRef]
49. Papadia, G.; Pacella, M.; Perrone, M.; Giliberti, V. A Comparison of Different Topic Modeling Methods through a Real Case Study of Italian Customer Care. *Algorithms* **2023**, *16*, 94. [CrossRef]
50. Hwang, S.J.; Lee, Y.K.; Kim, J.D.; Park, C.Y.; Kim, Y.S. Topic Modeling for Analyzing Topic Manipulation Skills. *Information* **2021**, *12*, 359. [CrossRef]
51. Chen, W.; Rabhi, F.; Liao, W.; Al-Qudah, I. Leveraging State-of-the-Art Topic Modeling for News Impact Analysis on Financial Markets: A Comparative Study. *Electronics* **2023**, *12*, 2605. [CrossRef]
52. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]
53. Rousseeuw, P.J. Least Median of Squares Regression. *J. Am. Stat. Assoc.* **1984**, *79*, 871–880. [CrossRef]
54. Padró, L.; Stanilovsky, E. FreeLing 3.0: Towards Wider Multilinguality. In Proceedings of the Language Resources and Evaluation Conference (LREC 2012), ELRA, Istanbul, Turkey, 21–27 May 2012.
55. Lewis, D.D. Reuters-21578. 1987. Available online: https://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html (accessed on 18 October 2023).
56. OpenAI. ChatGPT (September 25 Version) [Large Language Model]. 2023. Available online: https://chat.openai.com/ (accessed on 18 October 2023).