sciendo Journal of Social and Economic Statistics

Vol. 14, No. 1, Summer 2025
***DOI**: 10.2478/jses-2025-0003*

Abstract versus Full Paper: A quantitative approach

Dimitri CALIPOLITIS [1*], Carla PLEȘTIU[2]

[1, 2] The Bucharest University of Economic Studies, Romania

Abstract

The purpose of this article is to observe the differences between the words from selected papers and the terms from summaries or introductory chapters, that correspond with the selected papers, using quantitative methods. To test the above hypothesis, we took the terms from abstracts and introductory chapters, of three different type of papers (a scientific article, a literature book and a methodological study), and we compared them with the most frequent words from the selected papers. Furthermore, we also applied different methods from text mining, such as calculating the phi coefficient or checking compliance with a specific empirical law (Zipf, Heaps), to spot the existent dissimilarities. Following this analysis, it can be observed, from a quantitative point of view, that certain frequent words from the complete writings are found in their introductory components, present a fairly high frequency and outline the main ideas from the respective texts. This paper was co-financed by The Bucharest University of Economic Studies during the PhD program.

Keywords: text mining, pairwise correlation, quantitative methods, words frequency

JEL Classification: C10, C12, C15

1. Introduction

Text mining is a continuously developing subfield, separated from data mining, which aims to extract information and process specific models, using any written text as the main database. The essential goal of text mining is to structure a given input text, determine the patterns that emerge at first glance from the structured data following the processing of the document and to provide an interpretation or a statistical evaluation based on the results obtained. Given the recent technological development, a complex text mining analysis can also be achieved by applying specific machine learning (ML) or natural language processing (NLP) algorithms.

The main purpose of the present article is to observe, from a quantitative point of view, whether the terms used in the introductory components of the papers differ greatly from the most frequently words of the initial texts and whether this quantitative difference has an impact on the main ideas within the analyzed writings. By introductory component, we refer to the text paragraphs, which compress the main ideas existing in the papers or books, such as introductory chapters, summaries or abstracts. In order to encompass this diversification of writings, we considered that it is useful to analyze a literature book together with its summary, a methodological work with its introductory chapter and a scientific article together with its

* Corresponding author, Dimitri Calipolitis – calipolitisdimitri20@stud.ase.ro

abstract. The entire analysis was carried out in R, using dedicated packages for applying the text mining methodology.

The present approach is new, since we were unable to find similar research on this subject. In *the Literature Review* section, we discussed articles that use methods similar to those applied by us, but their main theme differs from the one proposed by us. Therefore, in the *Data and Methodology*, we pointed out exactly the theoretical part used, while in *Results and Discussions*, we emphasized in detail the relevant results that were obtained. From our point of view, future statistical models for evaluating writings can be developed based on these results, so that their introductory components will contain as many words as possible, that have a high frequency within their entire texts. The applicability of such a model would help in evaluating scientific articles before publication (when comparing the abstract with the full text), in checking the introduction section of a collective volume (to reflect clearly and concisely the important topics discussed), or in creating a correct summary of a literature book (to capture the main ideas from the text).

2. Literature Review

The literature on the differences between abstracts and full-text articles in biomedical text mining reveals several key structural and content-based distinctions that influence natural language processing (NLP) approaches. Previous studies have noted that biomedical information retrieval, particularly in tasks such as protein-protein interaction extraction, faces challenges when relying solely on abstracts, with more comprehensive data often residing in full-text bodies (Friedman et al., 2001). Semantic content, including gene/protein names, mutations, and diseases, is typically more densely packed in article bodies than in abstracts, further complicating the mining process (Lamurias & Couto, 2019). Additionally, there is a growing recognition of the structural variations between the two genres, such as sentence length and the frequency of parenthesized material, which can impact the performance of syntactic and semantic parsing tools (Rzhetsky et al., 2004). Methods employed in these studies range from manual annotation of biomedical corpora (e.g., CRAFT) to computational analyses of sentence segmentation and parsing, with results often highlighting the need for specialized tools to handle the complexities of full-text data (Kilicoglu et al., 2024). These findings underscore the importance of developing robust NLP systems capable of processing full-text journal articles in order to fully exploit the vast, nuanced information available in biomedical literature.

The article by Fontelo et al. (2013) investigates the accuracy of data in structured abstracts compared to full-text journal articles and their implications for clinical decision-making. The literature highlights the increasing reliance on abstracts, especially in resource-limited settings, where full-text articles may not be accessible. It emphasizes that abstracts should accurately represent the full article, particularly since clinicians often depend on them for evidence-based decision-making. Previous studies have indicated that discrepancies between abstract data and full text are common, with structured abstracts, due to their clear format, generally seen as more reliable (Taddio et al., 1994). However, discrepancies, such as “spin” in clinical trials or numerical inconsistencies, can mislead clinicians (Boutron et al., 2010; Yavchitz et al., 2012). The authors applied a quantitative approach, comparing structured abstracts to full-text articles across multiple medical journals, to assess data accuracy, using statistical methods to determine whether discrepancies were clinically significant or not.

Data and Methodology

In order to analyze the concepts proposed in the introductory part of this article, we selected the following three writings: "*Dracula*" by Bram Stoker, "*Becoming a Data Head - How to think, speak, and understand data science, statistics, and machine learning*" written by Alex J. Gutman and Jordan Goldmeier, and "*Correlated Topic Models*" by David M. Blei and John D. Lafferty. The first writing represents a classic fiction text, the second is a scientific and methodological text for the application of various concepts and methods in data science, and the third mentioned refers to a scientific article, which explains the theoretical basis and the method of applying topic modeling. All listed texts were imported into RStudio from text files and processed, by applying the removal of stop words, punctuation marks, spacing and numbers, converting uppercase to lowercase and lemmantization. We did that by using the functions in the *tm* and *tidyverse* packages, to obtain tables containing the frequency of words and their rank, both for the full texts and for their introductory components. Subsequently, starting from these tables, another table was created with the common terms from the full texts and their introductory components, for each paper, so that it could be used further in the analysis. In addition, filtered tables of the full texts were saved separately, where the frequency of each term is higher than 0.0010 (basically, the respective words have more than one occurrence in the text).

To verify whether there is a positive correlation between the common words, previously extracted in the tables, and those that have a significant positive correlation (greater than 0.5), the texts were reimported and several transformations were applied, so as to calculate the respective coefficients depending on the segmentation of each text. For "*Becoming a Data Head…*", coefficients were calculated both in the case of dividing the text into sections and chapters. Furthermore, the words were taken separately from the two types of tables (those with common terms and those with terms between which there is a high positive correlation) and several counts and joins were applied, so as to extract the number of common and uncommon words per text. In addition, the same operations were performed for the filtered tables, to observe whether there are quantitative differences compared to those containing all the terms in the text. The resulting indicators were saved in a new table (Table no. 1 contains their definition).

Table no. 1: The definition of variables

Variable Name	Definition
Not Common Words from Paper (Full)	The number of words in the text that are not found in the common words dataset
Not Common Words from Paper (Filtered)	The number of words in the text that are not found in the common words dataset and for which the term frequency at the full text level is greater than 0.0010
Common Words from Paper (Full)	The number of words in the text that are found in the common words dataset
Common Words from Paper (Filtered)	The number of words in the text that are found in the common words dataset and for which the term frequency at the full text level is greater than 0.0010
Not Common Words from Summary/Introduction/Abstract (Full)	The number of words in the introductory part of the text (introductory chapter/summary/abstract) that are not found in the common words dataset
Not Common Words from Summary/Introduction/Abstract (Filtered)	The number of words from the introductory part of the text (introductory chapter/summary/abstract), which are not found in the common words dataset and for which the frequency of the terms at the full text level is higher than 0.0010
Common Words from Summary/Introduction/Abstract (Full)	The number of words from the introductory part of the text (introductory chapter/summary/abstract), which are found in the common words dataset
Common Words from Summary/Introduction/Abstract (Filtered)	The number of words from the introductory part of the text (introductory chapter/summary/abstract), which are found in the common words dataset and for which the frequency of the terms at the full text level is higher than 0.0010
Not Paired Correlated Common Words (Full)	The number of words, which present a high positive correlation (higher than 0.5), and which are not found in the common words dataset

Variable Name	Definition
Not Paired Correlated Common Words (Filtered)	The number of words, which show a high positive correlation (greater than 0.5), which are not found in the common words dataset and for which the term frequency at the full text level is greater than 0.0010
Paired Correlated Common Words (Full)	The number of words, which show a high positive correlation (greater than 0.5), and which are found in the common words dataset
Paired Correlated Common Words (Filtered)	The number of words, which show a high positive correlation (greater than 0.5), which are found in the common words dataset and for which the term frequency at the full text level is greater than 0.0010

Source: Authors

Based on the calculated weights at the paper level, a series of plots were created to display the obtained results. Subsequently, starting from these indicators, several transformations were applied, in order to finally calculate the correlation coefficients for a new series of indicators, reported on the following types of records: ***Full*** – all terms from the full words dataset, ***Filtered*** – all terms from the filtered words dataset for which the frequency of the terms is greater than 0.0010, ***All*** – all terms from the combined two datasets of words (by union, without their summation at the indicator level). In fact, the definition of the columns of this new dataframe (Table no. 2) is as follows:

- ***Full_Indicator_Name*** – refers to the names of the indicators; they are the same as those in the previously described datasets, but with references to the texts and their division method; for example: "*Dracula Not Common Words Chapter*" represents the number of uncommon words in "*Dracula*" if the text were divided into chapters.
- ***Paper*** – refers to the number of words belonging to the full version of the text, but without taking into account its introductory component (the introductory chapter for "*Becoming a Data Head…*", respectively the abstract section for "*Correlated Topic Models*")
- ***Summary*** – refers to the number of words in the introductory component of the analyzed text
- ***Pairwise_Correlation*** – refers to the number of words between which there is a high positive correlation coefficient (over 0.5)
- ***Is_Common*** – is dummy variable, which takes the values 0 for uncommon words, 1 for common ones
- ***Record_Type*** – is reference to the type of record for which the data was extracted; takes only two values: *Full* (for the full dataset) and *Filtered* (for the filtered dataset)

Table no. 2: The final dataframe, used to calculate the correlations between the analyzed components

Full Indicator Name	Paper	Summary	Pairwise Correlation	Is Common
Dracula Not Common Words Chapter	6048	31	12	0
Dracula Common Words Chapter	222	222	210	1
Becoming a Data Head Not Common Words Section	3035	158	222	0
Becoming a Data Head Common Words Section	644	644	422	1
Becoming a Data Head Not Common Words Chapter	3035	158	15	0
Becoming a Data Head Common Words Chapter	644	644	629	1
Correlated Topic Models Not Common Words Chapter	468	2	4	0
Correlated Topic Models Common Words Chapter	54	54	50	1

Source: Authors

The calculation of the correlation between pairs of words was performed using the *pairwise_cor()* function from the *widyr* package, which applies the formula for calculating the phi coefficient according to the theory proposed by Yule (1912), and applied on text mining analysis by Silge and Robinson (2017) using R. The phi coefficient is a statistical measure, used in determining the association between two binary variables (in

our case, two distinct words as variables) when a contingency table with specific frequencies has been generated for the respective variables. It is based on the following relation:

$$\phi = \frac{n_{11} * n_{00} - n_{10} * n_{01}}{\sqrt{n_{1.} * n_{0.} * n_{.0} * n_{.1}}} \quad (1)$$

where:

- n_{11} represents the number of sections/documents, where the both two words appear (for example: word A and word B).
- n_{00} is the number of sections/documents, where neither of the two mentioned words appear.
- n_{10} defines the number of sections/documents, where A appears, but not B.
- n_{01} represents the number of sections/documents, where B appears, but not A.
- $n_{1.}$ is the total number of sections/documents, which contains the A word.
- $n_{0.}$ defines the total number of sections/documents, which does not contain the A word.
- $n_{.0}$ is the total number of sections/documents, which does not contain the B word.
- $n_{.1}$ defines the total number of sections/documents, which contains the B word (Silge & Robinson, 2017).

In order to verify Zipf's law, the methodology from Zipf (1949) was adopted, by calculating the frequency of words in the texts (including stop words), and based on this, the rank of each term starting with the highest frequency. This empirical law is based on the following relation:

$$f = \frac{1}{r} \quad (2)$$

where:

- f represents the frequency of a given word
- r is the statistical rank of that given word

In the case of verifying Heaps' law, the theory from (Heaps, 1978) was applied, using the *Heaps_plot* function from the *tm* package after dividing the texts into specific sections. This empirical law is based on the following relation:

$$V = k \cdot \left(N^{\beta}\right) \quad (3)$$

where:

- V represents the number of distinct words from a document
- N defines the number of total words from the same document
- k and β are empirically determined parameters

To test the statistical hypotheses from the resulted regression models, the following functions were used: *white* from the *skedastic* package to verify homoscedasticity, *durbinWatsonTest* from the *car* package to test the autocorrelation of errors, and *jarque.bera.test* from the *tseries* package to verify the normality of errors. To calculate the correlations between the number of words corresponding to each category and by record types, the *cor()* function from the *stats* package was used.

3. Results and Discussions

Following the application of the methodology described above, we observed various aspects, which will be detailed below. At first glance, if we compare the frequent terms from the complete papers with those of

the same rank, but from their introductory sections, we can observe differences in both the weights and the words used, which is a normal behavior. This is reflected in the dataset corresponding to the common terms of each paper, so that some words with a high rank from the words dataset of complete papers are missing in the first described one. For example, for "*Dracula*", the terms belonging to the 2nd ("hand") and 9th ("dear") rank are missing (Figure no. 1), for "*Becoming a Data Head*", there are no words corresponding to the 9th rank ("figure") or to the range between the 11th and 14th rank ("dataset", "feature", "input", "test"), and for "*Correlated Topic Models*", the terms belonging to the 3rd ("log"), 10th ("parameter") and 12th ("exp") rank do not appear. Of course, there are more missing words, but for the top 10 word rankings in each paper, the examples above represent valid observations.

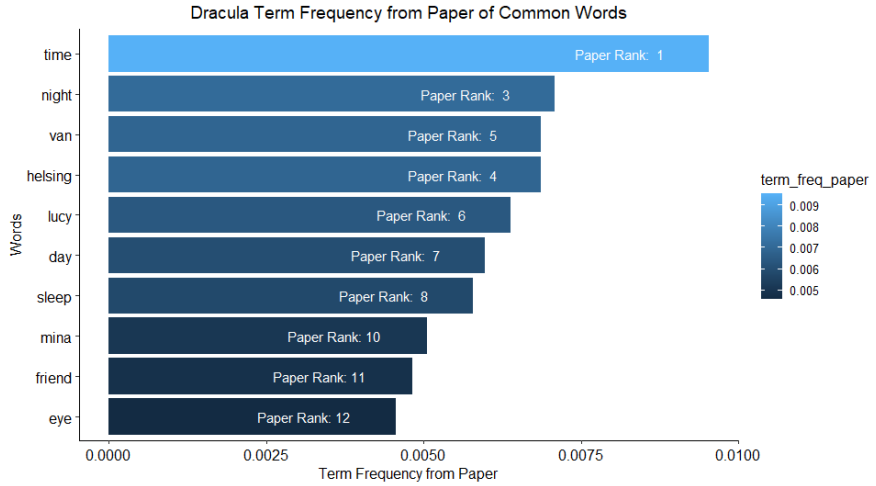

Figure no. 1: "*Dracula*" Term Frequency of Common Words

Source: Authors

Moreover, for all the writings analyzed, Zipf's and Heaps' laws are respected, which shows that the frequency of a word is inversely proportional to its statistical rank and, also there exists a connection between the distinct number of words and their total number in each document (Figure no. 2 as an example on "*Becoming a Data Head…*").

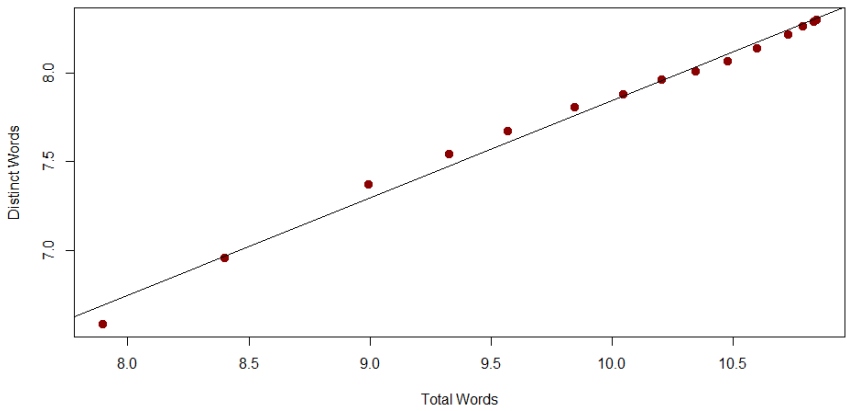

Figure no. 2: Heaps plot on "Becoming a Data Head…" if the book text is divided into chapters

Source: Authors

Analyzing the weight of word types (common/uncommon) for each component of the selected writings (words from full writing, terms from its introductory component and words between it exists a direct correlation), it is noted that there are differences depending on the type of chosen record (full compared to filtered). In other words, regardless of the terms analyzed in each selected paper, it is observed that at the full text level, the weight of uncommon words decreases for filtered records, while at the introductory components or direct correlation between word pairs level, it increases (Figure no. 3 as an example on

"*Correlated Topic Models*"). The value of the weights differs from one writing to another, but for writings of considerable size, the differences are quite significant. However, at the level of correlation between word pairs, these weights do not differ so much depending on the type of record (except for texts divided into a few sections or chapters) and the length of the analyzed text. At the same time, the percent of common words for the mentioned case does not fall below 83%, the exception being only for the words dataset which corresponds to the "*Becoming a Data Head*" if it were divided into sections.

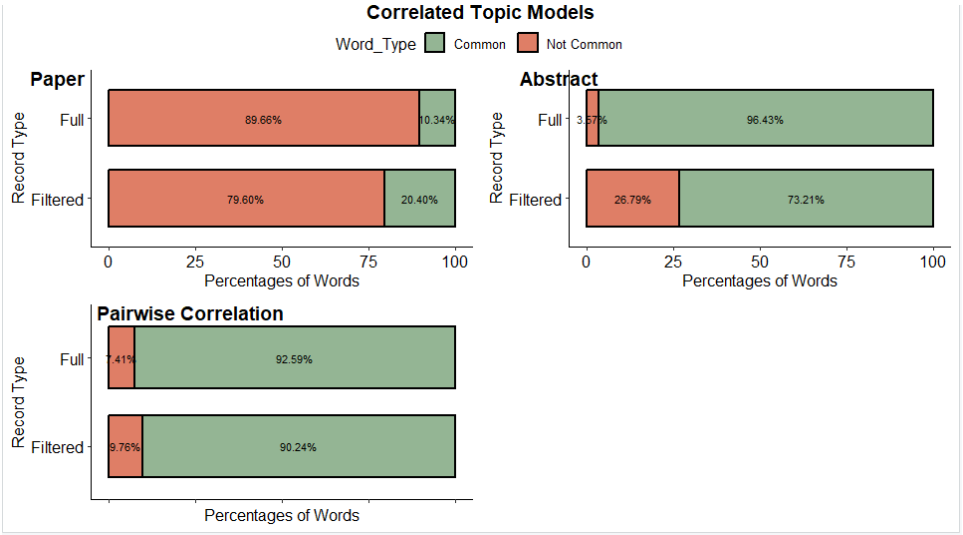

Figure no. 3: Percentages of Words by Word Type in "*Correlated Topic Models*"

Source: Authors

After applying several transformations on the created datasets, we were able to calculate the correlation between the number of words from the full text, the number of terms from the introductory component, the number of words which has a pairwise correlation and the dummy variable *Is_Common*. At the level of full records, a functional deterministic link is observed between *Pairwise_Correlation* and *Summary*, while between *Is_Common* and *Summary*, respectively between the same variable and *Pairwise_Correlation*, a direct link of medium intensity is recorded. Moreover, a direct link of medium intensity is noted between *Paper* and *Is_Common* (Figure no. 4). At the level of filtered records, correlation coefficients decrease significantly, so that only between *Is_Common* and *Summary* is found an indirect link of medium intensity. Compared to the total (both filtered and full records), a decrease in coefficients is also observed, but a direct link of medium intensity is maintained between *Pairwise_Correlation* and *Summary*.

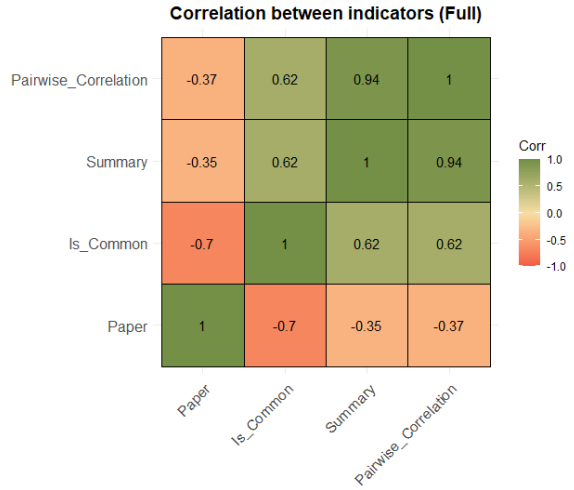

Figure no. 4: Correlation between final indicators on full records

Source: Authors

Starting from the previously extracted correlations, we created a regression model, through which we determined that the *Summary* variable is influenced by *Pairwise_Correlation*, and this was applied to all three previously obtained datasets (on full, filtered and all records). After analyzing the obtained models, we noted that, from a statistical point of view, the one applied to the full records is the best, having *R-Squared*=0.8587, *F-Statistic*=43.55 (*p-value* = 0.000582), and the coefficient of the independent variable being statistically valid (*T-test* = 6.599, *p-value* = 0.000582). However, the intercept is not statistically significant, but it can be removed from the model. Moreover, the classical statistical hypotheses are respected for all created models, but only the one applied to the full records remains the relevant one. In other words, the errors are homoscedastic, follow a normal distribution and are not autocorrelated. From this context, it emerges that the words in the introductory components are influenced by the high correlation found between certain pairs of words, which in fact reflect the main ideas in the analyzed writings.

4. Conclusion

Based on the undertaken analysis, it can be concluded that certain frequent words from the complete writings are found in their introductory components and present a fairly high frequency, compared to the total number of terms existing throughout those mentioned components. At the same time, there are terms from the complete writings that have a high frequency and do not appear in their introductory components, but this behavior is influenced by the human factor. We did not analyze closely how many words of this kind would be reported in total for each paper, because it would mean to establish a certain threshold on term frequency values and this approach needs a more detailed research. However, the calculated values for word frequency on each paper are determined correctly, given that the obtained data follow specific empirical laws (through the presence of Zips and Heaps power laws).

Moreover, if a correct filtering is applied to the words from full writings, which have a significant frequency (not just omitting the words that have only one appearance in the text) and they are compared with those between it exists a close correlation (greater than 0.5), then the extraction of important words, which outline the main ideas from the respective texts, will be achieved in a correct proportion. This statement is reinforced by the fact that the terms, between which there is that close correlation, influence the ideas existing in the respective introductory components, as demonstrated by the presented econometric model and the phi correlation coefficients, calculated for the defined variables in the analysis. Also, it can be observed that adding a dummy variable, which marks the type of word (common/uncommon), to the econometric model does not optimize it too much, especially if the dummy variable has a strong correlation with another similar indicator (as seen in the correlation between *Pairwise_Correlation* and *Is_Common*).

Although this field is in continuous development, the observations presented above can be used in the creation of a machine learning model, which would produce introduction texts or summaries very close to the initial ones, just based on the initial writings. In any case, to validate such a scenario and provide a model with high accuracy, other adjacent analyses must be performed, such as the impact of words between which there is a strong but indirect connection, or how much the percents between common and uncommon words would change if the sample of texts would increase.

References

Boutron, I., Dutton, S., Ravaud, P., & Altman, D. G. (2010). Reporting and Interpretation of Randomized Controlled Trials With Statistically Nonsignificant Results for Primary Outcomes. *JAMA*, *303*(20), 2058–2064. https://doi.org/10.1001/JAMA.2010.651

Fontelo, P., Gavino, A., & Sarmiento, R. F. (2013). Comparing data accuracy between structured abstracts and full-text journal articles: Implications in their use for informing clinical decisions. *Evidence-Based Medicine*, *18*(6), 207–211. https://doi.org/10.1136/EB-2013-101272,

Friedman, C., Kra, P., Yu, H., Krauthammer, M., & Rzhetsky, A. (2001). GENIES: A natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, *17*(SUPPL. 1). https://doi.org/10.1093/BIOINFORMATICS/17.SUPPL_1.S74,

Heaps, H. S. (1978). *Information Retrieval, Computational and Theoretical Aspects*. Academic Pr.

Kilicoglu, H., Ensan, F., McInnes, B., & Wang, L. L. (2024). Semantics-enabled biomedical literature analytics. *Journal of Biomedical Informatics*, *150*, 104588. https://doi.org/10.1016/J.JBI.2024.104588

Lamurias, A., & Couto, F. M. (2019). Text Mining for Bioinformatics Using Biomedical Literature. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, *1–3*, 602–611. https://doi.org/10.1016/B978-0-12-809633-8.20409-3

Rzhetsky, A., Iossifov, I., Koike, T., Krauthammer, M., Kra, P., Morris, M., Yu, H., Duboué, P. A., Weng, W., Wilbur, W. J., Hatzivassiloglou, V., & Friedman, C. (2004). GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatics*, *37*(1), 43–53. https://doi.org/10.1016/J.JBI.2003.10.001

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach* (1st edition). O'Reilly Media, Inc. https://www.oreilly.com/library/view/text-mining-with/9781491981641/

Taddio, A., Pain, T., Fassos, F. F., Boon, H., Ilersich, A. L., & Einarson, T. R. (1994). Quality of nonstructured and structured abstracts of original research articles in the British Medical Journal, the Canadian Medical Association Journal and the Journal of the American Medical Association. *CMAJ: Canadian Medical Association Journal*, *150*(10), 1611. https://pmc.ncbi.nlm.nih.gov/articles/PMC1336964/

Yavchitz, A., Boutron, I., Bafeta, A., Marroun, I., Charles, P., Mantz, J., & Ravaud, P. (2012). Misrepresentation of Randomized Controlled Trials in Press Releases and News Coverage: A Cohort Study. *PLOS Medicine*, *9*(9), e1001308. https://doi.org/10.1371/JOURNAL.PMED.1001308

Yule, G. U. (1912). On the Methods of Measuring Association Between Two Attributes. *Journal of the Royal Statistical Society*, *75*(6). https://doi.org/10.2307/2340126

Zipf, G. K. (1949). *Human Behavior And The Principle Of Least Effort*. Addison-Wesley Press, Inc. https://www.iqla.org/includes/basic_references/Zipf_1949_Human_Behavior_and_the_Principle_of_Least_Effort.pdf