FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

# Similarity Measures for Comparing and Measuring Diversity of News Feeds

**Luís Diogo dos Santos Teixeira da Silva**

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Sérgio Nunes

July 13, 2021

# Similarity Measures for Comparing and Measuring Diversity of News Feeds

**Luís Diogo dos Santos Teixeira da Silva**

Mestrado Integrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

Chair: Prof. José Magalhães Cruz

External Examiner: Prof. Nuno Escudeiro<br>
Supervisor: Prof. Sérgio Nunes

July 13, 2021

# Abstract

The current media culture, associated with the ease of information sharing that is derived from technological progress, has resulted in a substantial data flow. It is very easy for everyone to access news articles and reports on various subjects, written from anywhere in the world. An easy way of accessing this information is using news feeds, which are often associated with social media platforms and present reports from many different sources in an accessible way. This modern-day data paradigm has facilitated content distribution, as well as its consumption. As a result, it has become very hard for a common reader to filter relevant news from trustworthy sources from among the vast available information. This has created a need for tools that are able to facilitate the task of data aggregation and filtering. Particularly in news media, comparison of news feeds can have many different applications, such as news and article clustering, personalized news platforms, and media-related studies.

There are many developed systems in this area, with a variety of purposes. Our first objective was to collect and compare the different strategies used in news article and feed comparison, in order to understand which perform better in different environments. We were able to notice a lack of research in the comparison of news feeds, especially when matched against the abundance of systems that perform comparison of news articles. We were also able to conclude that most of these systems utilize text similarity techniques to compare articles.

Given these findings, we focused our research in strategies that could be used to compare news feeds. We defined four of them - with Feed Concatenation performing very well while maintaining a low temporal and spacial complexity -, and tested them alongside different text similarity measures. We were able to identify multiple strategies that performed well in controlled environments. Also, we came to understand that the strategies that compare the feeds themselves have a greater influence over the final results than the text similarity measures that were used, which proved to be interchangeable.

Finally, our last objective was to test the defined feed similarity strategies in more realistic environments - in our case, the Portuguese media landscape. We performed several experiments with data collected from Portuguese news sources. We were able to detect some shortcomings of defined algorithms, related to uneven datasets, as well as strategies to improve their performance, such as sampling processes. We were also capable of identifying some patterns amid the Portuguese media landscape, by grouping news sources according to their content's characteristics, as well as suggest some practical applications for the proposed algorithms in the field of smart feed construction.

**Keywords**: News feeds, text similarity, information retrieval

# Resumo

A atual cultura dos *media*, associada à facilidade de partilha de informação derivada do progresso tecnológico, resulta num fluxo substancial de dados. Todas as pessoas possuem hoje uma grande facilidade em aceder a artigos noticiosos e relatos dos mais variados assuntos, escritos de qualquer parte do mundo. Uma forma fácil de aceder a esta informação é através da utilização de *feeds* de notícias, que estão muitas vezes associados a plataformas de *social media*, apresentando artigos das mais vairadas fontes de uma forma acessível. Este paradigma moderno de dados facilitou a distribuição de conteúdo, assim como o seu consumo. Como resultado, tornou-se muito difícil para o leitor comum a tarefa de filtrar notícias relevantes, provenientes de fontes fidedignas, de entre a vasta informação disponível. Tudo isto criou uma necessidade de ferramentas que facilitem a tarefa de agregar e filtrar dados. Em particular, nos *media* noticiosos, a compação de *feeds* de notícias pode ter muitas aplicações diversas, que vão desde o aglomerar de artigos noticiosos, à personalização de plataformas de notícias e a estudos relativos aos *media*.

Há muitos sistemas desnvolvidos nesta área, com porpósitos muito distintos. O nosso primeiro objetivo foi recolher e comparar as diferentes estratégias utilizadas na comparação de artigos e *feeds* de notícias, de forma a melhor percebermos quais têm melhores resultados em diferentes cenários. Conseguimos identificar uma falta de pesquisa realizada na área de comparação de *feeds* de notícias, especialmente tendo em perspetiva a abundância de sistemas que realizam comparação de artigos noticiosos. Também pudemos concluir que a maior parte destes sistemas utilizam medidas de similaridade textual para comparar artigos.

Dados estes resultados, focámos a nossa pesquina em estratégias que podessem ser utilizadas para comparar *feeds* de notícias. Definimios quatro - sendo que Feed Concatenation teve bons resultados associados a uma baixa complexidade temporal e espacial -, e testámo-las associadas a diferentes medidas de similaridade textual. Fomos capazed de identificar várias estratégias que tiveram bom resultados em ambientes controlados. Para além disso, pudemos compreender que as estratégias responsáveis pela comparação dos *feeds* em si têm uma maior influência no resultado final do que as medidas de similaridade textual utilizadas, que provaram ser intersubstituíveis.

Finalmente, o nosso último objetivo foi testar as estratégias de similaridade de feeds anteriormente definidas em ambientes mais realistas - no nosso caso, o panorama dos *media* portugueses. Realizámos diversas experiências com dados recolhidos de fontes noticiosas portuguesas. Conseguimos detetar alguns defeitos dos algoritmos definido, relacionados com bases de dados desequilibradas, assim como estratégias para melhorar o seu desempenho, como, por exemplo, processos de amostragem. Fomos também capazes de identificar alguns padrões no panorama dos *media* portugueses, através do agrupamento das fontes de acordo com as características do conteúdo por estas produzido, assim como sugerir algumas aplicações práticas para os algoritmos propostos na área de construção inteligente de feeds.

**Keywords**: *Feeds* de notícias, similaridade textual, recuperação de informação

# Acknowledgements

I would like to thank:

Professor Sérgio Nunes, for his continued support and advice during the writing of this theses;

Professors Elisa Meireles and Joaquim Mendes, that, unbeknownst to them, played a critical role in the choice of my academic path;

My parents and my brother, who have been there for me my whole life and always stood behind my choices;

My grandparents, godparents, and cousin, for their continued love and support;

My friends, for their companionship;

Érica, for everything you've done for me this year and will do in the future.

Luís Diogo Silva

*“Choose a job you love,*
*and you will never have to work a day in your life.”*

Confucius

# Contents

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| AE | Article Elimination |
| AI | Article Injection |
| API | Application Programming Interface |
| AR | Article Repetition |
| AS | Article Shifting |
| CC | Cross Comparison |
| CCE | Cross Comparison with Elimination |
| CC-MRR | Cross Comparison with Mean Reciprocal Rank |
| ESA | Explicit Semantic Analysis |
| FC | Feed Concatenation |
| HAL | Hyperspace Analogue to Language |
| HTML | HyperText Markup Language |
| IC | Information Content |
| LCS | Least Common Subsumer |
| LDA | Latent Dirichlet Allocation |
| LSA | Latent Semantic Analysis |
| NCD | Normalized Compression Distance |
| NGD | Normalized Google Distance |
| NGF | National Global Feed |
| NID | Normalized Information Distance |
| RSS | Rich Site Summary |
| RT | Random Text |
| SW | Sliding Window |
| TF-IDF | Term Frequency-Inverse Document Frequency |
| XML | Extensible Markup Language |

# Chapter 1

# Introduction

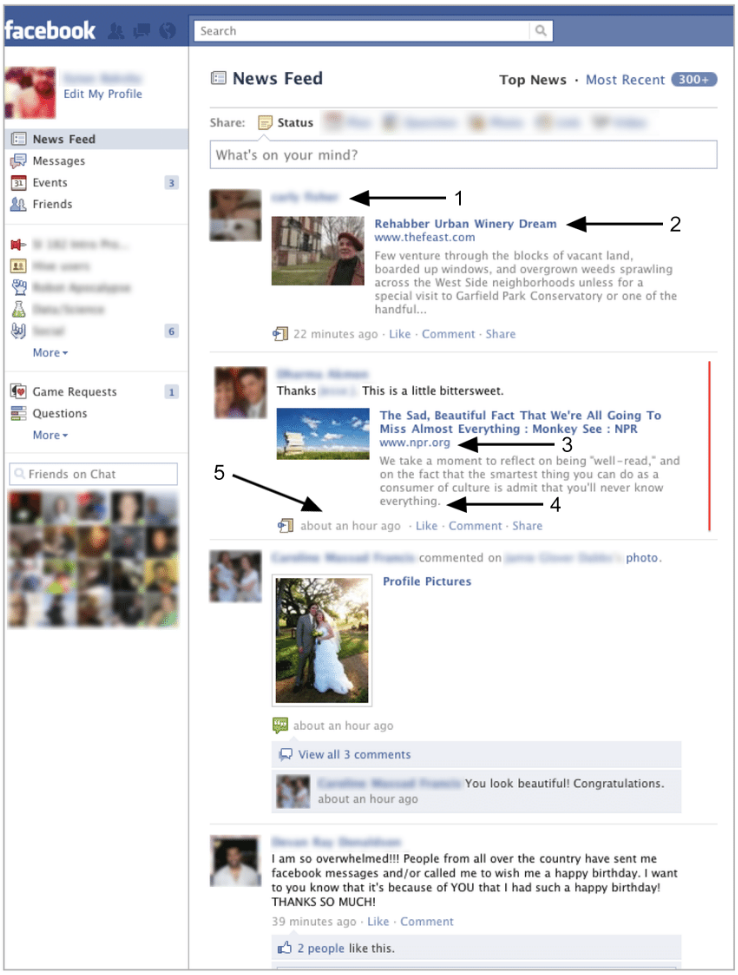
## 1.1 Problem Contextualization

Since very early in the history of humankind, there has been a need to satisfy the natural human curiosity and necessity of knowing the current state of world events, from the mythical story of the soldier that ran from Marathon to Athens to report a Greek victory over the Persians, only to die afterwards, to the roman *Acta Diurna*, meant to detail Senate meetings [37]. From early days, news reports have fulfilled their objective of registering, informing, controlling, and entertaining the masses, seeing their target audience and influence sphere expanded over the ages.

Data production, distribution and gathering have come a long way from these early days. We now live in an information society [88]. The evolution of technology and the trivialization of the use of the Internet has enabled a flux of information never seen before. Since the invention of the printing press, about one hundred billion books have been published across numerous languages and editions [26]. That amount of data corresponds to what was, as of 2012, being published online in a month. The ease of access to information is counterbalanced by the difficulty of selecting, filtering, and interpreting the different sources, resulting in what is called "information overload", caused by the sheer amount of data [56].

This is very noticeable in the news industry. News articles cover a variety of topics, from politics to economy, from sports to entertainment [27]. Despite the tendency for the slow disappearance of the traditional media organizations [63], the number of information producers, especially non-professional ones, has grown considerably [22], forcing readers to face articles covering similar themes and similar events and it can be a daunting task to decide what to read and where. News feeds have taken a major role in today's news dissemination, more than traditional news outlets and even email subscriptions to online newspapers [35].

A news feed can be characterized as a collection of news items, usually from different sources, updated frequently and often in real-time. The multiple news sources to which a user subscribes are aggregated and organized for ease of reading. Personal social network feeds are a good example of this mechanism. A user decides which news sources they wish to subscribe to, and the news items are shown integrated into their social feed. A good example of this data structure adapted to



Rehabber Urban Winery Dream
www.thefeast.com
Few venture through the blocks of vacant land, boarded up windows, and overgrown weeds sprawling across the West Side neighborhoods unless for a special visit to Garfield Park Conservatory or one of the handful...
22 minutes ago · Like · Comment · Share

Thanks This is a little bittersweet.
The Sad, Beautiful Fact That We're All Going To Miss Almost Everything : Monkey See : NPR
www.npr.org
We take a moment to reflect on being "well-read," and on the fact that the smartest thing you can do as a consumer of culture is admit that you'll never know everything.
about an hour ago · Like · Comment · Share

commented on photo.
Profile Pictures
about an hour ago
View all 3 comments
You look beautiful! Congratulations.
about an hour ago

I am so overwhelmed!!! People from all over the country have sent me facebook messages and/or called me to wish me a happy birthday. I want to you know that it's because of YOU that I had such a happy birthday! THANKS SO MUCH!
39 minutes ago · Like · Comment
2 people like this.

Figure 1.1: Example of a Facebook news feed [13].
1: Article sharer 2: Article title
3: Article source 4: Article lead
5: Article share date

news consumption is the Facebook feed. A user can "like" a news outlet page in order to add their articles to the user's personal feed. The feed items will appear organized and integrated with other posts shared by friends or other "liked" pages. At any point, a user can "unlike" a page, causing their contents to no longer appear in their feed in the future. An example of such a feed can be seen in Figure 1.1.

News feeds offer an alternative to other news distribution methods, such as traditional printed versions and even email subscriptions to online newspapers, and have shown to be a more convenient medium for some users, especially in networks. The news items are shown to a user in a centralized way, and some platforms make it very intuitive for a user to add or remove a news source from their news feed.

The news industry has, over the years, adapted to this reality with their 24-hours news cycle, a continuous news investigation and reporting, derived from competition for audience attention and advertising [51]. This model associated with the distribution and organization of the news industry has many consequences, ranging from the decrease of the news reports quality, in the urge to release more articles as fast as possible [27], to the production of articles targeting more generic audiences, which encourages the mainstreaming of covered subjects in determent of topics that could be considered niche or minor, impacting the diversity of news feeds [16].

In the wake of all these modern-day challenges, the necessity for systems and tools capable of filtering and processing large amounts of information (and, particularly, news articles and news feeds) has emerged. News feed comparison can be used in a variety of applications. The diversity of news sources and the high volume of articles frequently result in the same news to be repeated many times by different sources. Clustering news sources and similar news articles can help reduce such redundancy [83]. Also, as highlighted previously, information overload can be a deterrent for news readers, and automated selection for user-specific relevant news items can be useful [31]. The analysis and handling of this amount of news data can be used as a tool to perform different kinds of media-related studies, such as measuring information consumption diversity [32]. The search for new methodologies and algorithms capable of optimizing these processes is a developing field. Its purpose remains focused on helping news readers and, by extension, the general population to be informed and updated in an ever-changing world.

## 1.2 Objectives

This thesis aims to study, implement, and evaluate different measures to compute the degree of similarity between news feeds. First, we will collect information on current techniques for news feed comparison. This task will be subdivided into two parts. There will be an analysis of the methodologies used to process the news feeds as a data structure and how to approach comparison challenges related to their inherent segmented nature. Then, we will address the text similarity techniques used to compare the news articles' text, especially those used regularly in information retrieval systems.

Having described the state of the art in news feed comparison systems, this thesis aims to test the different techniques and measure their effectiveness in the comparison of news feeds. The approach (or set of approaches) that perform best will be subject to a broader set of tests in both professional and personal news feeds, such as mainstream media outlets and social media timelines.

## 1.3 Thesis Structure

Besides this introduction, this thesis is divided into six other chapters. Chapter 2 details the current state of the art in news feed comparison techniques. Chapter 3 explores different text similarity measures used in information retrieval systems. Chapter 4 will provide an overview of the datasets, including the metrics involved in the classification of similar news for control purposes and the methods we will use for their extraction, as well as the implementation details of the tests used for effectiveness measurements, including the results derived from them. Chapter 5 details the further testing conducted using the methodologies that performed the best in the previous tests, as well as the respective results. Chapter 6 presents the final conclusions of the theses, including possible improvements to be made to the overall process and possibilities of further work.

# Chapter 2

# Comparing News Articles and News Feeds

This chapter provides an overview of the methodologies used in systems that analyse and compare news articles and news feeds. We will delve into what source material is used to test these systems and how it is collected. We will then look into the different techniques used to preprocess the collected news articles and feeds before they can be used. We will finally take a deeper look at the methods used to compare and classify the news items in these systems and what conclusions we can derive from these results.

## 2.1 General Architecture

In this chapter, we will overview many different news processing systems. We define a news processing system as a system that uses information retrieved from news articles or news feeds to generate knowledge. Despite the variety of objectives and techniques implemented by these systems, they often follow mostly the same implementation steps and have similar architectures. Figure 2.1 shows a simple diagram representing the general architecture of these systems.

The first step for many systems is data collection. It is necessary to have enough data to test the proposed systems and improve them based on the results. Most systems obtain their data online, through dedicated APIs, news feeds, or web crawlers, since it bypasses the need to manually input data into the system. Most systems then implement multiple preprocessing techniques in order to prepare the collected data for the implemented methodologies. Even if these techniques are common, not all systems implement such techniques [24].

Having the data ready, the systems then implement a variety of methodologies. These are obviously system dependent and vary with the system's objective, ranging from clustering methods to text similarity measures. These methodologies then produce the system's results – again, what these results actually are varies with each system. It is even possible for the systems to use

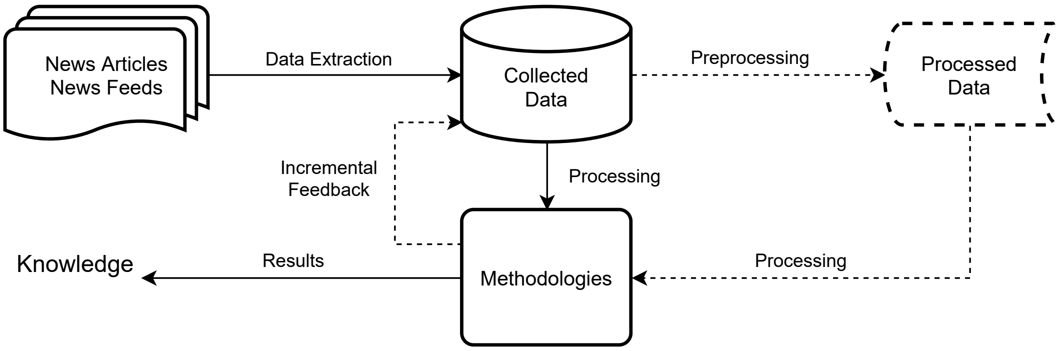


Figure 2.1: General architecture of news processing systems.

the obtained results to alter the available dataset [23], while other systems present a more linear approach.

## 2.2 Datasets and Data Extraction

The first step into testing an implementation of a news article analysis system involves collecting sample data, capable of mirroring the system's behaviour in a real environment. Given the aforementioned online nature of news dissemination nowadays, the Internet is where most of this test data is collected from. Even traditional news providers, such as newspaper and broadcast agencies, now have digital versions of their products [28]. As of 1998, more than 2,900 newspapers had online presence. In 2020, there are over 50,000 newspapers, magazines, radio stations, and television broadcast stations that operate partially or fully online around the world [6]. Moreover, the Internet has become a preferential source for news consumption, as shown by recent studies. In the United Kingdom, 77% claim to use the Internet as a source of news, leading other sources like television (51%) and print media (22%) [99].

The three main ways of obtaining news articles in order to build a testing dataset for a system are HTML page crawling, accessing RSS feeds, or using social network channels. We will delve into each one of these alternatives individually. There are other employed methods, though their use is rare. For example, sometimes data can be obtained from physical sources not available online, usually because of their old publication date, and coded manually into databases. For example, there are systems that utilize news articles from printed Belgian newspapers prior to 2003 [17] and handmade transcriptions from German public news television broadcasts [89]. These strategies are avoided when possible due to their significant manual labor cost.

### 2.2.1 HTML Documents

Many media organizations publish their news articles online on their own websites in the form of standard HTML (HyperText Markup Language) pages. This is how most online newspaper subscribers read their articles and thus is a way to obtain most articles available online. There has been much debate on whether or not web crawling is fair game, mostly due to authorship concerns. In late 2019, United State courts decided in favor of legalizing crawling when used to extract data that is publicly available and not copyrighted, even if the news items themselves are [96].

The usage of HTML documents as data sources faces some problems. HTML pages, as the base of online content, are not standardized by nature. There is no predefined structure that news articles written in HTML pages have to follow. This causes problems when retrieving information since the news item itself is “hidden” among all the other information the page contains, such as navigational elements, advertisements and images. It also has a scalability problem since complete HTML pages’ size can be a problem during data transfer [97].

This leads to some problems when analysing HTML pages for news extraction. Tools that perform the crawling and scrapping of the HTML documents have to select, in an automatic way, which parts of the document contain relevant information, having to process pages that do not contain static elements, including advertising. That said, many systems still use data from HTML documents for testing due to their variety and availability [97].

### 2.2.2 Web Feeds

HTML is not the only way news articles can be found online these days. Web feeds have become a common alternative for article publishing. For a regular news reader, these feeds have multiple advantages. It allows for news updates in real-time with no effort by the reader, it centers news in a single place, avoiding the need to visit multiple websites individually, and it does not require the user to subscribe to the traditional media newsletters. A web feed subscription is faster and can be edited or cancelled at any time. Web feeds can be read by using a feed aggregator. A feed aggregator is a software that allows the merging of feeds from various sources and displays the news items in a user-friendly way. A feed aggregator is capable of fetching the updates to the feeds and show to the user news items it considers relevant [91]. Many news websites provide web feeds, making it an alternative to other online media consumption methods. As of 2005, 31% of online users utilized web feed technology [41].

Regarding their application for data collection for news system testing, web feeds offer multiple advantages. Their structured nature overcomes many of the problems HTML parsing has. All data present in the feed is relevant and organized, facilitating its processing. Also, their relatively small document size, derivated from the lack of content that is unrelated to the feed items themselves, is an asset for large-scale analysis.

RSS (Rich Site Summary) has become a widely used data format for this purpose [7]. It is an XML-based format that can be used to describe and syndicate web content. Initially designed as a format for news sites, it has expanded its utility to help web blogging, podcasting, and other web-based publishing [10]. Another data format that is used for the same purposes is Atom [8]. Also an XML-based format, it tends to me more descriptive than RSS, containing more detailed information on each entry.

A recurring platform used by news systems as a data source is Media Cloud [3]. Media Cloud is an open-source platform that serves as an archive of news articles published online for research purposes. It updates its content continuously using many media sources web feeds and stores articles for future use. It currently stores over two hundred million news stories from over twenty-five thousand news sources. All of the information that can legally be distributed is accessible through Media Cloud's API [4]. Its use in news processing systems ranges from main news article source [27, 36, 18] to reference for comparison [40].

### 2.2.3 Social Network Channels

Another way for news readers to consume articles is through social networks. Many mainstream news providers have a social network presence and distribute their core articles on social platforms. Social networks are also a common outlet for articles produced by non-professional journalists, previously mostly present in web blogs [22]. Social networks can work as feed aggregators and integrate news articles produced by sources selected by the user into their personal social feed or simply use content exclusively produced by its own users in a similar way. Two of the social network platforms that were commonly used as news sources for research purposes in the articles we analyse later in this chapter are Facebook and Twitter.

Facebook allows limited access to its stored information through a dedicated API [1]. It allows direct access to public pages and posts, which is the common way for news providers to share their content. Besides this, it also allows for access to users, groups, and events, which can be useful for studies that need more information other than news articles themselves. Since it is structured in a graph, it also allows easy connection between the different social entities, if such is important for the task at hand.

Twitter also has its API for content access. It allows access to all public messages and conversations. Unlike Facebook's API, it also contains tools for real-time content streaming, random sampling, event categorizing, and timestamp filtering [9]. All these features can help to extract specific samples of information dedicated to a specific need, from data relating to a specific event to a representative sample of a population's messages in a specific period of time.

## 2.3 Data Preprocessing

Natural language processing by computers is an old challenge, due to the its inherit complexity and ambiguity. Thus, multiple preprocessing techniques are usually applied to the news articles datasets to facilitate the analysis process. The objectives of these methods are to identify important terms or phrases, remove extraneous information, or agglomerate similar concepts. The book *"Introduction to Information Retrieval"* by Manning et al. [61] will be used as a guide for this section since it contains a summary of multiple common techniques currently used in a variety of text analysis scenarios. We will detail the ones we found more common or otherwise relevant regarding news processing systems.

The first step in data preprocessing usually involves **Tokenization**. This method consists of dividing a character sequence into multiple logical units called tokens, representing a specific concept. This process frequently removes other text signs, such as accents, diacritics, and punctuation marks. The created units often correspond to single words, but not always. For example, the phrase *New York* represents a single concept and could be considered a single token. Tokenization is often useful when used to help information retrieval systems establishing a term dictionary.

Another frequently used technique is **Stop Words Removal**. Natural language documents usually have very common words such as pronouns and conjunctions that do not offer significant discriminatory value. The terms to be excluded can be determined through each term's collection frequency (the number of times a specific term appears in the documents of the collection) or through a *stop list*, a predefined list of words to be ignored. Even though stop lists have mostly fallen out of favour in modern search systems, it occasionally is still used in news processing systems.

Having identified the relevant content inside a document, many systems attempt to normalize the text, grouping similar tokens together. One existing technique for this effect is **Case-Folding**. The principle of this method usually consists of reducing letters to lower case, allowing the system to recognize instances of the same word with different capitalisation as equal. Case-folding should be used carefully in order to avoid clustering of different concepts under the same term. For instance, it would be unwise to eliminate the distinction between the word 'Windows', the operating system, and 'windows', the building feature in plural form. For that reason, this technique is ordinarily used selectively, affecting only words under specific circumstances, like at the beginning of sentences.

Finally, taking normalization a step further, it is useful to consider words with inflection variation, such as verb tenses, or words derived from the same origin and have a similar meaning, to be represented by the same base form. There are two main techniques aimed at accomplishing this: **Stemming** and **Lemmatization**. Stemming can be seen as a greedy approach to the problem, aiming to remove affixes from words with the use of crude heuristics, in an attempt to reduce similar words to the same form most of the time. Lemmatization uses a more precise approach, using vocabulary lists as well as word analysis in order to obtain a word's base form: the *lemma*. The distinction between the two methods is evident in the following example: given the word 'saw', stemming would probably remove the verb inflection and return 's', while lemmatization would attempt to analyze the context and return 'see', the verb, or 'saw', the noun [61]. Stemming is particularly common among news processing systems. The most used algorithm is *Porter's Algorithm* [77], which is designed for text documents written in English.

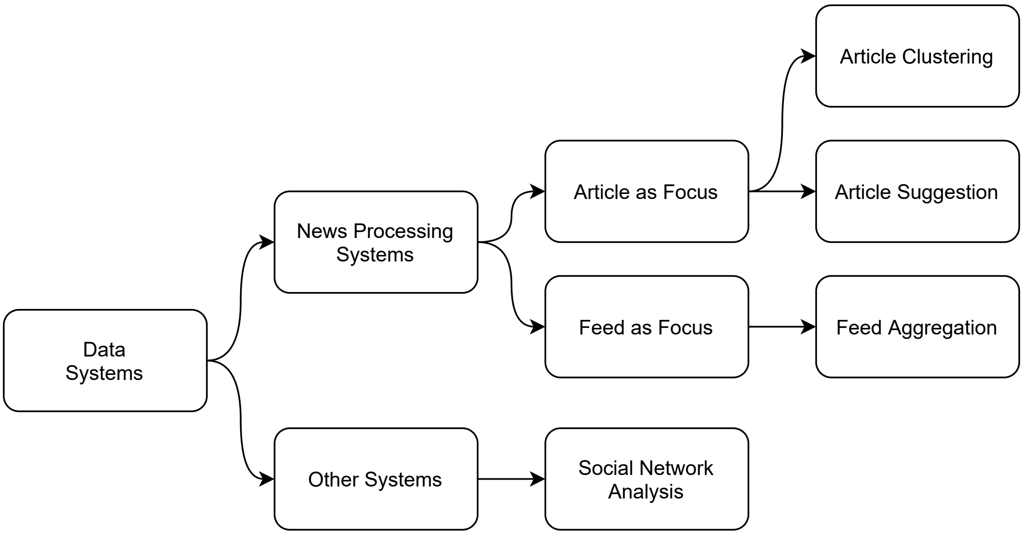


Figure 2.2: News processing systems grouping and structure.

## 2.4 News Comparison and Classification Techniques

Having described the data collecting process and the methodologies used to prepare the news dataset, we must now detail the systems that use this data, as well as their objectives. We have outlined some of the possible applications for these kinds of systems in Section 1.1. We will now take a deeper look at some of these systems, grouped by their objective: systems that cluster articles with similar themes or subjects, systems that suggest articles to users based on their preferences, and systems that look at news feeds as a unit and perform analysis with that point of view. Given the prevalence of social networks as a source of data, we will also take a brief look into systems that use similar techniques to the other systems but use non-news data collected from social networks. The overall structure and grouping of these systems is summarized on Figure 2.2.

### 2.4.1 Article Clustering

As mentioned in Section 1.1, the fact that there are many sources of news producing content can cause many articles to be similar or to be referring to similar subjects. Therefore, there is great utility in attempting to group articles for ease of access. This can be done on many different axes. Articles may be clustered based on the fact that they report the same event or, in a more general approach, they can be classified into subject categories, such as economics, politics, among others, and be grouped with articles that cover similar topics. This last approach is particularly useful when implementing a system that allows a user to browse for articles of a particular genre. The range of systems that attempt to cluster articles is very wide, from systems that attempt to understand news diversity in a specific region [87] to systems that perform spatial analysis and try to comprehend news diversity in the world [52]. We will discuss some examples of these systems and their main characteristics.

A common priority in many systems that perform article clustering of some sort is how to deal with dimensionality problems. Many of the techniques used in these systems have already been tested for small data collections, and the challenge many times is focused on finding strategies to have the same type of techniques to perform well in large data analysis, where most of the news article comparison problems lie. One such system attempts to perform RSS feeds classification combining Term Frequency-Inverse Document Frequency (TF-IDF) and a domain ontology based on metadata collected from news repositories [11]. This metadata is used to assign different weights to specific terms and concepts that can be used to classify and group articles. TF-IDF is a common information retrieval measure used to determine the relevance of specific terms in a document collection. It has inherent dimensionality problems, because it requires knowledge of term distribution in the whole document collection and can produce noise when applied to large data analysis. This system tackles this problem by testing with real users retrieving documents of their choice of category.

Another system that attempts to group articles in very large news article collections used information retrieved from Media Cloud [3], a news article repository mentioned in Section 2.2.2. This system attempts to perform topic extraction from large samples of news articles and classify them using a statistical topical model [27]. The technique utilized by the system for text comparison is Latent Dirichlet Allocation (LDA) [20], a common topic extraction method used in information retrieval systems. The model built with LDA uncovered a large variety of relevant topics – in far greater number than what would be reasonable for analysis. The topics were then hand-selected and aggregated into groups in order to reduce the overall number. Overall, the system was able to identify the topic similarity of different media sources in a data collection with over two hundred million entries.

The concern about dimensionality is not exclusive to systems that analyse web feeds and large news media repositories. Systems that use data collected from social media face a very similar problem. Social network presence is very high among the general population these days, and a lot of content is generated daily. As of 2015, 65% of American adults regularly use a social media website [74]. It is often the case that data derived from social network platforms – usually collected through APIs – is sampled semi-randomly from the available data, in an attempt to obtain a representation of the available content. One such system attempts to retrieve the most emergent topics discussed on Twitter within a specific time frame [24]. It used Twitter's API, which included a feature that allows for random sampling of public messages, as discussed in Section 2.2.3. The system extracted keywords from the collected data and used PageRank's authority system [68] to determine emergent topics. Despite collecting data from a sixteen-day period, the system was able to detect over 300,000 different keywords from a sample of over three million tweets.

Other systems attempt to bridge the gap between data collected from social media platforms and data from traditional media sources. One such system attempts to compare the two and understand the similarity between their content [31]. It behaves similarly to the other aforementioned article clustering systems. It performs keyword extraction using TextRank [64] and Term Document Frequency for graph construction. The authors performed tests with a variety of different text similarity measures, having Dice's Coefficient [34] and Jaccard Similarity [48] performing the best. The system was successful in analyzing the prevalence of topics in both social and traditional media over time, especially concerning variety.

A recurring topic in these systems is the use of text similarity techniques, common in information retrieval systems, as a means of article comparison. Many of the systems implement more traditional techniques like Cosine Similarity [31, 76], while others obtain better results with more recent methodologies, like the previously mentioned LDA [27, 87]. Some of the previously mentioned systems even attempt to implement multiple methodologies, trying to understand which performs better in the given scenario [31]. There are even some cases where the testing of different similarity measures is the focus. One such example is a system that attempts to test the effectiveness of a variety of such techniques when combined with different clustering methodologies [19]. Some of the tested techniques include Simple Matching [43], Russel & Rao Measure [82], and Tanimoto Coefficient [81]. The system utilizes the aforementioned Jaccard's Similarity as a baseline.

### 2.4.2 Article Suggestion

Another peril of modern-day news reading mentioned in Section 1.1 is information overload. It can be very hard for a reader to parse the information that they come across. Therefore, there are also many systems dedicated to select news articles that would be relevant to specific users. For this chapter's purpose, we consider an article suggestion system if it either considers a user's past reading experience and other obtainable information to decide what articles are relevant or select articles targeted to a specific demographic. Many of these systems result in building news feeds customized to a specific user: a smart news feed [58].

A system that is representative of article suggestion was developed by Burkepile et al. [23]. It suggests articles collected from RSS feeds based on a sample set previously found relevant by a user. This sample acts as a training set for future suggestions. Each candidate article is converted into multiple cells – sets of words randomly collected from the article – that are used for comparison. This approach was chosen in order to tackle the large size of articles collected from RSS feeds. Similar to many systems analyzed in Section 2.4.1, it uses Jaccard Similarity [48] to compare the training set and the generated cells. Each article that is found relevant is then integrated into the training set for future comparisons. The system is able to retrieve relevant articles with a very small number of false positives, even when the initial training set provided by the user is of very small sizes, including one example containing only five articles.

Another interesting example is a system that tackles a different problem also highlighted in Section 1.1 – the normalization of reported subjects and the difficulty of finding different points of view regarding a specific matter [92]. This system attempts to improve on other recommending systems by finding articles that offer a range of opinions regarding a specific topic. It attempts to analyze stylistic features in the articles using Linguistic Inquiry and Word Count [73]. The system proved not very effective when run autonomously but improved significantly when receiving user acceptance as feedback.

Despite not having to deal with the same scalability problems as the article clustering systems, there is some research done in article suggestion at large scale. One such example is a system that studies the generation of personal article recommendations to users of Google News [2]. The system's purpose is to test general article suggestion techniques in a large and dynamic environment [30]. It uses MinHash [21], Probabilistic Latent Semantic Indexing [46], and covisitation counts. The system is also built in a domain-independent way, meaning that the techniques utilized can be easily adapted to other environments.

All of the previously discussed systems utilize a user's past reading experience in order to suggest articles that may be relevant. This is not always the case. Systems that extract data from social networks can use other available user information to attempt to find relevant articles to suggest. One such example is a system that attempts to match user profiles from Twitter to news feeds articles using topical extraction [90]. It implements text similarity techniques, some of which were already mentioned in previous systems: Latent Semantic Analysis (LSA) [53] and Latent Dirichlet Allocation (LDA) [20]. Moreover, and unlike previously mentioned works, this system doesn't implement the methods separately and then compares results, but computes the average similarity value of the different techniques and uses it for comparison. The system also performs an interesting analysis regarding topic distribution among regions, allowing for article recommendations based on a user's geo-location.

Some systems also attempt to detect articles not targeting a specific user but instead a certain demographic or subject. One such system attempts to identify articles that would be suitable for children [75]. It compares two feeds from a British mainstream broadcaster: a normal unfiltered feed and one specifically designed for children and uses them to train classifiers like Naïve Bayes and Support Vector Machines. The system was able to identify specific terms inside the database that were strongly associated with child-appropriate articles – such as *cute*, *zoo* or *animals*, and terms that were associated with articles targeting an adult audience – like *Iraq* or *government*. There are other similar systems that implement the same techniques for identifying news relating to a specific subject, like articles with science-related concerns [91].

### 2.4.3 Feed Aggregation

A characteristic present in all the systems mentioned in the two previous sections is that the unit used for comparison and analysis is the article. Even when the data is obtained from news feeds, articles are seen as independent, and frequently the connection between articles of the same feed is not taken into consideration. This is not the case of all news processing systems. Some choose to take the news feed as the basic unit of the system, aggregating their respective articles. Despite the obvious loss of individuality of the articles themselves and, as a consequence, a higher difficulty to obtain results at the article level, such as article recommendations, this strategy can prove very helpful in developing systems that aim to obtain results at a higher level, such as similarity of content between news sources.

There are even some similar strategies implemented in projects that do not directly involve information retrieval. One example is an analysis of Belgian newspapers over time, searching for a possible decrease in content diversity over a thirty-year time span [17]. The interesting approach is that articles that are deemed to report the developments of the same event are grouped together in the same document, called a news story. These news stories are the unit that is used for comparison. However, most of the work performed is done manually and without the use of strategies like topical models or trained classifiers, so we will not discuss the used methodologies. A similar approach is used in a different system that compares German broadcasted television news with social media articles [89], where the comparison is also mostly manual.

A more relevant work for the context of this thesis is a paper describing possible methods for similarity computation of RSS feeds [97]. It performs an analysis of the different data formats and highlights the characteristics of RSS feeds that can be exploited to improve the results of traditional analysis of HTML pages. Specifically, it is highlighted that the stable structure of RSS feeds – all articles have a headline, links, and a summary – can be used to predict where relevant information can be found. The paper proposes a variation of a fuzzy measure used in a previous system [66] but does not implement or test its effectiveness. A similar fuzzy measure is implemented in a different system that compares different RSS feeds [102]. It uses an ontology as an aid to transform headlines collected from the feeds into a graph that is used to measure similarity. Despite being developed targeting RSS feeds in general, the system is not tested with news articles themselves. Instead, it was integrated into a job-seeking agent. In every test run, the system was able to find a job that the user classified as relevant in the first ten results.

The most interesting example that we found that uses feed as a comparison unit is a news clustering system that attempts to summarize similar news articles and presents the result to a user [83], similar to many systems in Section 2.4.2. It uses three similarity measures, many of which were already mentioned in previous systems: Edit Distance, Jaccard Similarity [48], and Cosine Similarity [84]. Each one of these measures is applied to each collected feed as if it was a single document, considering two feeds as similar if the calculated value crosses a threshold. The system attempts to discard feeds that are similar to others in order to avoid showing similar news to users. Another interesting feature of this system is that the measures are not only implemented separately and then compared with each other, but also used together. In fact, the system performs significantly better when the similarity value compared against the threshold is an average of the different used similarity measures. This effectiveness was measured against control values, determined by humans that manually compared different news feeds.

This feed aggregation technique is also very common in systems that use data collected from social networks, mainly because of the short text length of individual posts. Aggregated feeds have more information in a single document and can provide better results in many situations. An example of a system that states this argument as a justification for feed aggregation is a study that tries to detect the presence of filter bubbles in personal Facebook news feeds of Danish users [16]. A filter bubble can be defined as a group of people subject to a restricted subset of the available information derived from selective recommendation algorithms. Similar to many other systems that utilize topical modeling, it implements LDA to measure the similarity between concatenated personal feeds. Aside from topical similarity, the system also calculates link similarity using a simple matrix of overlapping links between users. The approach allowed the identification of possible filter bubbles in small subsets of the studied samples of feeds, concluding that under 10% of the sample could be included in this group.

Systems that use data obtained from Facebook showed concerns about how little information content there is in a single article. These considerations are aggravated in data collected from Twitter, a platform with a stronger character limitation. This is very noticeable in a system that attempts to compare the contents of Twitter feeds with conventional media articles [103]. In this specific case, LDA is once again used for topical extraction from the chosen mainstream media feeds. Similar to other previously discussed systems, no aggregation is performed at this stage. However, a variation of LDA specifically designed for Twitter, which includes the aggregation of tweets, is proposed for the second stage of topical extraction. The system was able to find that the analysed Twitter feeds cover a significantly narrower range of topics when compared to traditional media feeds but is able to spread major news topics very effectively.

Another system that presents a very similar approach also tries to evaluate information diversity between social and mainstream media [32]. The main difference is that, in this case, all data used is collected from Twitter. Two sets of feeds are analysed: feeds from random Twitter users and feeds from Twitter users that follow a required number of UK mainstream media accounts. Unlike the previous system, timelines from both groups are rebuilt and concatenated into individual documents, each representing a single timeline. The system uses Cosine Similarity [84] inside each test group to measure diversity and was able to find that there is greater topical variety in the feeds of random users when compared to the ones that follow mainstream accounts.

### 2.4.4 Other Social Network Analysis

Looking at the systems that perform analysis over news articles and news feeds that we described in the previous sections, we can observe that social networks are a common data source for training these systems. However, and unlike other used sources of information, social networks host a considerable amount of data that is not news related. For that reason, we will now give a very brief overview of other applications of social network data that is not directly news-related but still utilizes many common methodologies to the news processing systems.

One such system uses all messages present in one's Facebook feed and attempts to classify them in a system analogous to what we could see in an email inbox [85]. Many email hosting websites perform a classification of incoming emails: besides the regular ones, some are marked as important, while others are deemed as "spam". Similarly, this system attempts to divide the posts in a Facebook feed into two main categories: friends posts and liked pages posts. The first group is further subdivided into two more categories: life events, many times equivalent to friend's status updates, and entertainment posts. This classification is done by using different trained classifiers, such as Naïve Bayes and Support Vector Machines. The purpose of this system is to allow users to filter their Facebook feeds, removing content that they do not wish to see in a given moment.

One particular social media feature that is used in multiple systems is Facebook's status updates. It allows users to easily create posts regarding any content and have some desirable properties: they are shorter than full reviews or news articles but are longer and usually better written than tweets since the character restriction is much more forgiving [12]. For this reason, it is the focus of many systems. One such system attempts to classify the mood of the person writing the status update into good, bad, or average [86]. It used Support Vector Machines and was able to obtain 70% accuracy when compared with cross-validation files. Another similar system divides the status updates into only two categories: positive and negative [12]. It uses a multitude of different trained classifiers, some of them associated with pre-labeled data. It achieves better results than the previous system, with some of the classifiers reaching 85% accuracy.

More than analysing personal user feeds, other systems attempt to extract information about samples of users. For instance, one system proposes a new topical model for Twitter, specifically designed to work with health-related topics [70]. This new model is tested against LDA [20], showing better isolation of specific diseases, such as influenza, infections, and obesity. The same authors then expanded the system's functionalities, adding the ability to track illnesses over time, behavioral risks, illness spatial distribution, and medication usage [71]. The authors use this system's results to defend the use of social networks for research of social matters like public health.

Finally, another common use of social networks is publicity and marketing. One system performs analysis of companies' marketing posts on Facebook [101]. It proposes the division of such posts into two groups: marketing messages, that aim to improve direct sales or advertise promotions, and communication messages, that have no direct sales promotion and mention other events, such as holiday celebrations. This classification is done using multiple trained classifiers, some of them utilizing TF-IDF weights. The system measures a message's popularity based on the number of "likes" and comments that it has and found that, despite marketing messages being by far the most common, communication messages are much more popular.

## 2.5 Conclusions

Having described the systems that perform analysis of news articles and feeds, we can conclude that, despite having a great variety of objectives, the systems use many common methodologies. A summary of these systems can be found in Table 2.1. We can identify two main approaches: systems that focus on the article as a unit, which allows a more fine-grained analysis, and systems that use feeds as units, which have advantages when the project's scope is wider. We can see that the first approach is far more common and that the highlighted advantages of such methodology can be further explored.

Moreover, we can also identify two main methodologies to analyse the textual data itself: topical models and trained classifiers. Not only are topical models more common, but also the best performing trained classifiers often are combined with information retrieval measures to improve their results. However, when looking closer at what topical models were used, we can also find that the systems mostly use the same similarity measures – Cosine Similarity, Jaccard Similarity, and Latent Dirichlet Allocation. Also, the number of studies that compare the effectiveness of different measures is quite low [90, 83].

Table 2.1: Summary of news processing systems.

| **Article** | **Objectives** | **Techniques** |
|---|---|---|
| **Article Clustering** | | |
| Bergamaschi et al. (2007) [19] | Testing the effectiveness of different text similarity measures when combined with different clustering methodologies | Simple Matching, Russel & Rao Measure, Tanimoto Coefficient, Jaccard Similarity |
| Pons-Porrata et al. (2007) [76] | Proposal of a topic discovery system to extract knowledge from news streams | Cosine Similarity |
| Cataldi et al. (2010) [24] | Retrieval the most emergent topics discussed on Twitter within a specific time frame | PageRank |
| Agarwal et al. (2012) [11] | Classification of RSS feed news items | TF-IDF |
| Chuang et al. (2014) [27] | Topic extraction from Media Cloud and classification using a statistical topical model | LDA |
| Davis et al. (2016) [31] | Social media platforms and data from traditional media sources | TextRank, TF-IDF, Dice's Coefficient, Jaccard Similarity |
| Sjøvaag et al. (2018) [87] | Evaluation of the impact of direct press support on the diversity of online news in Norway | LDA |
| **Article Suggestion** | | |
| Das et al. (2007) [30] | Testing general article suggestion techniques in a large and dynamic environments | MinHash, Probabilistic Latent Semantic Indexing, Covisitation counts |
| Thelwall et al. (2007) [91] | Extraction public science-related fears from RSS feeds | Term extraction, Heuristic clustering |
| Burkepile et al. (2010) [23] | Suggestion of articles collected from RSS feeds based on a sample set previously found relevant by a user | Jaccard Similarity |
| Polajnar et al. (2012) [75] | Identification of articles suitable for children | Trained classifiers |
| Tabara et al. (2016)[90] | Matchmaking between Twitter user profiles and news feeds articles using topical extraction | LSA, LDA |
| Tintarev et al. (2018) [92] | Suggestion of articles offering a range of opinions regarding a specific topic | Linguistic inquiry, Word Count |
| **Feed Aggregation** | | |

| | | |
|---|---|---|
| Wegrzyn-Wolska et al. (2005) [97] | Proposal of methods for similarity computation of RSS feeds | Fuzzy measures |
| Yuan et al. (2010) [102] | Semantic matchmaking of RSS feeds for external applications' use | Fuzzy measures |
| Zhao et al. (2011) [103] | Comparison of the contents of Twitter feeds with conventional media articles | LDA, LDA variations |
| Bechmann et al. (2018) [16] | Detection of filter bubbles among Danish Facebook users | LDA |
| Devezas et al. (2018) [32] | Evaluation of information diversity between social and mainstream media | Cosine Similarity |
| Beckers et al. (2019)[17] | Analysis of Belgian newspaper content diversity over time | Manual classification |
| Steiner et al. (2019) [89] | Comparison of German broadcasted television news with social media articles | Manual classification |
| Sakhapara et al. (2019) [83] | Clustering similar news articles using the feed as a comparison unit | Edit Distance, Jaccard Similarity, Cosine Similarity |
| **Other Social Network Analysis** | | |
| Ahkter et al. (2010) [12] | Classification of Facebook's status updates | Trained classifiers, Pre-labeled data |
| Paul et al. (2011) [70] | Proposal of a health-related topical model for Twitter | LDA |
| Yu et al (2011) [101]. | Analysis of companies' marketing posts on Facebook | Trained classifiers, TF-IDF |
| Paul et al. (2012) [71] | Extension of a health-related topical model for Twitter | LDA |
| Shrivastava et al. (2012) [86] | Classification of Facebook's status updates | Trained classifiers |
| Setty et al. (2014) [85] | Classification of Facebook feed message content | Trained classifiers |

# Chapter 3

# Text Similarity Measures

Chapter 2 allowed us to understand what techniques are used in news article comparison. One of the drawn conclusions was the prevalence of text similarity measures. Moreover, there is not much research done on the effectiveness of different measures when applied to news article comparison, either in isolation or in combination with others.

For these reasons, we will now provide an overview of the available text similarity measures, especially the ones relevant for information retrieval systems. There are multiple surveys available that collect these measures. We will use multiple surveys [39, 79, 44, 25, 69] as a starting point, since they complement each other and focus on different techniques. An overview of the contents of these surveys can be found in Table 3.1. This overview presents all the techniques found in the surveys. We will cover in detail only the ones that we found relevant to this document - meaning, the ones that could possibly be applied to news feed comparison. We will follow a structure similar to these surveys and divide the measures into three main categories: string-based similarity, corpus-based similarity, and knowledge-based similarity. A brief description of what defines these categories can be found in the beginning of each respective section.

## 3.1 String-Based Similarity

String-based similarity measures attempt to measure similarity by analysing string and character sequences, ignoring the text's semantic properties. We will divide this section into two subsections: one will describe the measures that evaluate similarity at a character level, while the other one will detail the ones that evaluate it at a term level.

### 3.1.1 Character-Based Similarity

**Longest Common Substring** [47] is perhaps the simplest version of this type of measures. It determines the degree of similarity between two strings by measuring the longest of all the substrings that are present in both strings. Given two strings $x$ and $y$, Longest Common Substring can

Table 3.1: Overview of text similarity measures.

| **Measure** | **Gomaa et al. [39]** | **Pradhan et al.[79]** | **Han et al.[44]** | **Chandrasekara et al. [25]** | **Patwardhan et al. [69]** | **This Document** |
|---|---|---|---|---|---|---|
| **Character-Based Similarity** | | | | | | |
| Longest Common Substring | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Damerau-Levenshtein | ✓ | ✓ | | | | ✓ |
| Jaro | ✓ | ✓ | | | | ✓ |
| Jaro-Winkler | ✓ | | | | | ✓ |
| Needleman-Wunsch | ✓ | | | | | |
| Smith-Waterman | ✓ | | | | | |
| N-gram | ✓ | ✓ | | | | ✓ |
| **Term-Based Similarity** | | | | | | |
| Matching | ✓ | | | | | ✓ |
| Euclidean | ✓ | | | | | ✓ |
| Manhattan | ✓ | | | | | ✓ |
| Cosine | ✓ | ✓ | ✓ | | | ✓ |
| Centroid | | ✓ | | | | |
| Dice | ✓ | | | | | ✓ |
| Overlap | ✓ | | | | | ✓ |
| Jaccard | ✓ | | | | | ✓ |
| Web Jaccard | | ✓ | | | | ✓ |
| Web Simpson | | ✓ | | | | ✓ |
| Web PMI | | ✓ | | | | ✓ |
| **Corpus-Based Similarity** | | | | | | |
| HAL | ✓ | | | ✓ | | ✓ |
| LSA | ✓ | ✓ | ✓ | ✓ | | ✓ |
| GLSA | ✓ | | | | | |
| ESA | ✓ | | | ✓ | | ✓ |
| CL-ESA | ✓ | | | | | ✓ |
| PMI-IR | ✓ | | | | | |
| SOC-PMI | ✓ | | | | | |
| LDA | | | ✓ | ✓ | | ✓ |
| NGD | ✓ | ✓ | ✓ | ✓ | | ✓ |
| NID | | ✓ | | | | ✓ |
| NCD | | ✓ | | | | ✓ |
| DISCO | ✓ | | | | | |
| Wapth | | | | ✓ | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Knowledge-Based Similarity** | | | | | | |
| Resnik | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Jiang & Conrath | ✓ | | | ✓ | ✓ | ✓ |
| Lin | ✓ | | | ✓ | ✓ | ✓ |
| Leacock & Chodorow | ✓ | | | | ✓ | ✓ |
| Wu & Palmer | ✓ | | | | | ✓ |
| Path Length | ✓ | | | | | |
| Hirst & St. Onge | ✓ | | | | ✓ | ✓ |
| Lesk | ✓ | | | ✓ | ✓ | ✓ |
| Vector Pairs | ✓ | ✓ | | | | |
| Fusion | | ✓ | | | | |

be calculated with Equation 3.1.

$$LCS(i,j) = \begin{cases} 0 & if\ i=0\ or\ j=0 \\ 1+LCS(i-1,j-1) & if\ x[i] == y[j] \\ max \begin{cases} LCS(i,j-1) \\ LCS(i-1,j) \end{cases} & if\ x[i] \neq y[j] \end{cases} \quad [79] \quad (3.1)$$

**Damerau-Levenshtein** [42] takes the previous measure one step further, determining the distance between two strings to be equal to the number of single-character insertions, deletions and substitutions to transform the first string into the second. Given two strings *i* and *j*, the Damerau-Levenshtein measure can be calculated with Equation 3.2.

$$DL(i,j) = \begin{cases} max(i,j) & if\ min(i,j)=0 \\ min \begin{cases} DL(i-1,j)+1 \\ DL(i,j-1)+1 \\ DL(i-1,j-1)+1 \end{cases} & otherwise \end{cases} \quad [79] \quad (3.2)$$

**Jaro** [49] is a measure that bases similarity on the common characters of two strings and their relative position. It is mainly used for spelling correction and detection. Given two strings *i* and *j*, the matching character index *m*, and half the number of transpositions *t*, the Jaro measure can be calculated with Equation 3.3.

$$Jaro(i,j) = \begin{cases} 0 & if\ m=0 \\ \frac{1}{3}\left[\frac{m}{|i|}+\frac{m}{|j|}+\frac{m-t}{|m|}\right] & otherwise \end{cases} \quad [79] \quad (3.3)$$

**Jaro-Winkler** [98] is a variation of the previous measure. It alters the equal value of all the terms of the second branch of the formula, giving more weight to strings that match from the beginning.

**N-gram** [15] similarity is designed to evaluate the similarity of a subsequence of N units in a text. The similarity value is calculated by identifying similar units and dividing that number by the total number of units. Given *u, u1, u2, u3...* units, the N-gram similarity can be calculated with Equation 3.4.

$$N(u) = N(u1) \times N(\frac{u2}{u1}) \times N(\frac{u3}{u2}, u1)... \qquad [79] \quad (3.4)$$

### 3.1.2 Term-Based Similarity

Multiple measures in this category consider texts in a vector space. For brevity reasons, this approach will be described first. Each text fragment can be converted into a vector, having each term being a dimension of a vector space. The value of each dimension corresponds to the frequency of that term in the respective text.

**Matching Coefficient** [39] is the simplest measure that uses this vector representation. It counts the number of dimensions that are not null in both vectors – meaning, the number of different terms that are present in both texts. Given two texts $t_1$ and $t_2$, represented in the vector space of dimension $n$, the Matching Coefficient can be calculated using Equation 3.5.

$$MC(t_1, t_2) = \sum_{i=1}^{n} \begin{cases} 1 & if\ t_{1_i} \neq 0 \wedge t_{2_i} \neq 0 \\ 0 & otherwise \end{cases} \qquad [39] \quad (3.5)$$

**Euclidean Distance** [39] is a common way to measure the distance between two points or vectors. It consists of the direct distance between the data points in their respective dimensions. It can be applied to text similarity when the texts are represented in the aforementioned vector space. Given two texts $t_1$ and $t_2$, represented in the vector space of dimension $n$, the Euclidean Distance between the two texts can be calculated using Equation 3.6.

$$ED(t_1, t_2) = \sqrt{\sum_{i=1}^{n} (t_{1_i} - t_{2_i})^2} \qquad [39] \quad (3.6)$$

**Manhattan Distance** [39] is an alternative to the Euclidean Distance. It assumes that the distance between two points can only be calculated in a grid-like path – meaning, each "step" can only consist of an increment in a single dimension. In an analogous way to the Euclidean Distance, it can be applied to text similarity if the texts are represented in a vector space. Given two texts $t_1$ and $t_2$, the Manhattan Distance between the two texts can be calculated using Equation 3.7.

$$MD(t_1,t_2) = \sum_{i=1}^{n} |t_{1_i} - t_{2_i}| \qquad [39] \quad (3.7)$$

**Cosine Similarity** [84] is another measure that considers the same vector space. The similarity is evaluated by calculating the cosine of the angle between the defined vectors. Given two texts $t_1$ and $t_2$, the Cosine Similarity can be calculated using Equation 3.8.

$$Cos(t_1,t_2) = \frac{\sum_{i=1}^{n} t_{1_i} t_{2_i}}{\sqrt{\sum t_{1_i}^2} \times \sqrt{\sum t_{2_i}^2}} \qquad [79] \quad (3.8)$$

**Dice's Coefficient** [34] was initially proposed as a measure to determine the association between different species in an agnostic way. When used to compare textual elements, both the number of common terms and the total number of terms in both texts are used. Given the number of common terms $h$ and the total number of terms $t$, the Dice's Coefficient can be calculated using Equation 3.9.

$$DC = \frac{2h}{t} \qquad [34] \quad (3.9)$$

**Overlap Coefficient** [39] is a variation of the Dice's Coefficient, being the only difference the fact that it considers two strings a match if one is a substring of the other.

**Jaccard Similarity** [48] is a measure that evaluates similarity by comparing the number of elements that belong to the intersection of the texts and the elements of the union of the texts. Given two texts $t_1$ and $t_2$, the Jaccard Similarity can be calculated using Equation 3.10.

$$Jac(t_1,t_2) = \frac{|t_1 \cap t_2|}{|t_1| + |t_2| - |t_1 \cap t_2|} \qquad [79] \quad (3.10)$$

**Web Jaccard Measure**[79] is a variation of the Jaccard Similarity where the value is considered null if the intersection between the two texts is below a predefined threshold.

**Web Simpson Measure** [62] is a variation of the Web Jaccard Similarity where two strings are considered a match if one is a substring of the other. For efficiency reasons, and due to the large number of possible matches due to the substring rule, this technique usually requires the removal of unnecessary words from the documents. Given two texts $t_1$ and $t_2$, and a predefined threshold $C$, the Web Simpson Measure can be calculated using Equation 3.11.

$$WSM(t_1,t_2) = \begin{cases} 0 & if\ h(t_1 \cap t_2) < C \\ \frac{h(t_1 \cap t_2)}{min[h(t_1),h(t_2)]} & otherwise \end{cases} \qquad [79] \quad (3.11)$$

**Web Point Wise Mutual Information Similarity** [93] is yet another variation of the Web Jaccard Similarity, mainly applied in information theory and statistics. Given two texts $t_1$ and $t_2$ and a predefined threshold $C$, the Web Point Wise Mutual Information Similarity can be calculated using Equation 3.12.

$$PMI(t_1,t_2) = \begin{cases} 0 & if\ h(t_1 \cap t_2) < C \\ \log_2(\frac{h(t_1 \cap t_2)}{h(t_1)h(t_2)}) & otherwise \end{cases} \quad [79] \quad (3.12)$$

## 3.2 Corpus-Based Similarity

Corpus-Based similarity measures distinguish themselves from the ones described in the previous section because they are semantic measures – meaning, they evaluate similarity based on the meaning of the words instead of their character sequence. These measures in particular perform text comparison based on information gained from large corpora, which consists of large collections of texts from different knowledge domains.

**Hyperspace Analogue to Language** (HAL) [60] builds a co-occurence matrix, where both rows and columns represent vocabulary terms. The value of each element of the matrix indicates the strength of association between the two words. This means that words that have a tendency to appear near each other will have a higher strength of association. The matrix is built by sliding a window of variable size over the text and increment association values inversely proportional to the distance of each term to the focused word. To mitigate dimensionality problems, it is possible to drop out columns or rows with low entropy values. Having obtained the matrix, term vectors can be formed by considering the row and column of that term. The similarity between terms can then be calculated by measuring the distance between both vectors, using, for instance, Euclidean or Manhattan distance.

**Latent Semantic Analysis** (LSA) [53] also assumes that words that appear close together have a similar meaning. However, instead of directly identifying what words appear close to one another, it performs a word count in a predefined section of a text – a paragraph, for instance. It builds a matrix where rows are associated with unique terms and columns represent each text section. The value of each element of the matrix corresponds to the term-count of the term in the respective text section. Again, to avoid dimensionality problems, the matrix can be reduced in size with techniques like Singluar Value Decomposition. The matrix can then be converted into vectors – one for each row – and term similarity can be assessed by applying Cosine Similarity or Web Jaccard Similarity to those vectors.

**Explicit Semantic Analysis** (ESA) [38] is another measure that compares texts using a vector/matrix representation and was initially designed to compare Wikipedia entries. For each term, a vector is created, having each vector dimension represent a different text. The value of each of the vector's dimensions corresponds to the TF-IDF weight between the term and the corresponding text. The similarity of two terms is then calculated by measuring the Cosine Similarity between two vectors. It is possible to build the vectors the opposite way – meaning, having one vector for each text and each term corresponding to a different dimension -, allowing for direct text similarity calculation when two vectors are compared.

**Cross-Language Explicit Semantic Analysis** [78] is an extension of the previous method that takes advantage of document collections where each document is available in multiple languages, such as Wikipedia. It is, therefore, able to produce document vectors that are language-independent and allows for inter-language document comparison.

**Latent Dirichlet Allocation** (LDA) [20] attempts to represent each document not as a set of terms, but as a vector representing the topics that relate to the document. It assumes that the documents are composed of a random mixture of topics and the topics themselves can be represented by a distribution over all the available terms. It attempts to reverse engineer the process of creating the document in order to reach the underlying topics. After obtaining the topic vectors, the similarity between documents is usually measured using Cosine Similarity. This technique has the obvious advantage of reducing the dimensionality, given that the topics uncovered are always significantly less than the total number of terms.

**Normalized Google Distance** (NGD) [29] is calculated the similarity of two terms based on the hits returned by searching them in the Google search engine. Given two terms $i$ and $j$, the function $f$ that returns the number of page hits of a specific term, and $N$ the total number of searched pages, the NGD can be calculated using Equation 3.13.

$$NGD(i,j) = \frac{max[log\ f(i), log\ f(j)] - log\ f(i,j)}{log\ M - min[log\ f(i), log\ f(j)]} \qquad [79] \quad (3.13)$$

We can see that, if two terms only appear separate, their NGD is infinite. On the other hand, if two terms only appear together, their NGD is zero.

**Normalized Information Distance** (NID) [57] considers the distance between two strings as the length of the shortest binary program in the reference universal computing system that can output each of the strings when the other is given as input. Given two strings $i$ and $j$, and the binary program $k$, the NID can be calculated using Equation 3.14.

$$NID(i,j) = \frac{k(i,j) - min[k(i), k(j)]}{max[k(i), k(j)]} \qquad [79] \quad (3.14)$$

**Normalized Compression Distance** (NCD) [94] is presented as an alternative when NID is not computable. NCD uses a function that matches strings to a compressed version of them. Given two strings $i$ and $j$, and the compressor $C$, the NID can be calculated using Equation 3.15.

$$NCD(i,j) = \frac{C(i,j) - min[C(i), C(j)]}{max[C(i), C(j)]} \qquad [79] \quad (3.15)$$

## 3.3 Knowledge-Based Similarity

Knowledge-based similarity measures attempt to assess the similarity between terms using information obtained from semantic networks. The idea behind these networks is to link nouns, verbs, adjectives and adverbs into nets of cognitive synonyms, each expressing a different concept [39]. The most popular semantic network is WordNet [65], an English lexical database.

Many of these measures use the concept of Least Common Subsumer (LCS). The LCS of two terms can be described as the most specific concept which is an ancestor of both terms [72]. For example, assuming the following relationships:

- a cat is a feline
- a lion is a feline
- a feline is an animal
- a zebra is an animal

we can say that the LCS of “cat” and “lion” is “feline”, while the LCS of “cat” and “zebra” is “animal”. For this chapter, we will assume that there is a function *lcs* that, given two concepts, returns their LCS.

Another important concept that is recurrent is Information Content (IC). IC can be defined as a measure of specificity of a concept – meaning, more specific concepts have a higher information content. As an example, “carving fork” has a high IC, while “entity” has a low IC [69]. Given a concept $a$ and $P$ the estimated probability of finding a given concept in a large corpus, IC can be calculated using Equation 3.16.

$$IC(a) = -log(P(a)) \qquad [69] \quad (3.16)$$

**Resnik Measure** [80] considers the similarity between to concepts to be equal to the information content of the LCS of those concepts. Given two concepts $a$ and $b$, the Resnik Measure can be calculated using Equation 3.17.

$$Res(a,b) = IC(lcs(a,b)) \qquad [69] \quad (3.17)$$

**Jiang & Conrath Measure** [50] extend the Resnik measure by considering the path length between two concepts, resulting in a combination of the IC of the concepts themselves in addition to the IC of the concept’s LCS. Given two concepts $a$ and $b$, the Jiang & Conrath Measure can be calculated using Equation 3.18.

$$JC(a,b) = \frac{1}{IC(a)+IC(b)-2\times IC(lcs(a,b))} \qquad [69] \quad (3.18)$$

**Lin Measure** [59] has some similarities with the Jiang & Conrath Measure, simply for the fact that both consider the IC of both the concepts themselves as well as the IC of their LCS. The Lin Measure evaluates similarity by considering the ratio of the information required to understand the concept's similarity to the information of the concepts themselves. Given two concepts *a* and *b*, the Lin Measure can be calculated using Equation 3.19.

$$Lin(a,b) = \frac{2 \times IC(lcs(a,b))}{IC(a)+IC(b)} \quad [69] \quad (3.19)$$

**Leacock & Chodorow Measure** [54] assumes concepts to be organized in an *is-a* hierarchy and defines the similarity of concepts as their proximity in such hierarchy. Given two concepts *a* and *b*, the function *sl* that returns in the shortest path between two concepts in the hierarchy tree, and *D* the maximum depth of the tree, the Leacock & Chodorow Measure can be calculated using Equation 3.20.

$$LC(a,b) = max[-log(\frac{sl(a,b)}{2D})] \quad [69] \quad (3.20)$$

**Wu & Palmer Measure** [100] consider a very similar hierarchy to the Leacock & Chodorow Measure. It considers that similarity is based not only on the depth of the concepts in the hierarchy tree, but also in the depth of their LCS. Given two concepts *a* and *b*, *root* the root of the hierarchy tree, and the function *sl* that returns in the shortest path between two concepts in the hierarchy tree, the Wu & Palmer Measure Measure can be calculated using Equation 3.21.

$$WP(a,b) = \frac{2 \times sl(lcs(a,b),root)}{sl(a,lcs(a,b))+sl(b,lcs(a,b))+2 \times sl(lcs(a,b),root)} \quad [100] \quad (3.21)$$

**Hirst & St. Onge Measure** [45] has a different point of view from many of the previous measures, in the sense that it does not consider words distributed in an *is-a* hierarchy. Instead, it considers a broader range of relations. Specifically, three of them are defined: upward relations, connecting a more specific concept to a more general one; downward relations, connecting a more general concept to a more specific one; horizontal relations, connecting concepts of similar specificity. Examples of such relations could be *is-a*, *contains*, and *antonyms*, respectively [69].

**Lesk Algorithm** [55] assumes that concepts are more related to one another if they have more words in common in their definition. The algorithm itself compares a word's definition to the definition of the words that surrounds them in a sentence. The idea behind it is that words that make up a single sentence should have some similarity, and overall sentence topics can be extracted by overlapping the definitions of its constituting words. Banerjee and Pedersen later expanded this algorithm [14] so that it also overlaps a word's definition with the definition of words close to it in WordNet's concept hierarchy.

Table 3.2: Similarity measures utilized in news processing systems.

| Measure | Used in |
|---|---|
| LDA | Bechmann et al. [16]<br>Chuang et al. [27]<br>Paul et al. [70]<br>Paul et al. [71]<br>Sjøvaag et al. [87]<br>Tabara et al. [90]<br>Zhao et al. [103] |
| Jaccard Similarity | Burkepile et al. [23]<br>Davis et al. [31]<br>Pons-Porrata et al. [76]<br>Sakhapara et al. [83] |
| Cosine Similarity | Devezas et al. [32]<br>Pons-Porrata et al. [76]<br>Sakhapara et al. [83] |
| Dice's Coefficient | Davis et al. [31] |
| LSA | Tabara et al. [90] |

## 3.4 Conclusions

Having surveyed multiple techniques for text similarity calculation, we can see a great variety in the requirements for said techniques and often multiple techniques that can be applied in similar scenarios. Moreover, we can see that the vast majority of techniques were not used in the systems described in Chapter 2. A cross-reference between the news processing systems and the measures described in this chapter can be seen in Table 3.2. This table only includes the measures that are used at least once.

This is particularly relevant since we have previously identified the prevalence of text similarity techniques and topical models in the comparison of news articles. We can now affirm that there is a large number of methodologies that are yet to be tested in news article comparison. This, coupled with the aforementioned lack of studies that directly compare techniques in news article comparison, lets us believe that evaluating the effectiveness of different text similarity measures in news comparison systems is not only mostly uncharted territory, as it could be an important asset in improving the performance of said systems.

# Chapter 4

# News Feed Similarity Computation

This chapter will detail the steps followed in the proposal and comparison of different news feed similarity techniques. We will first delve into the details of the implemented techniques, including the dataset used in the first set of experiments and the obtained results. Then, we will test the same techniques against a set of artificially generated feeds, attempting to simulate different edge cases. The techniques will then be tested against a selected sample of real news feeds, selected for their ease of human readability. These results will be compared with data obtained from a survey to test their validity. Finally, we will compare the obtained results and derive conclusions on the effectiveness of the studied techniques.

## 4.1 Similarity Computation

We will now delve into the calculation of the similarity of news feeds. First, we will detail the data that was selected for testing. We will discuss the preprocessing that was used to prepare the collected web feeds. We will then overview the measures and strategies that were used to compare said feeds. Then, we will propose a method for comparing the results of the different used techniques. Finally, we will discuss the overall conclusions of the experiment.

### 4.1.1 Dataset Characterization

The data used to test the different news feed comparison techniques comes from a preexisting database, collected through the MediaViz platform [33]. It contains over 10 million articles from 68 different news sources, spanning from December 2014 to February 2021, resulting from a continuous crawling of web feeds. A data model of the dataset can be seen in Figure 4.1.

For our experiments, we are mainly interested in the information contained in the articles themselves: the fields *title* and *summary* from the table *articles*. Moreover, and in order to be able to compare different news sources, it is necessary to analyze the origin of said articles. The *articles* table is directly connected to the *feeds* table, which represents the web feeds from which

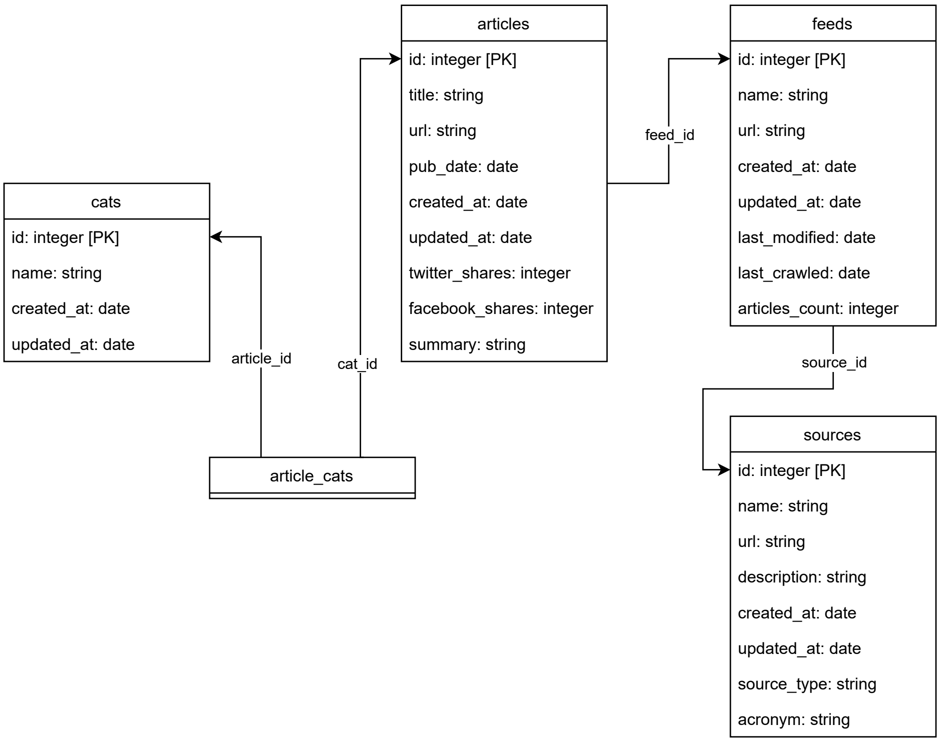


Figure 4.1: Data model of the MediaViz news item database.

articles are extracted. This table is then related to the *sources* table, whose rows each represent a different news source.

The distinction between feeds and sources is important, since each source can publish its articles in several different feeds. In the conducted experiments, we decided to aggregate all the feeds published by the same source. The reason for this is twofold. First, there is already a significant number of different sources, and reducing the number of individual news aggregates allows us to conduct fewer similarity computations while still taking into consideration the full spectrum of news sources. Secondly, the separation of news among different feeds by a single source is usually related to the news content, having each feed relating to a certain category, such as politics, arts, sports, and so on. Comparing such feeds among themselves would not be very productive, as it is not to be expected to be a significant value of similarity between them.

Another important piece of information present in the dataset is in the field *source_type* of the table *sources*. Each source is classified into one of five types: national, international, blogs, research, and archive. Since the objective of one of the case studies is to study the Portuguese media landscape, we have opted to only consider articles from sources tagged as national. As a sample to use in the experiments, we selected news articles that followed the specified requirement published in five consecutive days, from December 8th, 2014 to December 12th, 2014. Details about the number of feeds and articles present in the collected sample can be seen in Table 4.1.

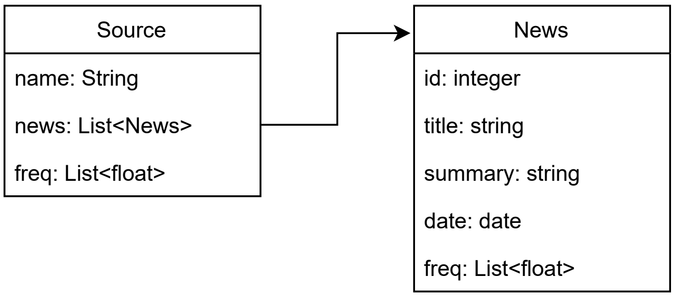


Figure 4.2: Data model of the news data structure.

The majority of the news sources have in their main body of text - the field *summary* - a short article lead, that does not extend beyond five or six lines of text. This is not true for the items published by the source *iOnline*, which contain full articles. For that reason, and to avoid discrepancies in the results due to the unmatching text length, we decided to exclude *iOnline* from the experiment, alongside the sources that had no articles in the selected time span. We also removed the sources that have no articles published in the selected time frame.

### 4.1.2 Preprocessing

The collected data detailed in the previous section was internally organized in a data structure. At the top level, there was an entry for each news source, since the purpose of this experiment was to compare different sources. Then, each source contained a list of news items, published by that source. Each news item contained information not only about its content, but also its publication. An overview of this data structure can be seen in Figure 4.2.

The objective of this preprocessing is to build the *freq* attributes of both classes. These vectors represent the frequency of terms in each document, in an analogous way to the vectorial representation described in Chapter 3. The *freq* attribute in the *Source* class is the result of the sum of the frequency vectors of the corresponding news items.

In order to calculate the frequency vector for a specific news item, first, it is necessary to define what constitutes a news item. For this experiment, we decided that the usable textual information of a news item consists of the concatenation of the *title* and *summary* attributes. In an analogous way to many systems discussed in Chapter 2, stopwords are then removed, and the remaining text is then tokenized, having each token correspond to a term, and the tokens then stemmed, in order to improve the term consistency. In order to achieve this, the NLTK Python library [5] and its implementation of the RSLP Stemmer [67] were utilized.

Table 4.1: Details of selected news article sources.

| | | **Nº of Articles per Day** | | | | |
|---|---|---|---|---|---|---|
| **Source** | **Nº of Feeds** | **8th** | **9th** | **10th** | **11th** | **12th** |
| A Bola | 1 | 98 | 124 | 119 | 112 | 116 |
| Antena 1 | 1 | 115 | 194 | 153 | 154 | 156 |
| Correio da Manhã | 1 | 140 | 201 | 137 | 133 | 146 |
| Diário de Notícias | 12 | 51 | 94 | 101 | 70 | 89 |
| Diário Digital | 11 | 0 | 0 | 0 | 0 | 0 |
| Económico | 7 | 20 | 109 | 98 | 99 | 91 |
| Expresso | 6 | 29 | 72 | 28 | 38 | 24 |
| iOnline | 1 | 10 | 25 | 18 | 15 | 11 |
| Jornal de Negócios | 1 | 68 | 125 | 96 | 93 | 89 |
| Jornal de Notícias | 20 | 39 | 72 | 61 | 69 | 65 |
| MaisFutebol | 1 | 96 | 93 | 79 | 72 | 72 |
| Observador | 1 | 67 | 87 | 54 | 52 | 71 |
| O Jogo | 8 | 87 | 73 | 85 | 79 | 97 |
| P3 | 4 | 5 | 14 | 17 | 17 | 15 |
| Porto Canal | 1 | 34 | 45 | 77 | 76 | 82 |
| Público | 1 | 57 | 92 | 81 | 66 | 89 |
| Record | 1 | 173 | 169 | 156 | 145 | 161 |
| Renascença | 12 | 61 | 119 | 113 | 98 | 119 |
| RTP | 20 | 40 | 56 | 39 | 61 | 51 |
| Sábado | 1 | 2 | 155 | 105 | 117 | 82 |
| SAPO Desporto | 1 | 0 | 0 | 0 | 0 | 0 |
| SAPO Notícias | 1 | 0 | 0 | 0 | 0 | 0 |
| SIC Notícias | 1 | 0 | 0 | 0 | 0 | 0 |
| Sol | 1 | 37 | 67 | 61 | 79 | 73 |
| TSF | 8 | 43 | 39 | 53 | 56 | 49 |
| TVI24 | 1 | 54 | 118 | 107 | 103 | 109 |
| Visão | 1 | 80 | 119 | 75 | 92 | 78 |

Table 4.2: Example of news item preprocessing.

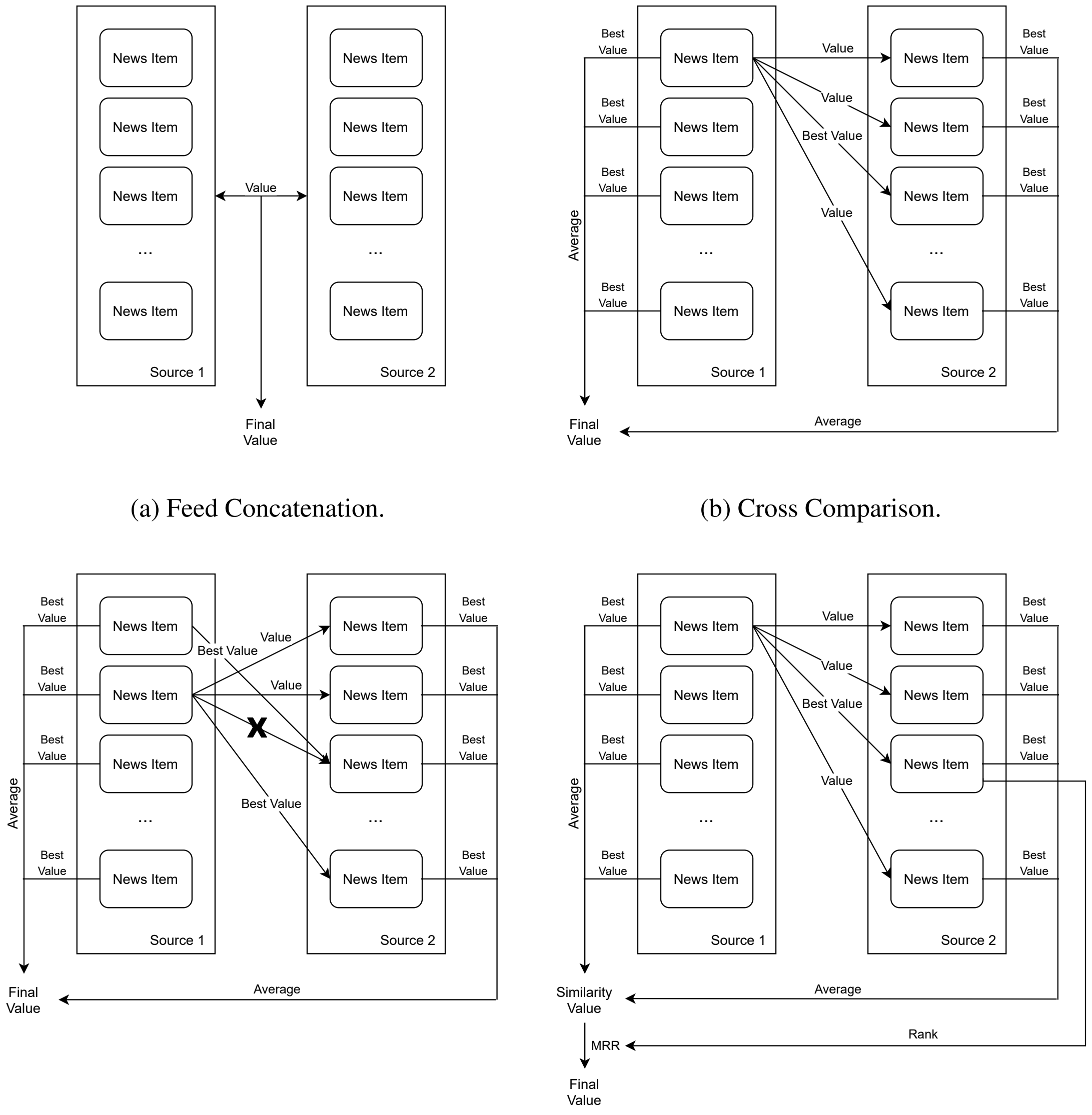
| | |
|---|---|
| **News Item Text** | O papel dos artistas nos Jogos da Fome do século XXI? O de resistente. Richard Florida como vendedor de banha da cobra e os artistas como catalisadores de energias anárquicas, focos de resistência – os Jogos da Fome da arte e do poder vistos pela conceituada artista plástica e académica norte-americana Martha Rosler |
| **Stemmed Tokens** | [‘o’, ‘papel’, ‘artist’, ‘jog’, ‘fom’, ‘sécul’, ‘xxi’, ‘o’, ‘resistente’, ‘richard’, ‘flor’, ‘vend’, ‘banh’, ‘cobr’, ‘artist’, ‘catalis’, ‘energ’, ‘anárquicas’, ‘foc’, ‘resist’, ‘jog’, ‘fom’, ‘art’, ‘pod’, ‘vist’, ‘conceitu’, ‘artist’, ‘plás’, ‘académ’, ‘norte’, ‘americ’, ‘marth’, ‘rosl’] |
| **Vector Representation** | [‘o’: 2, ‘papel’: 1, ‘artist’: 3, ‘jog’: 2, ‘fom’: 2, ‘sécul’: 1, ‘xxi’: 1, ‘resistente’: 1, ‘richard’: 1, ‘flor’: 1, ‘vend’: 1, ‘banh’: 1, ‘cobr’: 1, ‘catalis’: 1, ‘energ’: 1, ‘anárquicas’: 1, ‘foc’: 1, ‘resist’: 1, ‘art’: 1, ‘pod’: 1, ‘vist’: 1, ‘conceitu’: 1, ‘plás’: 1, ‘académ’: 1, ‘norte’: 1, ‘americ’: 1, ‘marth’: 1, ‘rosl’: 1] |

The tokens obtained from all the existing news articles were used to build a term dictionary. Each term in this dictionary corresponds to a dimension of the vector space that was used to represent the news items. The *freq* attribute was then calculated for each news item, having each vector dimension correspond to the frequency of the corresponding term in the article’s usable text. An example of the different stages of this process can be seen in Table 4.2. Having calculated the *freq* attribute for all news items from a source, the *freq* attribute of that source is then calculated by performing a vectorial sum of all the *freq* vectors that correspond to news items of that source.

### 4.1.3 Measures and Strategies

After the preparation of the data for analysis, it was then necessary to define the strategies to compare the different web feeds. Six text similarity measures were selected: **Matching Coefficient**, **Euclidean Distance**, **Manhattan Distance**, **Cosine Similarity**, **Dice’s Coefficient**, and **Jaccard Similarity**. The details of these similarity measures can be found in Chapter 3. Additionally, and following the conclusions of Sakhapara et al. [83], where an average of multiple measures was used to an increase of performance, a seventh measure was proposed, consisting of the arithmetic average of Cosine Similarity, Dice’s Coefficient, and Jaccard Similarity. We will refer to that measure as **Combined Measure**.

Additionally, four different strategies were utilized for comparing the feeds in a macro level, applying the selected algorithms, which can be seen in Figure 4.3. First, and akin to most of the systems discussed in Chapter 2, we opted to ignore the individual structure of the news items, as well as the structure of the feed itself, and consider the web feeds as units and compare them directly using the selected measures. This utilized the term frequency that was calculated for each source instead of the news articles’ individual term frequencies. We will refer to this strategy



(a) Feed Concatenation.

(b) Cross Comparison.

(c) Cross Comparison with Elimination.

(d) Cross Comparison with MRR.

Figure 4.3: Diagrams of the similarity comparison strategies.

as **Feed Concatenation** (FC). A diagram of this strategy can be seen in Figure 4.3a. Its general implementation can also be seen in Algorithm 1.

**Algorithm 1:** Feed Concatenation

```
FC $(source1, source2)$
    inputs : Two processed feeds $source1$ and $source2$, with global feed word frequency
    $sim_{val} \leftarrow sim\_measure(source1.freq, source2.freq)$;
    return $sim_{val}$;
```

The second strategy tries to emulate the one that was used by Sakhapara et al. [83], in which they manually pair news items (that they consider similar) together and apply the algorithms to those pairs. Our strategy attempts to simulate this process without the need to manually sort and pair large amounts of data. To compare two news sources, each news item of one of the sources is compared individually with all the news items of the other source using the selected measures, choosing as its final similarity value the one obtained when comparing to the news item it found most similar. This assumes that, if there is a news article in the other source that refers to the same event as the article in the first source, it will be the one that produces the best similarity value. If there is none, it is expected that the best-obtained value is reasonably low. The final similarity value between both sources is the arithmetic average of the values selected by all the articles. The algorithm is then executed again, starting with the second feed, averaging both results, so that the algorithm is symmetrical. We will refer to this strategy as **Cross Comparison** (CC). A diagram of this strategy can be seen in Figure 4.3b. Its general implementation can also be seen in Algorithm 2.

**Algorithm 2:** Cross Comparison

```
CC $(source1, source2)$
    inputs : Two processed feeds $source1$ and $source2$, with news word frequency
    $sim\_val \leftarrow (CalcSim(source1, source2) + CalcSim(source2, source1))/2$;
    return $sim_{val}$;
CalcSim $(source1, source2)$
    inputs : Two processed feeds, with news word frequency $source1$ and $source2$
    $total_{sim} \leftarrow 0$;
    $total_n \leftarrow 0$;
    foreach $n1 \in source1.news$ do
        $best \leftarrow 0$;
        foreach $n2 \in source2.news$ do
            $sim_{val} \leftarrow sim\_measure(n1.freq, n2.freq)$;
            if $sim_{val} > best$ then
                $best \leftarrow sim_{val}$;
        $total_{sim} \leftarrow total_{sim} + best$;
        $total_n \leftarrow total_n + 1$;
    return $total_{sim}/total_n$;
```

The third strategy is a variation of Cross Comparison. The main difference lies in the fact that, in CC, multiple articles from one of the sources can select the same article in the other source as their most similar counterpart. This means that, in an extreme scenario, all articles from one source could select the same article in the other source, completely disregarding the remaining articles. In order to ensure better coverage of all the articles from both feeds, the third strategy blocks articles that were previously selected as having the most similarity from being selected again, removing them from the cross comparison. In order to avoid blocking good news parings in favor of worse ones, it is given priority to the matches that have higher similarity values. We will refer to this strategy as **Cross Comparison with Elimination** (CCE). A diagram of this strategy can be seen in Figure 4.3c. Its general implementation can also be seen in Algorithm 3.

**Algorithm 3:** Cross Comparison with Elimination

```
**CCE** $(source1, source2)$
    **inputs :** Two processed feeds, with news word frequency $source1$ and $source2$
    $sim\_val \leftarrow (CalcSim(source1, source2) + CalcSim(source2, source1))/2$;
    **return** $sim_{val}$;
**CalcSim** $(source1, source2)$
    **inputs :** Two processed feeds, with news word frequency $source1$ and $source2$
    $total_{sim} \leftarrow 0$;
    $total_n \leftarrow 0$;
    $values \leftarrow []$;
    $values_{n1} \leftarrow -1$;
    **foreach** $n1 \in source1.news$ **do**
        $values.append([])$;
        $value_{n1} \leftarrow value_{n1} + 1$;
        $values_{n2} \leftarrow -1$;
        **foreach** $n2 \in source2.news$ **do**
            $value_{n2} \leftarrow value_{n2} + 1$;
            $values[value_{n1}].append([sim\_measure(n1.freq, n2.freq), value_{n2}])$
    **foreach** $v \in values$ **do**
        $v.sort()$;
    **while** $values.length > 0$ **do**
        **if** $values[0].length = 0$ **then**
            $break$;
        $values.sort()$;
        $values.pop()$;
        $total \leftarrow total + values[0][0][0]$;
        $total_n \leftarrow total_n + 1$;
    **return** $total_{sim}/total_n$;
```

The final strategy is again a variation of Cross Comparison. This time, the focus is on the order and positioning of the articles inside their respective feeds. For a user, it is not equal for two similar articles from different feeds to be published very close to each other and to be published with other articles in between. The similarity of the feeds is much more apparent if the similar articles are close to each other. In order to simulate this behaviour, this strategy introduces a coefficient that decreases the similarity value of the feeds with the increase of the relative distance of similar articles. Is calculation was inspired by Mean Reciprocal Rank [95], a measure utilized in rank evaluation when ranks are ordered by probability of correction. This measure is given by Equation 4.1:

$$MRR = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{rank_i}$$ [95] (4.1)

In our scenario, $n$ is the number of articles and $rank_i$ is the relative distance between article $i$ and its most similar counterpart. Meaning that, for example, if both articles occupy the second position in their respective feeds, $rank_i$ for this article pair will be 1. If one of the articles is in the second position and the other is in the fifth position, $rank_i$ for this article pair will then be 4. We will refer to this strategy as **Cross Comparison with Mean Reciprocal Rank** (CC-MRR). A diagram of this strategy can be seen in Figure 4.3d. Its general implementation can also be seen in Algorithm 4.

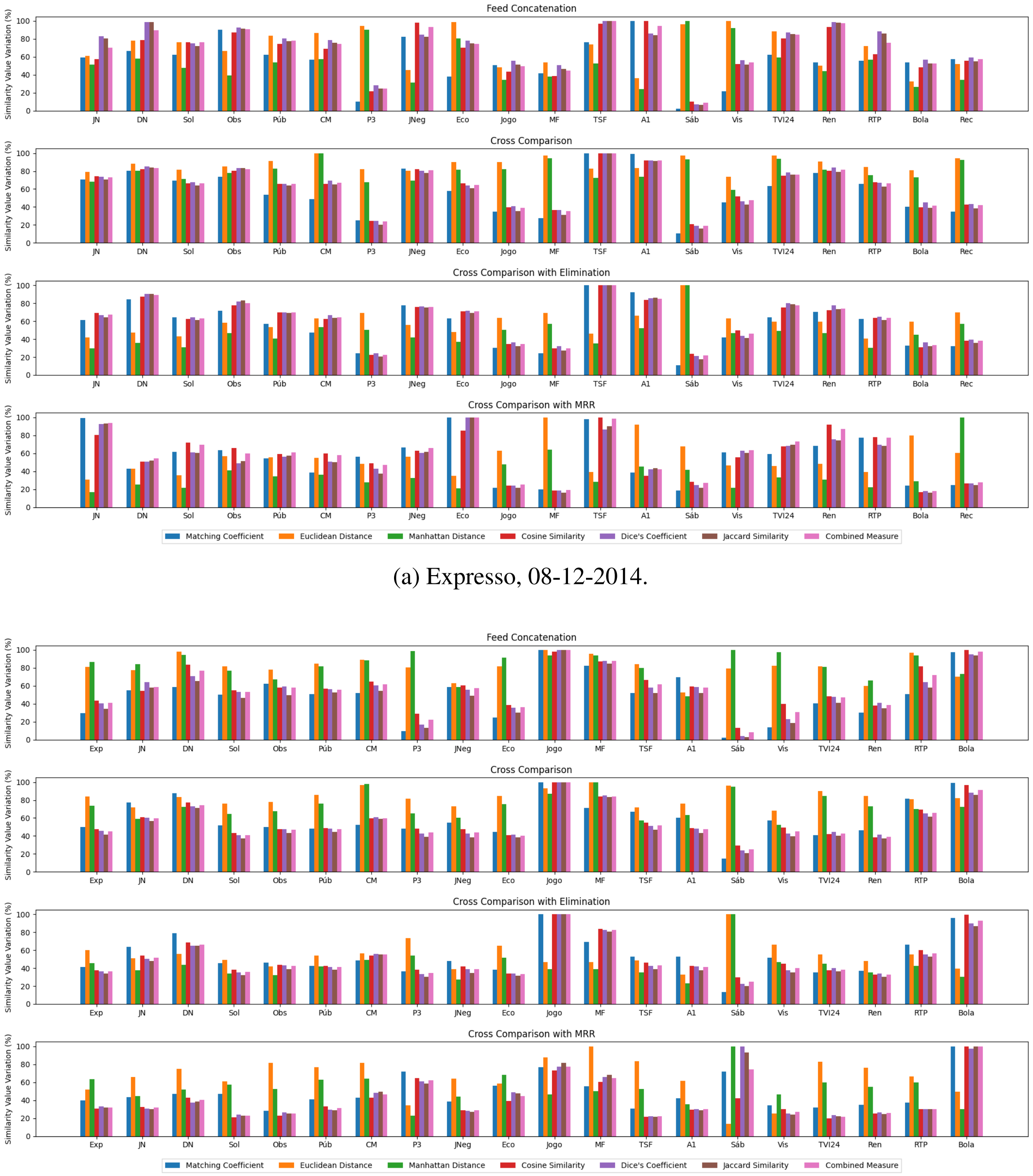
**Algorithm 4:** Cross Comparison with MRR

```
CC-MRR ($source1, source2$)
    inputs : Two processed feeds, with news word frequency $source1$ and $source2$
    $sim\_val \leftarrow (CalcSim(source1, source2) + CalcSim(source2, source1))/2$;
    return $sim_{val}$;
CalcSim ($source1, source2$)
    inputs : Two processed feeds, with news word frequency $source1$ and $source2$
    $total_{sim} \leftarrow 0$;
    $total_n \leftarrow 0$;
    $total_d \leftarrow 0$;
    $index_i \leftarrow 0$;
    foreach $n1 \in source1.news$ do
        $best \leftarrow 0$;
        $best_d \leftarrow 0$;
        $index_i \leftarrow index_i + 1$;
        $index_j \leftarrow 0$;
        foreach $n2 \in source2.news$ do
            $index_j \leftarrow index_j + 1$;
            $sim_{val} \leftarrow sim\_measure(n1.freq, n2.freq)$;
            if $sim_{val} > best$ then
                $best \leftarrow sim_{val}$;
                $best_d \leftarrow |index_i - index_j|$;
        $total_{sim} \leftarrow total_{sim} + best$;
        $total_n \leftarrow total_n + 1$;
    return $(total_{sim}/total_n) * (total_d/total_n)$;
```

These strategies were applied with the seven similarity measures on the twenty-two news sources, individually in each of the five selected days. This means that news articles published on a certain day could only be compared with other articles also published on that same day. This was done in order to maximize the likelihood of finding similar news items that cover the same events, since such news articles are likely to be produced on the same day.

### 4.1.4 Results

In order to compare the different measures, the values for each one were converted to percentage points, having the greatest obtained value equivalent to 100%. The reason for this comes from



(a) Expresso, 08-12-2014.

(b) Record, 08-12-2014.

Figure 4.4: Results of news feed similarity computation.

the fact that different strategies produce similarity values that are in different orders of magnitude, making result comparison difficult. Furthermore, both Euclidean Distance and Manhattan Distance cannot be directly converted to percentage points, since the greater the value is, the less similar the pair of feeds is. Therefore, the results for Euclidean Distance and Manhattan Distance were inverted before being converted to percentage points. As an example, we will discuss the results obtained from the comparison of the feeds obtained from *Expresso* and *Record* with all other sources. These results can be seen in Figures 4.4a and 4.4b, respectively.

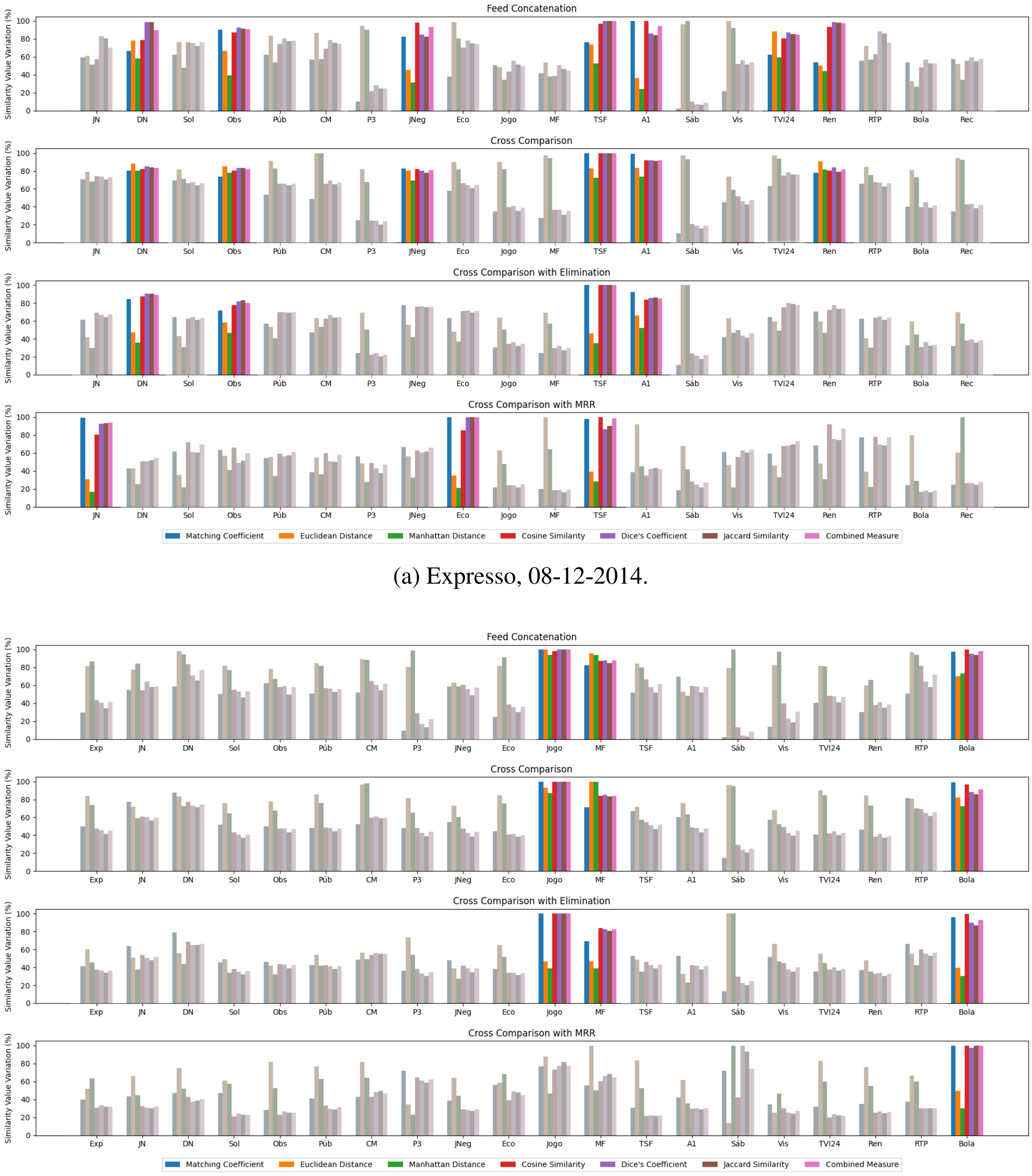
Looking at the overall results and the relative results of the text similarity measures, we can see that both ED and MD have higher results variation when compared to the remaining measures, showing consistently high values. This can be attributed to the inversion of the results, followed by their conversion to percentage. We will disregard them both in the discussion of the remaining results, so that they do not influence our conclusions regarding the comparison strategies. The remaining five measures obtained very similar results among them. Overall, we can say that they could mostly have been used interchangeably and obtained similar results.

In regards to *Expresso*, which can be considered a generalist news source, that covers a variety of different topics, it would be expected that it would show high similarity with other generalist news sources, in which the majority of the available sources can be included. On the other hand, *Record* is a source dedicated to sports news. We would expect it to be similar to other sport news sources, which, in our experiment, are *O Jogo* (Jogo), *MaisFutebol* (MF), and *A Bola* (Bola).

In Figures 4.5a and 4.5b, we can see highlighted the news sources whose majority of obtained similarity values were 80% or higher (excluding ED and MD). In both scenarios, we can see that FC, CC, and CCE obtained very similar results, given the higher similarity values to *Diário de Notícias* (DN), *Observador* (Obs), *TSF* (TSF), and *Antena 1* (A1), when compared to *Expresso*, and to *O Jogo* (Jogo), *MaisFutebol* (MF), and *A Bola* (Bola), when compared to *Record*. This mostly followed our expectations of similarity. The same is not true for CC-MRR, where the sources considered similar are somewhat different.

It is not easy to differentiate between FC, CC, and CCE in the *Record*, since the gap between the selected news sources and the rest is considerable. This is to be expected, since both the main source as well as the selected sources are all sports-oriented. With *Expresso*, however, there were many non-selected news sources that were close to the threshold, as well as some sources that were selected only by some of the algorithms. Generally, we can see that FC was less discriminating, having the values of more sources closer to the threshold. CCE, on the other hand, has a somewhat greater gap between the top scoring sources and the rest. CC showed results between the other two algorithms. It is however important to note that these are, overall, small differences and that all three algorithms performed similarly on both scenarios.

The discrepancies between the results obtained with CC-MRR are, however, expected. MRR reduced the similarity values exponentially with the relative distance between paired news items and it is normal that the results differ significantly from methods like FC, CC, and CCE, which completely disregard the article order inside the news feeds. This is even more prominent in our specific experiment, since we are using daily feeds, some of each have over 100 news items. This seems to indicate that, more than a tool for direct similarity calculation between feeds, CC-MRR can possibly be more useful in extracting information about the relative position of news articles. Two feeds that are considered similar by FC, CC, or CCE but have a lower CC-MRR either have similar articles published farther apart in the feed or one of the sources is significantly more prolific than the other, having published many more news items than the other, and therefore spacing the paired news items apart. For example, both FC, CC, and CCE consider *Record* very similar to both



(a) Expresso, 08-12-2014.

(b) Record, 08-12-2014.

Figure 4.5: Selection of similar news feeds.

*O Jogo* and *A Bola*, but CC-MRR is significantly higher for *A Bola*. As we can see in Table 4.1, both sources have a similar number of news items to one another (87 and 98, respectively), from which we can infer that *O Jogo* probably has its articles published farther apart in time.

In conclusion, we can say that we have observed evidences that indicate that FC, CC, and CCE can be suitable for feed similarity calculation, especially when using one of the more stable text similarity measures, such as DC or JS. Moreover, CC-MRR can be used in conjunction with the feed length and one of the other strategies to further infer about the feed's relative news item positioning.

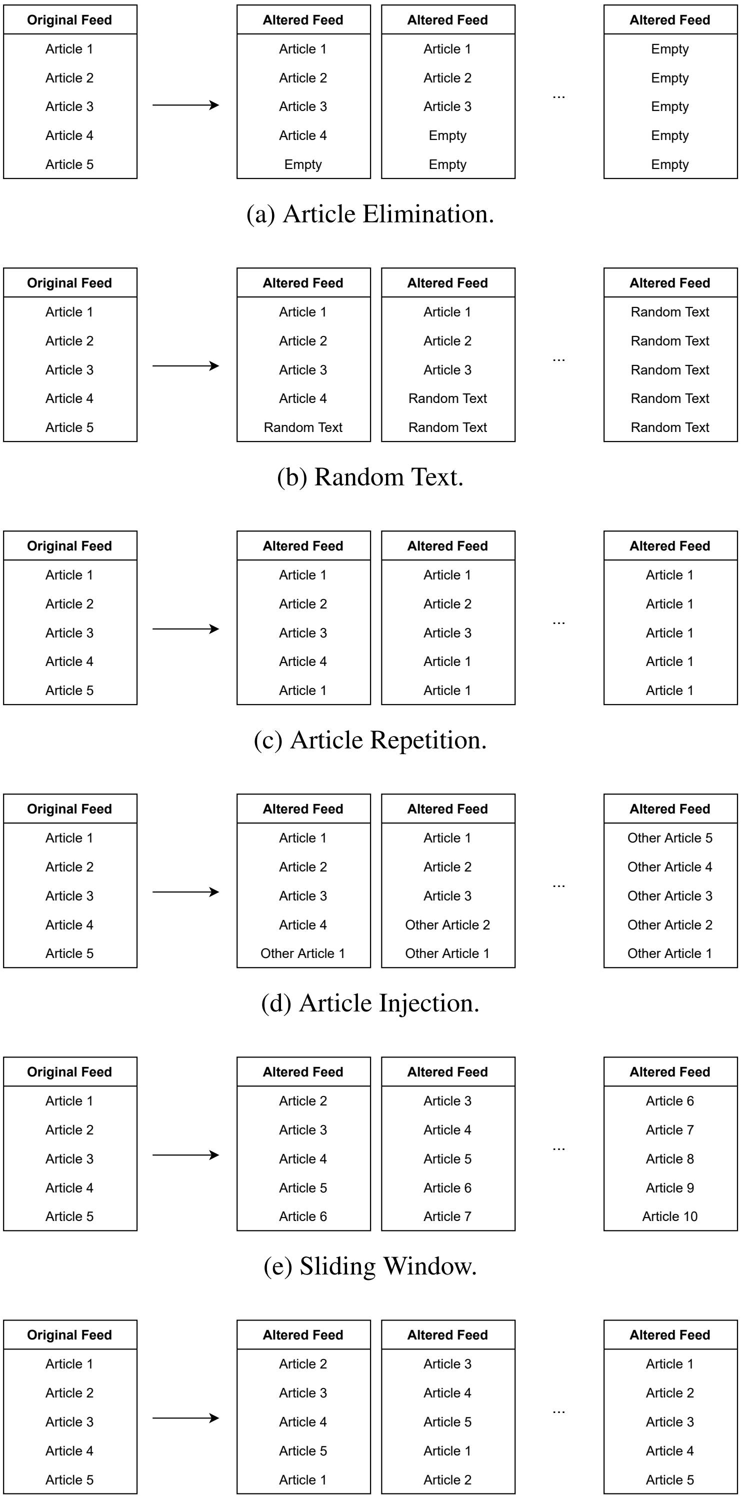
## 4.2 Synthetic Feed Generation

In order to better understand the behaviour of the utilized measures, we have decided to generate artificial news feeds and apply the similarity calculation strategies to them. The purpose of this experiment is to use the algorithms in scenarios where we can expect a certain result and, from that, discern conclusions regarding the effectiveness of the measures.

### 4.2.1 Generation Strategies

The different proposed scenarios are derivations of a reference feed extracted from the data detailed in Section 4.1.1. The used feed consists of the first five news articles published after 8 AM of December 8th, 2014 by the source *Expresso*. From this feed, six different sets of artificial feeds were generated. The overview of this process can be seen in Figure 4.6. Each of the sets of artificial feeds consisted of five items. Each new altered feed is generated by performing some operation on the previous one. Therefore, each newly generated feed is more different when compared to the original one.

In **Article Elimination** (AE), the news articles are progressively being replaced with articles whose title and summary consist of empty text strings, as seen in Figure 4.6a. In **Random Text** (RT), the newly inserted articles' title and summary consist of random words in Portuguese, having the same number of words as the corresponding removed article, as seen in Figure 4.6b. In both these generation strategies, we expect the similarity values to stedily decline with the tampering of the news feeds, since the articles are not being replaced with coherent articles. Moreover, and since the feeds are being represented in a vectorial space, we will try to understand if there is a diference between simply removing values from the vector (AE) and replacing some values with others that are expected to not overlap with existing ones (RT). In **Article Repetition** (AR), the first article of the feed is used to replace the removed articles. In this particular case, the last generated feed is redundant, since it replaces the first article with itself, as seen in Figure 4.6c. With this strategy we will try to understand the behaviour of each compaison strategy when faced with repeated information. In **Article Injection** (AI), the articles inserted in the fabricated feeds come from other feeds present in the dataset, as seen in Figure 4.6d. In this experiment, we used articles from the source *RTP*, collected on the same date. In **Sliding Window** (SW), we consider not only the five articles present in the original feed, but also the five articles that were published by the same source immediately afterwards, as seen in Figure 4.6e. Considering this order of articles, the generated feeds are created by progressively removing the oldest article still present in the feed and replacing it with the next one in the order. That means that the first generated feed contains articles two through six, and the next feed the articles three through seven, and so on. With the last two strategies, we try to emulate real world feeds by replacing the articles with different but coherent articles. We expect that the drop of similarity is therefore less pronounced



(a) Article Elimination.

(b) Random Text.

(c) Article Repetition.

(d) Article Injection.

(e) Sliding Window.

(f) Article Shifting.

Figure 4.6: Diagrams of artificial feed generation.

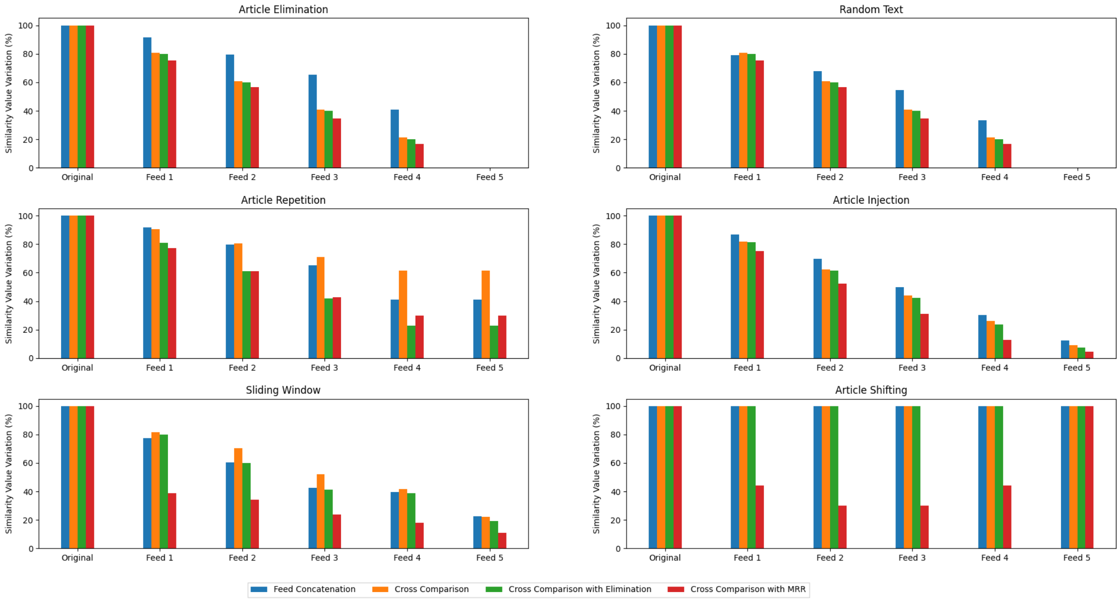


Figure 4.7: Results of fabricated feed similarity computation.

when compared to AE or RT. Finally, in **Article Shifting** (AS), we use a technique similar to SW, but instead of inserting new articles at the end of the feed, it uses the one that was just removed, as seen in Figure 4.6f. That means that the generated feeds do not differ from the original one in content, but only in the article order. The last generated feed is, similar to what happens in AR, redundant, since it repeats the original article order. This last scenario is expected to affect only CC-MRR, since no other strategy cares about article order.

### 4.2.2 Results

Each generated feed was compared to its corresponding original one using the strategies described in Section 4.1.3: Feed Concatenation, Cross Comparison, Cross Comparison with Elimination, and Cross Comparison with MRR. Given the results discussed in Section 4.1.4, in which we concluded that the text similarity measures were mostly interchangeable, we decided to only use Dice's Coefficient in this experiment. In all groups of generated feeds, the original feed was also compared to itself, as a way to ground for the comparisons, since it expected that all comparisons of a feed with itself are evaluated with a similarity value of 1 by all strategies.

Before delving into the specific scenarios, we can see that, overall, the results are consistent with the ones obtained in Section 4.1.4. FC, CC, and CCE are very close to each other in most scenarios, with CC-MRR having more variable results, which is to be expected. It was also expected that the results in these scenarios are more consistent, since these experiments were conducted in a very controlled environment.

Both **AE** and **RT** are similar in nature: articles are replaced with empty strings and random words, respectively. It is not likely that random words from a dictionary can form a text string that text similarity algorithms consider much more similar to a news item than the value obtained from comparing the same article to an empty string. Therefore, we expect that the strategies obtain similar values in both artificial feed sets. Moreover, we expect that the similarity value is somewhat equal to the ratio of original articles still present in the generated feed - meaning, 0.8 for the first generated feed, 0.6 for the second, and so on. This is consistent with the results of all the strategies except for FC. This is probably derived from the fact that CC, CCE, and CC-MRR are better at isolating the adulterated portions of the feed, since they select only the best articles for comparison, being therefore able to mostly ignore the fabricated articles. This is not true for FC, which has the entire feed in consideration at all times. Even so, the values of FC still decrease the more adulterated is the feed and are not so that separated from the expected result to be considered inaccurate.

Another pair of similar feed sets is **AI** and **SW**. In both cases, the news items are replaced with different ones in each feed generation. We can also expect similar results in both scenarios, once again somewhat close to the ratio of original articles remaining in the generated feed. However, we can expect the values to deviate a little more from these values than AE and RT did, since the articles that are being inserted into the feeds are real news items and therefore more likely to contain common English words, which would probably result in greater overlap with the original news items, causing the overall similarity values to rise. The results obtained using FC, CC, and CCE overall fit in what was expected. The ones calculated using CC-MRR, however, have a significant similarity drop in SW when compared to AI. This is likely a product of the way the feeds were generated. In fact, it is intended behaviour for CC-MRR, since the articles that remain from the original feed are further apart from their original positions with each feed generated, something that does not occur in AI. The observed drop is consistent with the CC-MRR results in AS, which we will discuss later on.

**AR** is a very specific scenario. Unlike the previous ones, the information that is artificially inserted in the generated feeds is not random, but is in fact data that is guaranteed to be similar to part of the original feed. This is the scenario where FC, CC, and CCE have the most disparate results. CCE performs as it would be expected, since it remains very close to the ratio of original articles. CC is at the opposite end of the spectrum, showing high similarity values for all the generated feeds. FC stands between the other two strategies. It is not immediately obvious which is performing better, as it is not obvious what makes two feeds similar. It is largely open to interpretation whether one feed with five news items similar to a singular item on the other feed should be considered similar, dissimilar, or something in between.

Finally, **AS** is mostly a scenario that is used to test the range of values that can be obtained by using CC-MRR. As expected, FC, CC, and CCE consider all the feeds in this group completely similar to each other. Also, as expected, CC-MRR shows a significant drop, that is maximum in Feeds 2 and 3, where each news item is the farthest from their original position. The obtained values are consistent with the drop observed in the results from SW, as well as the ones obtained in Section 4.1.4. These results from CC-MRR can be used as a benchmark on how to use this strategy to evaluate news item positioning in feeds.

In conclusion, we can further reinforce the usefulness of FC, CC, and CCE as strategies to compare news item feeds. They all showed results that matched the expectations in all of the studied scenarios. We also gained a deeper understanding of the situations where each measure behaves differently from the others, when facing very particular sets of data. Finally, we were able to better comprehend the behaviour of CC-MRR regarding news feed item positioning and structure.

## 4.3 User Study

As a final test to the feed similarity strategies developed so far, we decided to compare their results with the ones obtained through an user study. For that, a simple questionnaire was developed where participants were asked to evaluate the similarity between two small feeds. These feeds were then analyzed by the similarity strategies in order to both understand which ones obtained results closer to the ones from the survey, as well as get some insights on what the general public understands by “feed similarity”.

### 4.3.1 Experiment Design

For this experiment, five pairs of feeds were prepared. Each feed contains five news items, all extracted from the same source and published on January 16th, 2021. The news items were hand-picked in order to create different and contrasting scenarios. Inside each feed, the news items were ordered chronologically according to their publishing time. A complete list of the pairs of feeds used in the study can be found, as written in their original Portuguese, in Appendix A. A summary of the pairs of feeds utilized, including their design intention and our expected user rating, can be found in Table 4.3.

The study had as its targeted audience the Portuguese-speaking students of Faculdade de Engenharia da Universidade do Porto. Each participant was presented with a series of instructions regarding the survey. Mainly, it was explained to them: what is a news feed; that, in this survey, the purpose is to compare two feeds, considering each feed as a unit; and that similar feeds are the ones that contain news items that cover the same events or discuss the same themes. More details regarding these instructions, as written in their original Portuguese, as well as the user interface of the survey, can be found in Appendix B. After the initial instructions, the users were presented with one of the pairs of feeds and asked to rate their similarity between 0 and 10, where 0 is the value for two completely dissimilar feeds and 10 for two completely similar ones. After submitting their answer, the participants were given the choice of analyzing more pairs of feeds, if they so desired.

### 4.3.2 Survey Results

The survey had a total of 142 valid responses. The distribution of the results can be seen in Figure 4.8. Looking at the overall results, we can conclude that they were skewed towards the

Table 4.3: News feed pairs summary.

| Pair | Source 1 | Source 2 | Content Description | Expected Rating |
|---|---|---|---|---|
| Pair 1 | Expresso | Jornal de Notícias | All five news items in one feed have a similar counterpart in the other feed | 9-10 |
| Pair 2 | Expresso | Observador | All five news items in one feed have a somewhat related counterpart in the other feed | 6-8 |
| Pair 3 | Expresso | Público | All items in the second feed are somewhat related to one of three items in the first feed. The other two in the first feed have no similar counterpart | 4-6 |
| Pair 4 | Expresso | iOnline | Only some news items have a somewhat related counterpart in the other feed | 1-3 |
| Pair 5 | Expresso | TSF | Four out of five news items have a similar counterpart in the other feed. The remaining one has no similar counterpart | 7-8 |

middle of the scale (meaning, the more similar feeds had ratings below what was expected and the less similar ones showed a higher rating). However, the relative positioning of the feed pairs among themselves still was as expected, with Pair 1 having the highest rating and Pair 4 the lowest. This flattening of results is probably derived from the high amplitude of response values. For instance, Pair 2 was rated as both 0 and 10. This overall skewing of results is likely a product of the fact that news feed similarity is ambiguous. If for one person two articles about sports could be considered similar, another person could think that they need to talk about the same specific event to be similar. It is not clear which one is best or worst. If this is true for article similarity, the discrepancies can be even greater in feed similarity, where multiple news articles need to be taken into consideration. Nevertheless, we can still see that, even if the survey aren't the exact rating values that were expected, it was still able to place the pairs of feeds in the intended relative similarity order.

### 4.3.3 Similarity Computation Results

Now that we analyzed the results obtained through the survey, it is time to compare them to the feed similarity strategies. To do so, we ran all strategies defined in 4.1.3, using Dice's Coefficient as the text similarity measure. In order to better compare the results, and similar to what was done in Section 4.1.4, the values were converted to percentage points, having the greatest obtained value equivalent to 100%. The results obtained in this experiment can be seen in Figure 4.9.

First, we can see that the results obtained through the feed similarity strategies are all relatively low in absolute value when compared to the user rating and the expected results. Nevertheless, and in line with what happened with the user ratings, the relative positioning of the feed pairs among

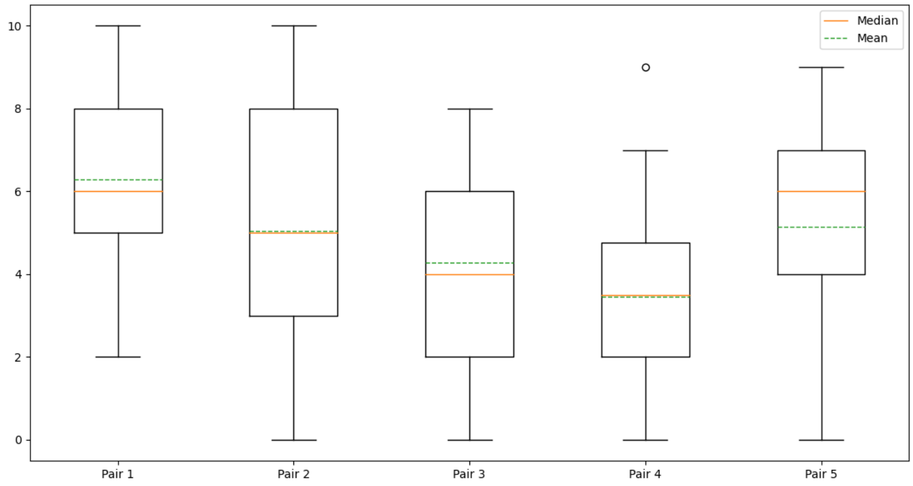


Figure 4.8: User study results distribution.

themselves still was as expected, mainly with the strategies that have been the most consistent - FF, CC, and CCE. CC-MRR continues to show its application as a complementary measure. As an example, FC, CC, and CCE produced similar results for Pairs 2 and 5. However, the value of CC-MRR is higher for Pair 5. By looking and the pairs of feeds themselves, which can be seen in Appendix A, we can see that the news items that were probably matched with each other during the cross comparison portion of the algorithm are significantly farther apart in Pair 2 when compared to Pair 5. This is consistent with the values given by CC-MRR.

The main takeaway from this experiment is, perhaps, that it is not very useful to look at the feed similarity values in an absolute way. For instance, the fact that Pair 1 has an FC value of 0.455 does not easily convey the fact that the two feeds were, in fact, very similar to each other. On the other hand, by saying that Pair 1 has an FC value of 0.455 and Pair 2 a value of 0.213, we can immediately understand that Pair 1 is probably much more similar than Pair 2. The true usefulness of these feed similarity measures probably doesn't lie in calculating absolute similarity values, but instead, in comparing different sets of feeds, ordering them based on the degree of similarity they have amongst each other.

## 4.4 Considerations on Complexity

Given the extent to which the different strategies were tested and how close they were to each other in terms of results, it is important to briefly discuss the algorithms' complexity, both temporal and spacial. We will base ourselves in the algorithms' pseudocode as shown in Section 4.1.3.

Starting with the simplest option, FC is completely independent of the number of news present in each source, having a temporal complexity of *O(1)*. This means that the algorithm should

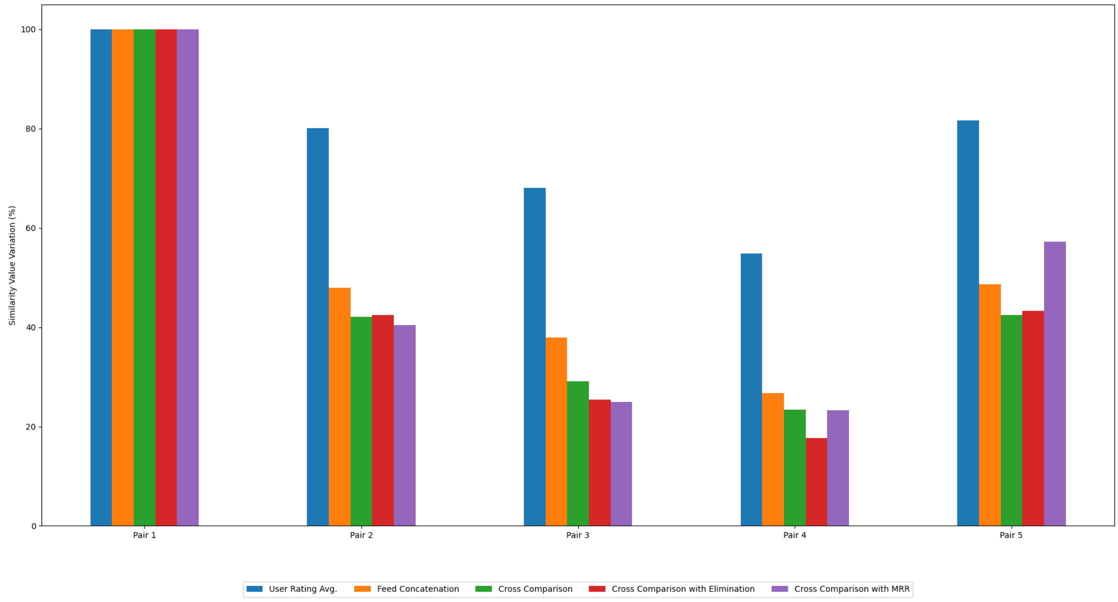


Figure 4.9: User study results comparison.

behave similarly in terms of execution time if we use news collected from a period of some hours or a month, as seen in the next chapter. Moreover, the algorithm also only requires word frequency at the feed level, which means that it also has a spacial complexity of *O(1)*.

CC, CCE, and CC-MRR are all variations of each other. They are very similar in their structure and we will analyse them together. Since these algorithms perform the comparisons at the article level, each article needs to be compared to each article present in a different feed. Nevertheless, we avoid repeating computations that were previously done, since the similarity calculation is symmetrical. This means that, if Article A has been compared to Article B, we skip comparing Article B to Article A. Therefore, these algorithms have a temporal complexity of *O(n log(n))*. Moreover, they require word frequency at the article level, which means that they also have a spacial complexity of *O(n)*.

Given how similar the results are between all four algorithms, as well as this complexity analysis, it is hard to argue against using FC in all situations. We are yet to encounter a situation in which the difference of performance among the algorithms is such that it justifies the loss of temporal and spacial efficiency.

## 4.5 Conclusions

In this chapter, we defined a series of feed similarity strategies. Each of these strategies had two different depths levels. First, there is a feed level, which is responsible for handling the web feed's naturally segmented structure. Then, there is a news item level, which consists of the text similarity measures that are used to compare the news articles themselves.

In regards to the news item level, we concluded that the text similarity measures could mostly be used interchangeably. Despite altering the absolute values obtained, swapping the text similarity used in our strategies did not produce a significant impact on the obtained results, especially when they are compared relative to each other.

When it comes to the feed level, FC, CC, and CCE all proved to be reliable strategies in all the different test scenarios, deviating little from each other. Any of them would be adequate to be used under most circumstances. CC-MRR also showed to be a good complementary measure, to be used in conjunction with one of the others in order to obtain further information regarding the feeds' relative structure.

Finally, we were able to observe that the tested feed similarity measures are less useful when used to obtain absolute similarity values, being more effective when applied to sets of feeds, with the purpose of comparing them amongst each other.

# Chapter 5

# Analysis of the Portuguese Media Landscape

In this chapter, we will analyse the Portuguese media landscape, trying to perform a characterization of its different news sources. To do that, we will first construct a National Global Feed and attempt to construct subfeeds that are similar to it. Then, we will perform a series of experiments to better understand the Portuguese landscape. These experiments are conducted throughout a month and include average source similarity, themed feed comparison, and subfeed construction with forced inclusion of news sources. Afterwards, we will repeat these experiments after sampling the available news, in order to mitigate the difference of article production among the different news sources. Finally, we will draw conclusions both on the Portuguese media landscape, as well as on the behaviour of the similarity strategies in the experiments.

## 5.1 National Global Feed and Subfeeds

In order to better understand the importance of each news source in the global news landscape, we decided to construct a National Global Feed (NGF). This feed is simply a concatenation of all the available single news source feeds in a given time frame. This will allow us to perform multiple experiments where we compare different feeds against a representative of the landscape as a whole. We will also be using different subfeeds of the NGF. These feeds will contain only a portion of the news sources that form the larger feed. The objective is to construct subfeeds that can be considered good substitutes to the NGF, eliminating some of the redundancy that is naturally present in the news industry.

### 5.1.1 Experiment Design

The purpose of the subfeeds is to help to identify sets of news sources that, together, are capable of emulating the news landscape as whole. If this is possible, the overall redudancy should be lower, since we achieved an high degree of similarity between the subfeed and the NGF with a smaller number of sources and, therefore, articles. So, we want the subfeeds to be as similar as possible to the NGF with the least amount of news sources as possible. In order to do that, we want the included news sources to be as different and varied as possible. This way, we avoid as much as possibly to include redundant articles in the subfeed and reach an high-degree of similarity with less news sources.

Our experiment is based on the premise that, if two sources are very different from each other, if we combine them into a subfeed, we will be closed to the NGF than if we combine two more similar sources. The idea is that dissimilar sources are more likely to cover different subjects and, as a result, their combination to be more similar to the NFG. Therefore, we start by comparing all available news sources with each other and selecting the two most different ones to be the base of the subfeed. These are merged together to form the first subfeed. Afterwards, the subfeed is compared to all unused news source feeds, with the most different news source being incorporated into the subfeed. This process is repeated until all news sources are included into the subfeed. At that point, the subfeed should be equal to the NGF.

After each news source inclusion, the newly-created subfeed is compared to the NGF, in order to track the progress of the capability that the subfeed has of emulating and representing the national news landscape. We expect that this similarity value will increase with each news source inclusion, until it reaches 1 when it contains all available news sources. Also, this experiment will be repeated for news feeds obtained during periods of four hours, a day, a week, and a month, all with data collected from January, 2015. Given the large amount of data that is expected during the latter time periods, we will only use FC paired with Dice's Coefficient, as used in the previous experiments, for the reasons discussed in Section 4.4.

### 5.1.2 Results and Conclusions

The results of subfeed construction similarity when compared to the NGF can be seen in Figure 5.1. The first noticeable result is that the average similarity increases with the increase of the time period from which the news articles were collected. This can be explained by the fact that, in long time periods like a week or a month, the amount of news articles per source increases significantly and, therefore, the likelihood of finding common words with other feeds increases as well, simply because the overall number of words increased. This highlights a limitation of the system, indicating that it is likely to show better results in smaller samples of data. Nevertheless, in all four scenarios, the behaviour of the system was as expected, with the similarity values converging towards 1 with the inclusion of new news sources. Moreover, we can also see that the system was able to obtain values of similarity in the range of 0.7-0.8 using about half the available sources. This not only confirms our perception that redundancy is common among news sources, but also reveals that the system is able to properly combine news sources in a way that is able to emulate the NGF.

If Figure 5.2, we can see the average entry position of each source in the subfeed during the described experiments. Meaning, if a source was a part of the initial two sources that form the

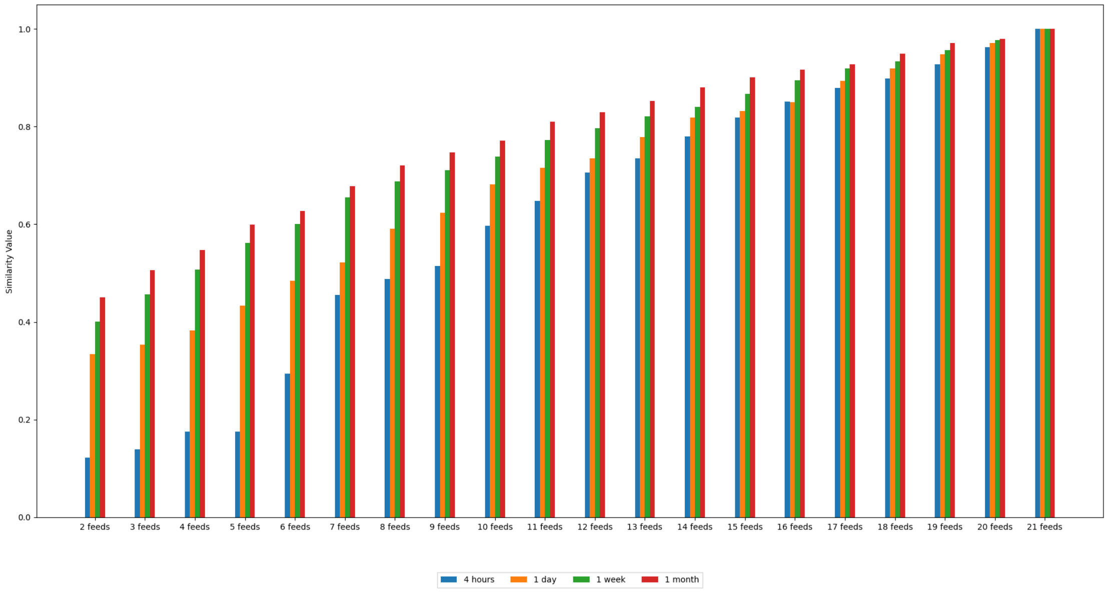


Figure 5.1: Results of subfeed construction.

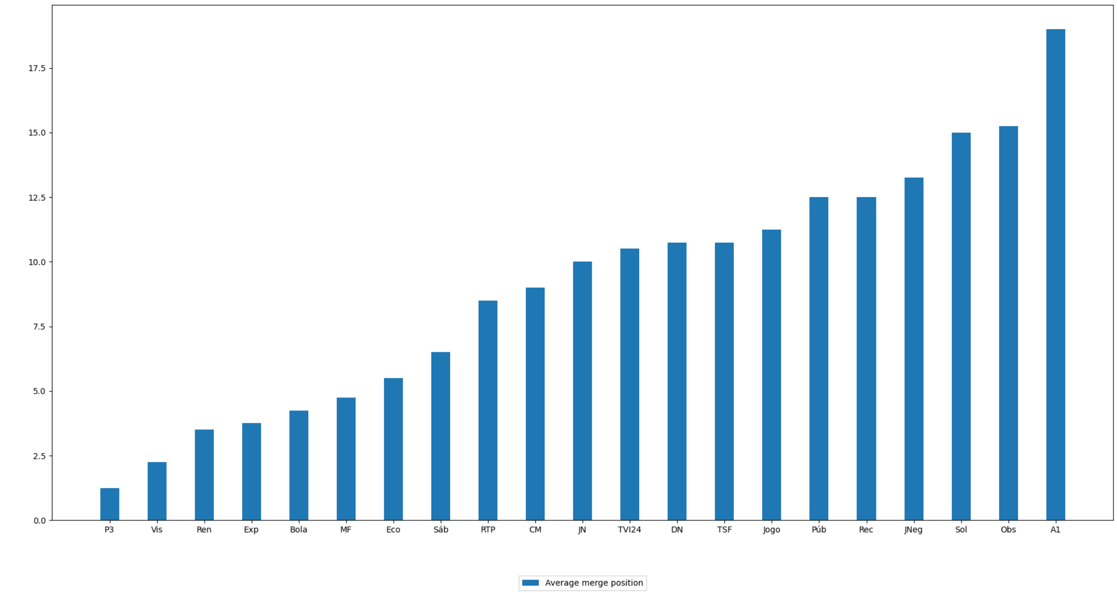


Figure 5.2: Subfeed construction source entry average.

initial subfeed, its entry position is 0. If it is the next selected source, its entry position is 1, and so on. As we can see, there are sources that constantly are included in the early stages of the subfeed construction, as well as sources that are constantly left for last. A source that is constantly chosen in the beginning is a source that is likely very different from many other sources, since it shows low similarity with the subfeed that is being constructed. This is, for instance, expected behaviour for the source that showed the lowest average entry position, *P3*, since it is a supplement of a larger

Table 5.1: Comparison of subfeed's source entry position.

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **4 Hours** | JN | MF | Sab | P3 | DN | Ren | Bola | Exp | Vis | Pub | Eco | CM | TSF | Jogo | RTP | Rec | JNeg | TVI24 | Sol | Obs | A1 |
| **Day** | Sab | Bola | P3 | Vis | Eco | Ren | Exp | CM | RTP | Pub | TVI24 | DN | MF | Sol | JN | TSF | JNeg | Jogo | Obs | Rec | A1 |
| **Week** | Vis | Bola | P3 | Exp | Ren | MF | Eco | RTP | Sab | TSF | CM | Jogo | TVI24 | JN | Obs | JNeg | Sol | DN | Pub | Rec | A1 |
| **Month** | Vis | Rec | P3 | Exp | Ren | MF | Eco | TVI24 | Jogo | RTP | JNeg | TSF | CM | Bola | Obs | DN | Jn | Sol | Pub | Sab | A1 |
| **Forced** | Rec | Vis | Sab | P3 | Eco | Ren | Exp | CM | RTP | Pub | TVI24 | MF | Sol | DN | JN | TSF | JNeg | Bola | Jogo | Obs | A1 |

newspaper. In the other hand, a source that is left for last does not necessarily mean that it needs to be very similar to all other sources, but that it is never very dissimilar to anything, since only the most dissimilar source is chosen to integrate the subfeed at each iteration. This is expected behaviour for more generalist news sources, that have some amount of coverage of most subjects.

Another interesting result was discovered after analysing the entry position of specific news sources. The list of sources, ordered by their entry position in each of the scenarios, can be seen in Table 5.1. Each row of this table indicates the order in which the sources were included into the subfeed in the given scenario, from left to right. We have also highlighted the entry of the sources *A Bola* and *Record* - both sports-based news sources - in red and green, respectively. As we can see, there is a tendency for these two sources to enter the subfeed at opposite ends of the spectrum, which is to be expected, given the similar topics that they cover on their news articles.

To confirm this, we repeated the experiment done with the news collected in the time period of a day, but forcing *Record*, which was left to near last in the regular experiment, to be part of the initial pair of news that form the initial subfeed. As expected, this resulted in the source *A Bola* to be included in the subfeed much later. This shows that this methodology could have some application in systems where a user selects some sources that they would like to be part of their news diet, and the system would then be able to suggest other news sources to be part of a balanced news feed that not only included the sources selected by the user, but other complementary news sources, being this way able to emulate the NGF without much of the unnecessary redundancy.

## 5.2 Month-Long Analysis

Given the limitations found when dealing with large amounts of news articles per source identified in the previous experiments, we decided to change the approach to the remaining experiments. We will focus on limiting the collected news articles to periods of one day. We will then repeat the experiments throughout a 30-day period, to avoid basing the results on the specificities of a single day. The experiments were conducted using data from January, 2015. Our goal is to identify patterns withing the available sources, such as groups of sources that complement each other well. Moreover, we also aim to test which methodologies best fit the comparison strategies that we previously defines

We will first compare all news sources against each other, as well as against the NGF. As said, the results are obtained daily and averaged throughout the 30 days. The purpose is to generally understand which sources are similar to others, as well as what sources are closer to the general landscape. Afterwards, we will construct themed subfeeds based on the source's type of news articles produced - daily newspapers, weekly newspapers, radio and TV, and sports newspapers. The purpose is to understand, inside these categories, which ones are more distinct and differ most from the themed subfeed. Finally, we will repeat the subfeed construction of the previous example, but forcing each source to be a part of the starting subfeed. The purpose is to try to match news sources to other complementary sources that form varied subfeeds.

### 5.2.1 Source Similarity

The results of the average of similarity between all the sources can be seen in Table 5.2. In this table, the cells are highlighter with colors based on their value: red for values below 0.1; orange for values between 0.1 and 0.2; yellow for values between 0.2 and 0.3; green for values equal to 0.3 or greater. Also, the table is arranged vertically accordingly to groups of sources that produce similar content. The groups are daily newspapers, weekly newspapers, radio and television, sports, economics and finance, and others.

We can immediately see some patterns related to the groups of sources defined in the table. First, we can see that most source groups are similar amongst themselves. The exception is the weekly newspapers, that have low similarity. Moreover, we can see that sources belonging to daily newspapers are very similar to radio and television sources. This is somewhat to be expected, since they all are dedicated to the production of generalist news articles. Other source groups also behave as expected. Sports sources, for instance, have high similarity values only when compared against other Sports sources. We can also see that there are two sources that are consistently dissimilar. One of them is *P3*, which is mostly expected behaviour, for reasons that we discussed previously. The other one is *Visão*. This is more unexpected, especially because it is the one that has the lowest overall similarity values, which cannot be explained only with the weekly newspapers news content. In Section 5.3 we will perform further analysis to the dataset in order to better understand these low similarity results.

The results related to the the NGF must be carefully analysed. It is likely that they suffer from the same problems of dimensionality identified in Section 5.1.2, since the NGF has a considerable larger amount of news articles when compared to each individual source. It is likely that the results of this comparison can be affected by each source's article representativity in the NGF.

### 5.2.2 Themed Feeds

The results for the themed subfeed construction can be seen in Table 5.3. We used the same groups defined in the previous section, using only the ones that have at least four news sources.

In the first group, the daily newspapers, we can see that the comparison of each source with the themed subfeed mostly follows that is to be expected. *Jornal de Notícias*, *Diário de Notícias*, and *Público* all have similar values, and all of them are high. This not only means that they are similar to the themed subfeed, but also that they are similar to each other - which was already concluded in the previous section -, since the only way to reach such high values of similarity is to find similarities not only with the articles in the themed subfeed that are originated from the

Table 5.2: Results of month-long average similarity calculation.

| | | Correio da Manhã | Diário de Notícias | Jornal de Notícias | Público | Expresso | Sábado | Sol | Visão | Antena 1 | Renascença | RTP | TSF | TVI24 | A Bola | O Jogo | MaisFutebol | Record | Económico | Jornal de Negócios | Observador | P3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **All** | 0.197 | 0.282 | 0.270 | 0.259 | 0.105 | 0.311 | 0.285 | 0.044 | 0.412 | 0.179 | 0.241 | 0.234 | 0.251 | 0.344 | 0.279 | 0.211 | 0.272 | 0.130 | 0.279 | 0.249 | 0.105 |
| **Daily** | **Correio da Manhã** | - | 0.306 | 0.338 | 0.289 | 0.193 | 0.342 | 0.287 | 0.106 | 0.286 | 0.263 | 0.291 | 0.307 | 0.324 | 0.229 | 0.245 | 0.223 | 0.244 | 0.199 | 0.242 | 0.293 | 0.128 |
| | **Diário de Notícias** | - | - | 0.365 | 0.330 | 0.212 | 0.315 | 0.323 | 0.094 | 0.332 | 0.263 | 0.336 | 0.338 | 0.332 | 0.332 | 0.343 | 0.297 | 0.323 | 0.216 | 0.289 | 0.313 | 0.140 |
| | **Jornal de Notícias** | - | - | - | 0.338 | 0.207 | 0.297 | 0.377 | 0.113 | 0.410 | 0.301 | 0.371 | 0.424 | 0.385 | 0.274 | 0.274 | 0.263 | 0.268 | 0.226 | 0.290 | 0.350 | 0.131 |
| | **Público** | - | - | - | - | 0.205 | 0.283 | 0.316 | 0.100 | 0.338 | 0.283 | 0.315 | 0.323 | 0.328 | 0.260 | 0.259 | 0.238 | 0.263 | 0.222 | 0.298 | 0.315 | 0.149 |
| **Weekly** | **Expresso** | - | - | - | - | - | 0.183 | 0.204 | 0.108 | 0.184 | 0.203 | 0.214 | 0.219 | 0.215 | 0.147 | 0.155 | 0.155 | 0.164 | 0.221 | 0.214 | 0.204 | 0.216 |
| | **Sábado** | - | - | - | - | - | - | 0.289 | 0.089 | 0.303 | 0.243 | 0.289 | 0.284 | 0.297 | 0.300 | 0.296 | 0.262 | 0.427 | 0.203 | 0.292 | 0.278 | 0.143 |
| | **Sol** | - | - | - | - | - | - | - | 0.109 | 0.373 | 0.289 | 0.321 | 0.348 | 0.348 | 0.245 | 0.244 | 0.211 | 0.246 | 0.226 | 0.318 | 0.329 | 0.145 |
| | **Visão** | - | - | - | - | - | - | - | - | 0.094 | 0.123 | 0.116 | 0.116 | 0.112 | 0.068 | 0.072 | 0.078 | 0.082 | 0.091 | 0.094 | 0.110 | 0.080 |
| **Radio and TV** | **Antena 1** | - | - | - | - | - | - | - | - | - | 0.300 | 0.359 | 0.376 | 0.399 | 0.258 | 0.237 | 0.198 | 0.242 | 0.224 | 0.340 | 0.357 | 0.131 |
| | **Renascença** | - | - | - | - | - | - | - | - | - | - | 0.288 | 0.306 | 0.319 | 0.186 | 0.190 | 0.180 | 0.198 | 0.215 | 0.261 | 0.282 | 0.137 |
| | **RTP** | - | - | - | - | - | - | - | - | - | - | - | 0.366 | 0.325 | 0.289 | 0.289 | 0.260 | 0.284 | 0.238 | 0.296 | 0.322 | 0.147 |
| | **TSF** | - | - | - | - | - | - | - | - | - | - | - | - | 0.399 | 0.258 | 0.264 | 0.234 | 0.242 | 0.238 | 0.298 | 0.345 | 0.140 |
| | **TVI24** | - | - | - | - | - | - | - | - | - | - | - | - | - | 0.224 | 0.219 | 0.197 | 0.229 | 0.243 | 0.326 | 0.342 | 0.141 |
| **Sports** | **A Bola** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 0.432 | 0.379 | 0.394 | 0.158 | 0.236 | 0.241 | 0.131 |
| | **O Jogo** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 0.398 | 0.397 | 0.162 | 0.224 | 0.242 | 0.136 |
| | **MaisFutebol** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 0.394 | 0.152 | 0.199 | 0.216 | 0.131 |
| | **Record** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 0.164 | 0.232 | 0.242 | 0.139 |
| **E&F** | **Económico** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 0.281 | 0.219 | 0.127 |
| | **Jornal de Negócios** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 0.303 | 0.140 |
| **Other** | **Observador** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 0.143 |
| | **P3** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |

source, but also to be similar to articles contributed by other sources. *Correio da Manhã* is the one source that presents a lower similarity value when compared to the themed subfeed. This is likely a product of this source's more sensationalist line of article producing.

The second group, the weekly newspapers, has considerably different results. Two of the

| | All | JN | DN | Pub | CM |
|---|---|---|---|---|---|
| **All** | 1.000 | 0.581 | 0.599 | 0.561 | 0.443 |
| **JN** | - | 1.000 | 0.363 | 0.338 | 0.338 |
| **DN** | - | - | 1.000 | 0.330 | 0.306 |
| **Pub** | - | - | - | 1.000 | 0.289 |
| **CM** | - | - | - | - | 1.000 |

(a) Daily.

| | All | Exp | Vis | Sol | Sab |
|---|---|---|---|---|---|
| **All** | 1.000 | 0.312 | 0.143 | 0.729 | 0.684 |
| **Exp** | - | 1.000 | 0.108 | 0.204 | 0.183 |
| **Vis** | - | - | 1.000 | 0.109 | 0.089 |
| **Sol** | - | - | - | 1.000 | 0.289 |
| **Sab** | - | - | - | - | 1.000 |

(b) Weekly.

| | All | TSF | A1 | TVI24 | Ren | RTP |
|---|---|---|---|---|---|---|
| **All** | 1.000 | 0.454 | 0.728 | 0.477 | 0.354 | 0.463 |
| **TSF** | - | 1.000 | 0.376 | 0.352 | 0.306 | 0.366 |
| **A1** | - | - | 1.000 | 0.399 | 0.300 | 0.359 |
| **TVI24** | - | - | - | 1.000 | 0.319 | 0.325 |
| **Ren** | - | - | - | - | 1.000 | 0.288 |
| **RTP** | - | - | - | - | - | 1.000 |

(c) TV and Radio.

| | All | Jogo | MF | Bola | Rec |
|---|---|---|---|---|---|
| **All** | 1.000 | 0.583 | 0.461 | 0.689 | 0.570 |
| **Jogo** | - | 1.000 | 0.381 | 0.432 | 0.397 |
| **MF** | - | - | 1.000 | 0.378 | 0.364 |
| **Bola** | - | - | - | 1.000 | 0.394 |
| **Rec** | - | - | - | - | 1.000 |

(d) Sports.

Table 5.3: Results of themed subfeed construction.

sources - *Sol* and *Sábado* - have very high similarity values with the themed subfeed, while the other two sources - *Expresso* and *Visão*, have very low similarity values. We had already discussed reasons why these sources have less similarity amongst each other, unlike the other groups. However, this is unlikely to fully explain this slated results. This drop of similarity likely means that the number of articles that each source contributes to the themed subfeed is considerably different, since only that way values this low could be reached. We will discuss this further in Section 5.3.

The third group, radio and television, mostly has results similar to the first group. Four of the five news sources - *TSF*, *TVI24*, *Renascença*, and *RTP* all have similar values when compared to the themed subfeed. The exception is *Antena 1*, which has a considerably higher similarity value. Unlike the second group, all sources in this group are reasonably similar to each other and, therefore, higher similarity values when compared to the themed subfeed are expected. Nevertheless, such a discrepancy is probably only explained by an also significant difference in the number of articles produced by each source.

Finally, the fourth group, the sports newspapers, also have results equal to what would be expected, with three of the four sources - *O Jogo*, *A Bola*, and *Record* all having high similarity values, all close to each other, since they all produce news articles about a specific set of events. The fourth source, *MaisFutebol*, lags slightly behind. This difference is, however, not as pronounced as in the other groups. This may be explained by the fact that *MaisFutebol* is even more dedicated to the football sport that the others - despite the fact that, in Portugal, most of the sports news coverage is dedicated to football. Nevertheless, the difference in the coverge of other sports' events might explain this slight drop of similarity.

### 5.2.3 Subfeed Construction with Forced Entry

Table 5.4: Results of subfeed construction with forced entry.

| | | Correio da Manhã | Diário de Notícias | Jornal de Notícias | Público | Expresso | Sábado | Sol | Visão | Antena 1 | Renascença | RTP | TSF | TVI24 | A Bola | O Jogo | MaisFutebol | Record | Económico | Jornal de Negócios | Observador | P3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Daily | **Correio da Manhã** | - | 0 | 3 | 0 | 66 | 0 | 0 | 93 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 21 | 0 | 17 | 0 | 0 | 97 |
| | **Diário de Notícias** | 3 | - | 0 | 0 | 72 | 0 | 0 | 97 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 7 | 0 | 17 | 3 | 0 | 97 |
| | **Jornal de Notícias** | 0 | 0 | - | 0 | 76 | 0 | 0 | 93 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 14 | 0 | 17 | 0 | 0 | 97 |
| | **Público** | 0 | 0 | 0 | - | 62 | 0 | 0 | 93 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 97 |
| Weekly | **Expresso** | 3 | 0 | 0 | 0 | - | 7 | 0 | 93 | 3 | 7 | 0 | 0 | 0 | 21 | 14 | 38 | 3 | 14 | 0 | 0 | 97 |
| | **Sábado** | 0 | 0 | 0 | 0 | 72 | - | 0 | 100 | 0 | 3 | 0 | 0 | 0 | 3 | 0 | 3 | 0 | 17 | 0 | 3 | 97 |
| | **Sol** | 0 | 0 | 0 | 0 | 48 | 3 | - | 93 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 38 | 0 | 14 | 3 | 0 | 97 |
| | **Visão** | 3 | 0 | 0 | 0 | 79 | 0 | 0 | - | 0 | 0 | 0 | 0 | 0 | 72 | 0 | 0 | 28 | 24 | 0 | 0 | 93 |
| Radio and TV | **Antena 1** | 0 | 0 | 0 | 0 | 59 | 7 | 0 | 93 | - | 0 | 0 | 0 | 0 | 3 | 0 | 24 | 3 | 14 | 0 | 0 | 97 |
| | **Renascença** | 0 | 0 | 0 | 0 | 24 | 0 | 0 | 86 | 0 | - | 0 | 0 | 0 | 90 | 0 | 0 | 10 | 14 | 0 | 0 | 76 |
| | **RTP** | 0 | 0 | 0 | 0 | 24 | 0 | 0 | 86 | 0 | 0 | - | 0 | 0 | 41 | 0 | 0 | 59 | 14 | 0 | 0 | 76 |
| | **TSF** | 0 | 0 | 0 | 0 | 62 | 3 | 0 | 100 | 0 | 0 | 0 | - | 0 | 3 | 0 | 21 | 0 | 14 | 0 | 0 | 97 |
| | **TVI24** | 0 | 0 | 0 | 0 | 17 | 0 | 0 | 86 | 0 | 0 | 0 | 0 | - | 93 | 0 | 0 | 7 | 14 | 0 | 0 | 83 |
| Sports | **A Bola** | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 86 | 0 | 0 | 0 | 0 | 0 | - | 0 | 0 | 38 | 14 | 0 | 0 | 62 |
| | **O Jogo** | 0 | 0 | 0 | 0 | 79 | 0 | 0 | 100 | 0 | 3 | 0 | 0 | 0 | 3 | - | 0 | 0 | 17 | 0 | 0 | 97 |
| | **MaisFutebol** | 3 | 0 | 0 | 0 | 66 | 0 | 0 | 100 | 3 | 0 | 0 | 0 | 0 | 3 | 0 | - | 0 | 24 | 0 | 3 | 97 |
| | **Record** | 0 | 0 | 0 | 0 | 100 | 3 | 0 | 90 | 3 | 0 | 0 | 0 | 0 | 14 | 10 | 7 | - | 10 | 0 | 0 | 62 |
| E&F | **Económico** | 0 | 0 | 0 | 0 | 34 | 7 | 0 | 97 | 0 | 0 | 0 | 0 | 0 | 14 | 0 | 45 | 3 | - | 3 | 0 | 97 |
| | **Jornal de Negócios** | 3 | 0 | 0 | 0 | 45 | 3 | 0 | 93 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 38 | 3 | 10 | - | 0 | 100 |
| Other | **Observador** | 0 | 0 | 3 | 0 | 55 | 0 | 0 | 93 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 31 | 0 | 17 | 0 | - | 97 |
| | **P3** | 14 | 3 | 3 | 0 | 76 | 7 | 0 | 93 | 7 | 3 | 0 | 0 | 3 | 17 | 3 | 45 | 3 | 17 | 0 | 3 | - |

The results for subfeed construction with forced selection of news sources as starters can be seen in Table 5.4. In this table, the source indicated in each row was used as the starting point to the construction of the subfeed. The value of each cell represents the percentage of days in which the source indicated by the column was included **withing the first three sources** integrated into the subfeed that started with the source indicated by the row. The cells are highlighted with colors based on their value: red for values below 20%; yellow for values between 20% and 35%; green for values equal to 35% or greater.

We can easily spot some patterns. First, we can see that in the vast majority of the time, the sources selected at the early stages of the subfeed construction are *Expresso*, *Visão*, and *P3*. Moreover, most of the sources are never selected in those iterations of the subfeed construction, even though this is probably a product of the fact that the other ones were always given priority

Table 5.5: Distribution of news articles among news sources.

| | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Correio da Manhã** | 43 | 66 | 63 | 53 | 53 | 55 | 41 | 45 | 62 | 61 | 54 | 53 | 36 | 47 | 69 | 81 | 49 | 49 | 33 | 42 | 53 | 47 | 57 | 64 | 63 | 79 | 66 | 92 |
| **Diário de Notícias** | 27 | 39 | 41 | 38 | 56 | 70 | 54 | 38 | 46 | 52 | 33 | 89 | 97 | 77 | 101 | 77 | 51 | 46 | 65 | 62 | 120 | 87 | 77 | 70 | 63 | 70 | 114 | 155 |
| **Jornal de Notícias** | 40 | 54 | 39 | 50 | 43 | 49 | 63 | 60 | 71 | 39 | 40 | 67 | 64 | 73 | 72 | 57 | 47 | 38 | 45 | 51 | 65 | 61 | 53 | 35 | 41 | 51 | 63 | 68 |
| **Público** | 40 | 48 | 45 | 46 | 50 | 52 | 55 | 49 | 55 | 55 | 44 | 49 | 50 | 52 | 64 | 62 | 45 | 42 | 43 | 48 | 49 | 50 | 46 | 47 | 40 | 81 | 79 | 90 |
| **Expresso** | 8 | 11 | 8 | 13 | 17 | 30 | 12 | 22 | 19 | 11 | 3 | 26 | 27 | 24 | 20 | 11 | 5 | 5 | 21 | 29 | 35 | 27 | 17 | 9 | 8 | 14 | 17 | 25 |
| **Sábado** | 86 | 54 | 13 | 12 | 48 | 66 | 74 | 55 | 50 | 21 | 21 | 61 | 171 | 171 | 101 | 185 | 133 | 79 | 147 | 145 | 124 | 121 | 127 | 183 | 124 | 219 | 154 | 249 |
| **Sol** | 29 | 43 | 33 | 41 | 51 | 64 | 40 | 67 | 47 | 34 | 24 | 40 | 49 | 63 | 62 | 55 | 40 | 30 | 61 | 49 | 66 | 74 | 62 | 36 | 26 | 59 | 66 | 57 |
| **Visão** | 10 | 8 | 11 | 7 | 7 | 8 | 5 | 4 | 8 | 1 | 4 | 5 | 4 | 6 | 8 | 7 | 4 | 3 | 4 | 10 | 8 | 6 | 4 | 3 | 3 | 10 | 7 | 12 |
| **Antena 1** | 57 | 67 | 41 | 57 | 96 | 114 | 145 | 126 | 134 | 70 | 66 | 130 | 160 | 132 | 147 | 128 | 76 | 55 | 122 | 100 | 126 | 128 | 134 | 66 | 78 | 133 | 124 | 157 |
| **Renascença** | 36 | 72 | 37 | 46 | 72 | 93 | 112 | 77 | 75 | 66 | 63 | 104 | 73 | 91 | 80 | 106 | 35 | 39 | 82 | 95 | 84 | 99 | 115 | 42 | 40 | 101 | 90 | 81 |
| **RTP** | 40 | 47 | 30 | 18 | 38 | 40 | 50 | 57 | 40 | 25 | 28 | 31 | 54 | 57 | 67 | 74 | 34 | 28 | 40 | 54 | 58 | 52 | 48 | 32 | 21 | 43 | 42 | 59 |
| **TSF** | 24 | 29 | 26 | 30 | 44 | 54 | 41 | 38 | 43 | 29 | 28 | 42 | 45 | 47 | 49 | 60 | 38 | 35 | 49 | 38 | 52 | 47 | 60 | 30 | 30 | 47 | 52 | 50 |
| **TVI24** | 38 | 85 | 27 | 26 | 89 | 99 | 95 | 79 | 82 | 51 | 47 | 70 | 93 | 73 | 105 | 115 | 40 | 34 | 88 | 94 | 70 | 88 | 84 | 33 | 33 | 101 | 101 | 94 |
| **A Bola** | 70 | 101 | 84 | 75 | 91 | 108 | 94 | 97 | 94 | 100 | 88 | 110 | 102 | 99 | 93 | 106 | 93 | 72 | 101 | 99 | 80 | 92 | 91 | 80 | 103 | 113 | 118 | 123 |
| **O Jogo** | 36 | 69 | 79 | 68 | 79 | 91 | 80 | 88 | 88 | 82 | 66 | 113 | 81 | 89 | 96 | 95 | 85 | 71 | 82 | 79 | 88 | 85 | 95 | 102 | 81 | 93 | 97 | 79 |
| **MaisFutebol** | 26 | 74 | 71 | 65 | 79 | 80 | 80 | 79 | 82 | 72 | 76 | 75 | 76 | 89 | 91 | 74 | 74 | 72 | 81 | 87 | 74 | 80 | 63 | 70 | 95 | 92 | 90 | 105 |
| **Record** | 50 | 74 | 71 | 66 | 70 | 75 | 76 | 70 | 81 | 75 | 71 | 73 | 71 | 74 | 91 | 86 | 78 | 70 | 80 | 80 | 67 | 74 | 79 | 79 | 93 | 92 | 115 | 104 |
| **Económico** | 2 | 19 | 2 | 9 | 37 | 53 | 53 | 48 | 36 | 3 | 6 | 53 | 47 | 37 | 48 | 17 | 1 | 4 | 32 | 39 | 46 | 47 | 42 | 8 | 9 | 18 | 23 | 35 |
| **Jornal de Negócios** | 9 | 52 | 9 | 24 | 68 | 73 | 69 | 68 | 67 | 27 | 38 | 73 | 72 | 66 | 99 | 90 | 22 | 27 | 69 | 69 | 70 | 70 | 57 | 26 | 46 | 111 | 94 | 100 |
| **Observador** | 29 | 47 | 36 | 38 | 41 | 39 | 49 | 50 | 46 | 45 | 46 | 35 | 55 | 42 | 58 | 59 | 46 | 50 | 41 | 25 | 41 | 32 | 47 | 40 | 41 | 82 | 93 | 87 |
| **P3** | 1 | 22 | 1 | 1 | 11 | 11 | 11 | 13 | 18 | 2 | 2 | 19 | 20 | 17 | 9 | 14 | 3 | 6 | 20 | 10 | 18 | 15 | 15 | 3 | 3 | 16 | 15 | 16 |

The purpose of this experiment was to find sources that complement others. This is useless if, at all times, the same sources are selected. This is, however, consistent with the results calculated in Section 5.2.1, where *Visão* and *P3* had low similarity values when compared to all other sources. As explained in Section 5.1.1, the source selected to integrate the subfeed at a given iteration is the one that is less similar to the current subfeed. It is, therefore, expected that these sources, that are so different from the rest, are constantly selected. Even if these results confirm what we already knew about these sources, they are not useful to detect possible groups of complementary news sources.

## 5.3 Month-Long Analysis with Sampling

Given the high gaps of similarity values discovered through the last few sections, we decided to perform some statistical analysis on the data used in the experiments. In Table 5.5, we can see the number of articles produced by each source on each day used in the experiments conducted in the previous sections of this chapter. Highlighted in red are the cells with values lower than 20 news articles.

As we can see, there are significant discrepancies among the different news sources. Some constantly publish over a hundred articles per day, while others produce less that ten. This difference is consistent throughout the month. The sources with less articles coincide with the ones that had consistently low similarity values in the previous experiments, mainly *Expresso*, *Visão*, and *P3*.

In order to attempt to mitigate the differences, we decided to repeat the experiments conducted throughout Section 5.2, except that, for each source, in each day, if that source produced over 20 news articles, 20 news articles are randomly selected from the available ones, and only these ones are considered. This is bound to introduce some randomness into the results, since many articles will no longer find their counterparts in other news sources, dropping the overall similarity values. Nevertheless, we expect that this sampling method attenuates the influence of the number of articles in the similarity value computation.

### 5.3.1 Source Similarity

The results of the average of similarity between all the sources after sampling can be seen in Table 5.6. As expected, the similarity results are overall significantly lower than the ones obtained in Section 5.2.1. Therefore, in order to better represent the obtained results, we decided to update the color scale. As a results, the cells are highlighter with colors based on their value: red for values below 0.1; orange for values between 0.1 and 0.15; yellow for values between 0.15 and 0.2; green for values equal to 0.2 or greater. The table is sorted using the same groups as before.

Overall, the patterns observed in the previous experiment repeat themselves in this one. In particular, both *Visão* and *P3* still have the lowest overall similarity values. The gap has, however, been reduced significantly, since the overall results are also lower. The sampling process seems to have overall attenuated the difference between the sources without affecting the observed patterns and general similarity. One source that has dropped its overall similarity values is *Correio da Manhã*, which now has values on par with the ones of *P3*. Given the previously discussed nature of the articles produced by this news source, it is likely that it has been affected by the sampling process more than most, loosing some of the more mainstream news items it produces to the random selection of articles.

### 5.3.2 Themed Feeds

The results for the themed subfeed construction after sampling can be seen in Table 5.7. The groups used are the same as the ones described in Section 5.2.2.

The results of the first group, the daily newspapers, seem to have suffered little change when compared to the previous experiment. The overall similarity values are lower, which is consistent with the source similarity calculation obtained after sampling. *Jornal de Notícias*, *Diário de Notícias*, and *Público* have produced similar values when compared to the themed subfeed, with *Correio da Manhã* having a slightly lower similarity value. The drop of similarity values observed in the previous experiment does not seem to have significantly affected its behaviour in this experiment.

The results of the second group appear to have improved when compared to the previously conducted experiment without sampling. *Sol* has the highest similarity when compared to the themed feed, with *Sábado* and *Expresso* being slightly behind. This coincides more with the behaviour expected from the latter, which appears to have benefitted from the sampling process. *Visão*, however, still has a significantly lower similarity value, which is coherent with its low similarity values observed in Section 5.3.1. It is expected that *Expresso* benefits more from the sampling than *Visão*, since the latter still has significantly less daily news articles. These results appear to corroborate the hypothesis that the sampling process was not enough to normalize the results of the source *Visão*.

Table 5.6: Results of month-long average similarity calculation after sampling.

| | | Correio da Manhã | Diário de Notícias | Jornal de Notícias | Público | Expresso | Sábado | Sol | Visão | Antena 1 | Renascença | RTP | TSF | TVI24 | A Bola | O Jogo | MaisFutebol | Record | Económico | Jornal de Negócios | Observador | P3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **All** | 0.144 | 0.200 | 0.225 | 0.208 | 0.161 | 0.163 | 0.249 | 0.074 | 0.243 | 0.163 | 0.244 | 0.235 | 0.169 | 0.203 | 0.169 | 0.133 | 0.162 | 0.148 | 0.232 | 0.218 | 0.179 |
| **Daily** | **Correio da Manhã** | - | 0.179 | 0.197 | 0.168 | 0.150 | 0.170 | 0.167 | 0.105 | 0.173 | 0.159 | 0.191 | 0.191 | 0.183 | 0.145 | 0.152 | 0.128 | 0.143 | 0.136 | 0.145 | 0.178 | 0.105 |
| | **Diário de Notícias** | - | - | 0.224 | 0.200 | 0.191 | 0.178 | 0.208 | 0.107 | 0.201 | 0.171 | 0.233 | 0.219 | 0.186 | 0.203 | 0.208 | 0.172 | 0.191 | 0.160 | 0.179 | 0.195 | 0.119 |
| | **Jornal de Notícias** | - | - | - | 0.215 | 0.190 | 0.171 | 0.244 | 0.125 | 0.254 | 0.204 | 0.256 | 0.286 | 0.236 | 0.181 | 0.175 | 0.141 | 0.162 | 0.176 | 0.189 | 0.232 | 0.117 |
| | **Público** | - | - | - | - | 0.181 | 0.171 | 0.204 | 0.112 | 0.219 | 0.189 | 0.216 | 0.213 | 0.199 | 0.166 | 0.156 | 0.142 | 0.151 | 0.169 | 0.200 | 0.204 | 0.127 |
| **Weekly** | **Expresso** | - | - | - | - | - | 0.153 | 0.190 | 0.108 | 0.188 | 0.177 | 0.195 | 0.201 | 0.173 | 0.135 | 0.136 | 0.122 | 0.139 | 0.198 | 0.206 | 0.181 | 0.122 |
| | **Sábado** | - | - | - | - | - | - | 0.190 | 0.108 | 0.188 | 0.177 | 0.195 | 0.201 | 0.173 | 0.135 | 0.136 | 0.122 | 0.139 | 0.198 | 0.206 | 0.181 | 0.122 |
| | **Sol** | - | - | - | - | - | - | - | 0.128 | 0.244 | 0.203 | 0.235 | 0.256 | 0.214 | 0.160 | 0.148 | 0.124 | 0.146 | 0.188 | 0.224 | 0.219 | 0.132 |
| | **Visão** | - | - | - | - | - | - | - | - | 0.120 | 0.127 | 0.132 | 0.124 | 0.126 | 0.080 | 0.078 | 0.072 | 0.086 | 0.096 | 0.109 | 0.120 | 0.080 |
| **Radio and TV** | **Antena 1** | - | - | - | - | - | - | - | - | - | 0.219 | 0.256 | 0.262 | 0.239 | 0.145 | 0.137 | 0.113 | 0.134 | 0.194 | 0.217 | 0.241 | 0.128 |
| | **Renascença** | - | - | - | - | - | - | - | - | - | - | 0.207 | 0.217 | 0.215 | 0.124 | 0.119 | 0.110 | 0.121 | 0.172 | 0.185 | 0.200 | 0.115 |
| | **RTP** | - | - | - | - | - | - | - | - | - | - | - | 0.276 | 0.212 | 0.202 | 0.193 | 0.160 | 0.183 | 0.193 | 0.214 | 0.232 | 0.133 |
| | **TSF** | - | - | - | - | - | - | - | - | - | - | - | - | 0.225 | 0.177 | 0.174 | 0.148 | 0.159 | 0.193 | 0.214 | 0.239 | 0.129 |
| | **TVI24** | - | - | - | - | - | - | - | - | - | - | - | - | - | 0.128 | 0.122 | 0.108 | 0.123 | 0.176 | 0.200 | 0.219 | 0.107 |
| **Sports** | **A Bola** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 0.252 | 0.221 | 0.219 | 0.120 | 0.148 | 0.153 | 0.116 |
| | **O Jogo** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 0.235 | 0.213 | 0.118 | 0.139 | 0.149 | 0.111 |
| | **MaisFutebol** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 0.204 | 0.104 | 0.116 | 0.122 | 0.098 |
| | **Record** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 0.113 | 0.136 | 0.142 | 0.115 |
| **E&F** | **Económico** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 0.224 | 0.176 | 0.115 |
| | **Jornal de Negócios** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 0.205 | 0.128 |
| **Other** | **Observador** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 0.121 |
| | **P3** | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |

The third group also appears to have better results after sampling. All five sources have relatively close similarity values when compared to the themed subfeed, which is what is expected. The significantly higher similarity value of *Antena 1* that was present in the same experiment before sampling has disappeared. As we can see in Section 5.3, *Antena 1* was one of the highest-

| | All | JN | DN | Pub | CM |
|---|---|---|---|---|---|
| **All** | 1.000 | 0.502 | 0.494 | 0.524 | 0.384 |
| **JN** | - | 1.000 | 0.227 | 0.211 | 0.211 |
| **DN** | - | - | 1.000 | 0.199 | 0.180 |
| **Pub** | - | - | - | 1.000 | 0.158 |
| **CM** | - | - | - | - | 1.000 |

(a) Daily.

| | All | Exp | Vis | Sol | Sab |
|---|---|---|---|---|---|
| **All** | 1.000 | 0.453 | 0.227 | 0.666 | 0.451 |
| **Exp** | - | 1.000 | 0.155 | 0.288 | 0.220 |
| **Vis** | - | - | 1.000 | 0.220 | 0.151 |
| **Sol** | - | - | - | 1.000 | 0.262 |
| **Sab** | - | - | - | - | 1.000 |

(b) Weekly.

| | All | TSF | A1 | TVI24 | Ren | RTP |
|---|---|---|---|---|---|---|
| **All** | 1.000 | 0.489 | 0.505 | 0.375 | 0.359 | 0.500 |
| **TSF** | - | 1.000 | 0.259 | 0.228 | 0.221 | 0.272 |
| **A1** | - | - | 1.000 | 0.229 | 0.219 | 0.254 |
| **TVI24** | - | - | - | 1.000 | 0.210 | 0.209 |
| **Ren** | - | - | - | - | 1.000 | 0.207 |
| **RTP** | - | - | - | - | - | 1.000 |

(c) TV and Radio.

| | All | Jogo | MF | Bola | Rec |
|---|---|---|---|---|---|
| **All** | 1.000 | 0.500 | 0.413 | 0.583 | 0.492 |
| **Jogo** | - | 1.000 | 0.225 | 0.251 | 0.220 |
| **MF** | - | - | 1.000 | 0.228 | 0.205 |
| **Bola** | - | - | - | 1.000 | 0.221 |
| **Rec** | - | - | - | - | 1.000 |

(d) Sports.

Table 5.7: Results of themed subfeed construction after sampling.

producing sources. This leads us to believe that the previously observed in similarity was mostly a product of the high density of articles from this source in the themed subfeed.

Finally, the fourth group seems to have suffered little change when compared to the previous experiment. Moreover, the gap that was observed between the *MaisFutebol* and the remaining sources diminished. This, however, might be an undesired side effect from the sampling process. Since most of the sports-related news articles in Portugal relate to football, the sampling is likely to have eliminated most of the non-football articles, since they are so few. This might have mitigated the content difference between the sports sources that we hypothesized in Section 5.2.2. Nevertheless, the difference is not that substantial, and the results are still consistent to what was to be expected.

Overall, the sampling process seems to have had a positive effect over the themed subfeed construction. This reinforces the conclusion that the difference in the amount of news articles produced has a heavy influence over the similarity results.

### 5.3.3 Subfeed Construction with Forced Entry

The results for subfeed construction with forced selection of news sources as starters after sampling can be seen in Table 5.8. The characteristics of the table have not been changed when compared to Table 5.4, including the color scheme.

When we compare both tables, we can easily see that there is significantly more variety of sources selected in the early stages of subfeed construction. Mainly, there are significantly less

Table 5.8: Results of subfeed construction with forced entry after sampling.

| | | Correio da Manhã | Diário de Notícias | Jornal de Notícias | Público | Expresso | Sábado | Sol | Visão | Antena 1 | Renascença | RTP | TSF | TVI24 | A Bola | O Jogo | MaisFutebol | Record | Económico | Jornal de Negócios | Observador | P3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Daily** | **Correio da Manhã** | - | 0 | 0 | 3 | 14 | 10 | 0 | 93 | 0 | 0 | 0 | 0 | 7 | 7 | 10 | 55 | 10 | 21 | 3 | 0 | 66 |
| | **Diário de Notícias** | 28 | - | 0 | 0 | 14 | 7 | 0 | 97 | 0 | 10 | 0 | 0 | 3 | 3 | 10 | 79 | 10 | 14 | 3 | 0 | 59 |
| | **Jornal de Notícias** | 17 | 0 | - | 0 | 24 | 10 | 0 | 93 | 0 | 0 | 0 | 0 | 3 | 3 | 7 | 59 | 17 | 14 | 0 | 0 | 52 |
| | **Público** | 34 | 0 | 0 | - | 7 | 14 | 0 | 93 | 0 | 3 | 0 | 0 | 3 | 7 | 17 | 45 | 14 | 14 | 3 | 0 | 45 |
| **Weekly** | **Expresso** | 28 | 0 | 0 | 3 | - | 17 | 0 | 90 | 0 | 3 | 0 | 0 | 7 | 14 | 10 | 59 | 14 | 14 | 0 | 0 | 41 |
| | **Sábado** | 14 | 0 | 3 | 7 | 7 | - | 0 | 93 | 3 | 14 | 0 | 0 | 17 | 10 | 10 | 41 | 7 | 21 | 3 | 3 | 45 |
| | **Sol** | 17 | 3 | 0 | 0 | 17 | 10 | 0 | 90 | 0 | 3 | 0 | 0 | 3 | 3 | 7 | 66 | 21 | 14 | 0 | 0 | 45 |
| | **Visão** | 34 | 0 | 3 | 0 | 10 | 21 | 0 | - | 0 | 17 | 0 | 0 | 14 | 17 | 21 | 48 | 10 | 28 | 3 | 3 | 69 |
| **Radio and TV** | **Antena 1** | 28 | 0 | 0 | 0 | 7 | 10 | 0 | 90 | - | 0 | 0 | 0 | 0 | 3 | 14 | 69 | 17 | 14 | 3 | 0 | 45 |
| | **Renascença** | 14 | 0 | 0 | 0 | 7 | 7 | 0 | 93 | 0 | - | 0 | 0 | 3 | 7 | 14 | 62 | 21 | 14 | 0 | 3 | 55 |
| | **RTP** | 41 | 0 | 0 | 0 | 14 | 14 | 0 | 93 | 0 | 3 | - | 0 | 3 | 7 | 7 | 45 | 10 | 14 | 3 | 0 | 45 |
| | **TSF** | 21 | 0 | 0 | 0 | 7 | 14 | 0 | 90 | 0 | 0 | 0 | - | 10 | 7 | 0 | 69 | 14 | 14 | 3 | 0 | 52 |
| | **TVI24** | 21 | 0 | 0 | 0 | 7 | 7 | 0 | 90 | 0 | 3 | 0 | 0 | - | 3 | 17 | 72 | 17 | 14 | 0 | 0 | 48 |
| **Sports** | **A Bola** | 38 | 0 | 0 | 3 | 24 | 7 | 0 | 93 | 0 | 14 | 0 | 0 | 24 | - | 0 | 10 | 3 | 24 | 3 | 0 | 55 |
| | **O Jogo** | 24 | 0 | 3 | 0 | 24 | 10 | 0 | 93 | 0 | 17 | 0 | 0 | 28 | 3 | - | 14 | 7 | 21 | 0 | 0 | 55 |
| | **MaisFutebol** | 31 | 0 | 3 | 0 | 10 | 14 | 0 | 93 | 3 | 14 | 0 | 0 | 28 | 3 | 0 | - | 10 | 28 | 3 | 3 | 55 |
| | **Record** | 31 | 0 | 0 | 10 | 17 | 7 | 0 | 90 | 3 | 24 | 0 | 0 | 28 | 3 | 7 | 28 | - | 14 | 0 | 0 | 38 |
| **E&F** | **Económico** | 31 | 0 | 0 | 1 | 7 | 7 | 0 | 86 | 0 | 7 | 0 | 0 | 10 | 7 | 10 | 76 | 17 | - | 3 | 0 | 34 |
| | **Jornal de Negócios** | 45 | 0 | 0 | 0 | 3 | 10 | 0 | 90 | 0 | 0 | 0 | 0 | 0 | 7 | 14 | 66 | 10 | 10 | - | 0 | 45 |
| **Other** | **Observador** | 31 | 0 | 0 | 0 | 7 | 10 | 0 | 90 | 0 | 7 | 0 | 0 | 7 | 3 | 7 | 69 | 14 | 14 | 3 | - | 38 |
| | **P3** | 48 | 0 | 3 | 0 | 7 | 7 | 3 | 93 | 0 | 10 | 0 | 0 | 10 | 7 | 7 | 72 | 0 | 24 | 7 | 0 | - |

sources that were never selected within the first three iterations of the subfeed. We can see a significant rise of most sources belonging to the weekly and sports groups.

The three sources that were most prevalent in the previous experiment - *Expresso*, *Visão*, and *P3* - have all reacted differently to the sampling process. *Expresso* has reduced significantly its appearance in the first three selected sources, in such way that it has now results similar to most other sources. This is more of what is expected from this news source, and is coincident with the its other results obtained after sampling. It is perhaps the source that reacted the best to sampling. *Visão*, on the other hand, was basically unaffected by the sampling process, which is also consistent with the its other results until now. It is safe to conclude at this point that the number of articles produced by this source is simply too low for this for this algorithm to obtain good results in its comparisons. *P3* still appears often in the early stages of subfeed construction, but has seen its percentage reduced to about half. These are more reasonable values for a source to have, and is another good indication in favor of the sampling process.

The overall drop of percentage seen in the top-appearing sources in the previous experiment allowed not only for most sources to have sporadic appearances, but also to some sources to considerably increase their percentages. The two most notable ones are *MaisFutebol* and *Correio da Manhã*. It is only natural that a sports source is consistently included in most subfeeds, since this group is consistently dissimilar from the others. Given that *MaisFutebol* is the source from this group that tends to have the lowest similarity values when compared to other sources, it is only natural that it is the one that is selected the most. This rise of the sports sources' inclusion percentage is another point in favor of the sampling. The incease of *Correio da Manhã*'s percentage is also consistent with its other results obtained after sampling.

These results still do not permit us to form groups of complementary news sources more than the results obtained in Sections 5.2 and 5.3. What we conclude is that the inclusion of sources from different thematic groups is a good way to obtain a varied news feed. These results, however, show that the sampling process is a great tool to help the construction of subfeeds and for its possible application in systems that construct feeds with reduced redundancy for users.

## 5.4 Conclusions

In this chapter, we were able to apply the strategies for news feed comparison that we had previously defined in a more realistic environment - the Portuguese news landscape. We conducted a series of experiments not only to try to gain a deeper understanding of this news panorama, but also to better understand the strengths and weakness of the developed algorithms.

First, we were able to detect some limitations of the implemented strategies related to the size of the datasets. The comparison of feeds containing articles from large periods of time, such as a week or a month, caused an increase of the average similarity values. This bloating of the results that is a consequence of the increased amount of content and greater chance of overlap proved to hinder the ability to compare sources. Averaging results obtained from daily comparisons proved to be a much more effective way to perform long-term analysis. The algorithms also proved to be very sensible to discrepancies in the number of articles produced by each source. Sources that have fewer articles have consistently lower similarity values. Merging this gap using a sampling process proved to be very helpful to mitigate these differences.

Finally, we were able to define news source groups withing the Portuguese media landscape based on the content each source produces. These groups proved to be good tools to better understand the diversity of the landscape, as well as a way to predict the results comparison. We were also able to identify the usefulness of subfeeds of the Global National Feed as a way to emulate the overall news panorama without the natural redundancy usually present. This may have applications in news suggesting systems.

# Chapter 6

# Conclusions and Further Work

In this dissertation, we have defined, developed, and tested different approached for the comparison of news feeds. We decided to focus on these strategies because, on our research of other systems that cover similarity of news articles, the vast majority of them focus on the comparison of the articles themselves. There is still very little research done in the comparison of the feeds themselves.

In this research, we also discovered that many of the systems that compare news articles do so using text similarity techniques. We then found out that the variety of such techniques far surpasses the ones that are used in such comparisons, and that there is little research done on the compared effectiveness of such measures in the evaluation of similarity of news articles. We would later observed that a great number of these text similarity measures behave very similar to each other and that their use is mostly interchangeable. It proved to be much more impactful the strategy used to handle the structure of a news feed rather than the comparison of the articles themselves. Of course, there is still value in a deeper comparative analysis of different text similarity measures and their impact on the comparison of news articles, but this should not be a priority.

Given the relative low impact that the text similarity measures proved to have in the overall results, the focus of our work turned to the feed comparison strategies. Given the sparse available research feed comparison, we defined four new measures to compare news feeds - Feed Concatenation, Cross Comparison, Cross Comparison with Elimination, and Cross Comparison with MRR. Initial testing in controlled environments, such as small feeds or artificial ones, showed that the first three measures not only behaved similarly, but also showed good results that matched our expectations. The latter measure proved to be ineffective on its own, even if it provided some insight on the feeds characteristics if used as a support algorithm. These algorithms, however, are still in their early stages and need refinement, which is why we believe that further research is more useful in the near future. Feed Concatenation proved to be very sensitive to uneven datasets or overall large amounts of data, which is likely a product of its *bag of words* nature. The remaining algorithms were not tested in such large-scale environments due to their temporal and spacial complexity. It is necessary to design and test improved versions of these algorithms, that are able to perform well in real-world environments, where data is not always balanced, while having an acceptable temporal and spacial complexity. Nevertheless, the obtained results are promising and lay a foundation from which further work can be done.

Finally, and more than conclusions about efficacy of similarity algorithms in a vacuum, we can also suggest some practical applications for their use. In particular, the construction of subfeeds of a global feed can be a great tool to help reducing redundancy in content feeds generated by feed aggregators. We found that it is possible to generate feeds by merging multiple sources from among the sources available that are capable of emulating the feed that contains all sources. The fact that it leaves some sources unused while presenting similar content indicates that some of the redundancy present in the global feed was removed. This could be the foundation for systems that suggest sets of news sources to a user. Moreover, we were able to generate these subfeeds starting from any set of different sources, which means that it could be customizable by the user to include some sources that they proritize. It is also important to note that none of the developed strategies, despite being tested only with Portuguese news items, are not limited to the Portuguese language. Only the preprocessing, which can be modified to accommodate other language, is specific to Portuguese. These strategies could easily be extended and utilized with any language.

## Appendix A

# News Feeds for User Study

Table A.1: News feeds for user study - Pair 1.

| Expresso | Jornal de Notícias |
|---|---|
| **Plano de recuperação português "precisa de reforço nas reformas", diz vice da Comissão Europeia**<br>Valdis Dombrovskis diz que Portugal está "muito ativo" a negociar o plano de recuperação com Bruxelas e quanto mais rápido foram aprovados os planos, mais rápido será o dinheiro distribuído. | **Liga dos Bombeiros diz que há doentes horas retidos nas macas em ambulâncias e que já houve uma morte**<br>O presidente da Liga dos Bombeiros Portugueses denunciou, este sábado, que há doentes transportados para os hospitais a passar "horas nas macas das ambulâncias", tendo sido já registada a morte de um paciente dentro da ambulância sem entrar na unidade hospitalar. |
| **Alemanha. Aliado de Merkel eleito presidente dos democratas-cristãos**<br>Calmo, simpático e "terra à terra", Armin Laschet é uma boa notícia para o Governo português. Ministro-presidente do estado mais populoso desde 2017, domina o aparelho do partido, tem experiência como vice-chanceler e uma sensibilidade reconhecida aos assuntos sociais. Será muito provavelmente o candidato a chanceler pela CDU nas legislativas de setembro próximo. | **Armin Laschet é o novo líder da CDU de Angela Merkel**<br>O presidente do estado alemão da Renânia do Norte - Vestefália, Armin Laschet, foi este sábado eleito o novo líder da CDU, força política à qual pertence a chanceler Angela Merkel, derrotando Friedrich Merz numa votação em congresso. |
| **Bombeiros dizem que morreu uma pessoa numa ambulância nas filas dos hospitais**<br>O presidente da Liga dos Bombeiros Portugueses denunciou hoje que há doentes transportados para os hospitais a passar "horas nas macas das ambulâncias", tendo sido já registada a morte de um paciente dentro da ambulância sem entrar na unidade hospitalar. | **Confinamento não reduziu movimento nas marginais do Porto, Gaia e Matosinhos** A regra obriga ao recolhimento no lar. Ainda assim, no primeiro sábado do novo confinamento, manteve-se o movimento nas marginais de Gaia, Porto e Matosinhos. De bicicleta, a correr, ou simplesmente a caminhar, várias foram as pessoas que passearam à beira mar esta manhã. Um cenário diferente do verificado no primeiro confinamento, decretado em março. |
| **Portugueses em casa, mas pouco. Apenas 39,5% estão mesmo confinados** Os dados da mobilidade medidos pela consultora PSE, indicam que os portugueses cumpriram "de forma muito ligeira" a orientação para "ficar em casa". Na verdade, o confinamento em casa de março de 2020 foi muitíssimo superior (61% vs. 39,5%) ao nível atingido desta vez, pelo menos neste primeiro dia. | **Plano de recuperação português "precisa de reforço nas reformas"** O vice-presidente da Comissão Europeia Valdis Dombrovskis considerou, em entrevista à Lusa, que o esboço do plano de recuperação e resiliência português precisa de reforço na área das reformas, salvaguardando porém que ainda não está fechado. |
| **Marcelo volta a testar negativo à covid-19** Presidente da República realizou mais um teste este sábado de manhã. | **Presidente da República volta a testar negativo à covid-19** O presidente da República, Marcelo Rebelo de Sousa, voltou a testar negativo à covid-19. |

Table A.2: News feeds for user study - Pair 2.

| Expresso | Observador |
|---|---|
| **Plano de recuperação português "precisa de reforço nas reformas", diz vice da Comissão Europeia**<br>Valdis Dombrovskis diz que Portugal está "muito ativo" a negociar o plano de recuperação com Bruxelas e quanto mais rápido foram aprovados os planos, mais rápido será o dinheiro distribuído. | **Meio confinamento**<br>Este meio confinamento não dá sentido ao nosso sofrimento. O que se perde, e quem perde nesta discriminação sem paridade, não justifica o pouco efeito que esta meia decisão entre Deus e o diabo trará. |
| **Alemanha. Aliado de Merkel eleito presidente dos democratas-cristãos**<br>Calmo, simpático e "terra à terra", Armin Laschet é uma boa notícia para o Governo português. Ministro-presidente do estado mais populoso desde 2017, domina o aparelho do partido, tem experiência como vice-chanceler e uma sensibilidade reconhecida aos assuntos sociais. Será muito provavelmente o candidato a chanceler pela CDU nas legislativas de setembro próximo. | **UE/Presidência. "Portugal vai pôr Índia na agenda política da UE", diz Josep Borrel**<br>O chefe da diplomacia europeia diz que é difícil alcançar um acordo comercial com a Índia mas classifica como "um grande passo" a intenção da presidência portuguesa de dar prioridade a esse dossiê. |
| **Bombeiros dizem que morreu uma pessoa numa ambulância nas filas dos hospitais**<br>O presidente da Liga dos Bombeiros Portugueses denunciou hoje que há doentes transportados para os hospitais a passar "horas nas macas das ambulâncias", tendo sido já registada a morte de um paciente dentro da ambulância sem entrar na unidade hospitalar. | **Torres Vedras. Longas filas de ambulâncias na urgência Covid duram horas** O hospital local viveu um pesadelo esta sexta-feira com uma longa fila de ambulâncias paradas em frente às urgências. A triagem dos doentes com Covid-19 chegou a ser feita dentro das ambulâncias. |
| **Portugueses em casa, mas pouco. Apenas 39,5% estão mesmo confinados** Os dados da mobilidade medidos pela consultora PSE, indicam que os portugueses cumpriram "de forma muito ligeira" a orientação para "ficar em casa". Na verdade, o confinamento em casa de março de 2020 foi muitíssimo superior (61% vs. 39,5%) ao nível atingido desta vez, pelo menos neste primeiro dia. | **Armin Laschet é o novo líder da CDU de Angela Merkel** O presidente do estado alemão da Renânia do Norte - Vestefália foi eleito o novo líder da CDU, derrotando Friedrich Merz numa votação em congresso. Será assim o sucessor de Angela Merkel. |
| **Marcelo volta a testar negativo à covid-19** Presidente da República realizou mais um teste este sábado de manhã. | **Teste a Marcelo. "Negativo? Sabem primeiro que eu"** De visita à Santa Casa da Misericórdia do Barreiro, Marcelo Rebelo de Sousa revelou o resultado de mais um teste à Covid-19. Mas, ainda antes de o fazer, já a informação circulava. "Não incomoda nada." |

Table A.3: News feeds for user study - Pair 3.

| **Expresso** | **Público** |
|---|---|
| **Plano de recuperação português "precisa de reforço nas reformas", diz vice da Comissão Europeia**<br>Valdis Dombrovskis diz que Portugal está "muito ativo" a negociar o plano de recuperação com Bruxelas e quanto mais rápido foram aprovados os planos, mais rápido será o dinheiro distribuído. | **Mundo ultrapassa dois milhões de mortes por covid-19. OMS rejeita passaporte para vacinados**<br>OMS apelou aos países que não exigissem provas de vacinação contra a covid-19 para que seja autorizada a entrada dos viajantes “uma vez que ainda se desconhece o impacto das vacinas na redução da transmissão e que a disponibilidade actual de vacinas é demasiado limitada”. Em todo o mundo, o coronavírus SARS-CoV-2 já infectou mais de 93,8 milhões de pessoas e matou dois milhões. |
| **Alemanha. Aliado de Merkel eleito presidente dos democratas-cristãos**<br>Calmo, simpático e "terra à terra", Armin Laschet é uma boa notícia para o Governo português. Ministro-presidente do estado mais populoso desde 2017, domina o aparelho do partido, tem experiência como vice-chanceler e uma sensibilidade reconhecida aos assuntos sociais. Será muito provavelmente o candidato a chanceler pela CDU nas legislativas de setembro próximo. | **Apertar os círculos da economia circular: os resíduos orgânicos como oportunidade**<br>Não basta a economia circular – há que atingir círculos cada vez mais fechados, assentes em cadeias de produção-consumo curtas e soluções de reutilização e reciclagem mais localizadas. |
| **Bombeiros dizem que morreu uma pessoa numa ambulância nas filas dos hospitais**<br>O presidente da Liga dos Bombeiros Portugueses denunciou hoje que há doentes transportados para os hospitais a passar "horas nas macas das ambulâncias", tendo sido já registada a morte de um paciente dentro da ambulância sem entrar na unidade hospitalar. | **Alemanha: Armin Laschet é o novo líder da CDU**<br>O candidato que representa a linha de continuidade de Angela Merkel venceu a segunda volta das eleições do partido conservador alemão. |
| **Portugueses em casa, mas pouco. Apenas 39,5% estão mesmo confinados** Os dados da mobilidade medidos pela consultora PSE, indicam que os portugueses cumpriram “de forma muito ligeira” a orientação para “ficar em casa”. Na verdade, o confinamento em casa de março de 2020 foi muitíssimo superior (61% vs. 39,5%) ao nível atingido desta vez, pelo menos neste primeiro dia. | **Portugueses cumpriram ordem para ficar em casa de forma “muito ligeira”** Nesta sexta-feira, o primeiro dia do novo confinamento, só 39,5% dos portugueses estiveram confinados em casa. Durante o confinamento de Março e Abril, a média de pessoas que ficaram em casa foi de 61%. |
| **Marcelo volta a testar negativo à covid-19** Presidente da República realizou mais um teste este sábado de manhã. | **Portugueses obedeceram à ordem para confinar em 2020. E agora?** Dados sobre ao confinamento médio durante o terrível ano de 2020 mostram que os portugueses não só cumprem as regras como às vezes até se antecipam às medidas. Porém, o “pára-arranca” de Novembro e Dezembro com medidas pontuais terá tornado as pessoas mais desconfiadas e, por isso, mais desconfinadas. |

Table A.4: News feeds for user study - Pair 4.

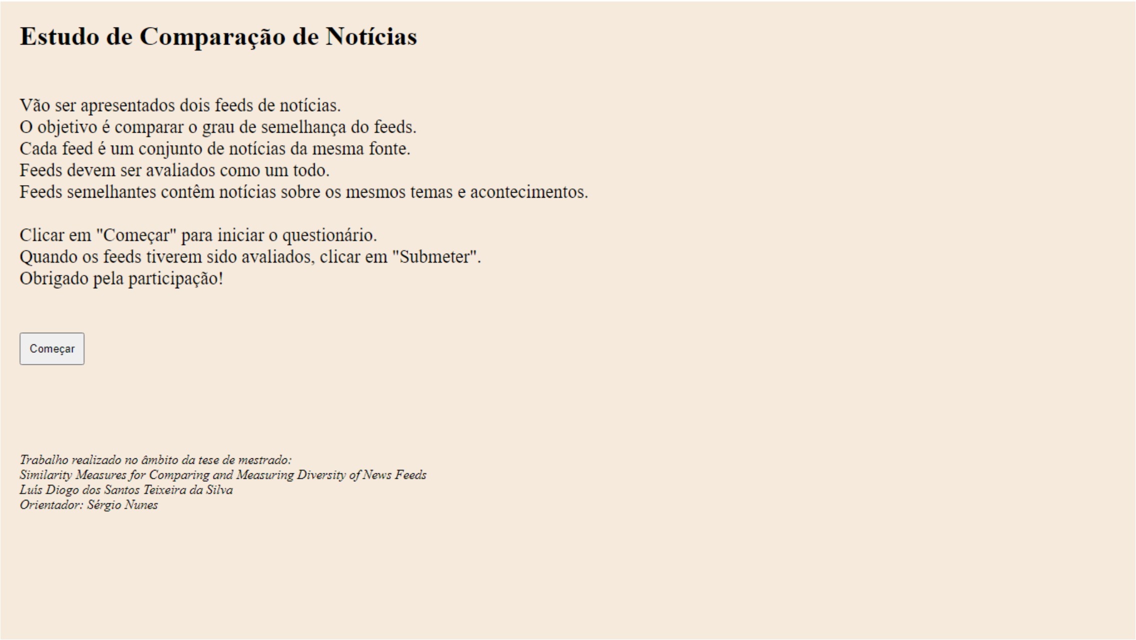
| **Expresso** | **iOnline** |
|---|---|
| **Plano de recuperação português "precisa de reforço nas reformas", diz vice da Comissão Europeia**<br>Valdis Dombrovskis diz que Portugal está "muito ativo" a negociar o plano de recuperação com Bruxelas e quanto mais rápido foram aprovados os planos, mais rápido será o dinheiro distribuído. | **Confinamento de um mês? 'Não vai chegar'. Há 200 vezes mais casos ativos do que em Março**<br>Na próxima semana, os cuidados intensivos podem passar os 700 doentes com covid-19. Internados podem subir para 6 mil. Especialistas temem que planalto de casos reportados seja artificial e se deva a limites na capacidade de testagem. Com milhares de inquéritos epidemiológicos pendentes, a região de Lisboa já está a ter o apoio de 200 militares nas chamadas para infetados que permitem alertar e isolar os contactos com maior risco de terem ficado infetados e ajudar a travar cadeias de transmissão. |
| **Alemanha. Aliado de Merkel eleito presidente dos democratas-cristãos**<br>Calmo, simpático e "terra à terra", Armin Laschet é uma boa notícia para o Governo português. Ministro-presidente do estado mais populoso desde 2017, domina o aparelho do partido, tem experiência como vice-chanceler e uma sensibilidade reconhecida aos assuntos sociais. Será muito provavelmente o candidato a chanceler pela CDU nas legislativas de setembro próximo. | **Covid-19. Campanha de vacinação atrasa-se**<br>Pfizer vai reduzir entregas das próximas semanas. Após investigar 13 mortes de idosos vacinados, Noruega alerta que reações comuns podem ser um risco para pessoas muito frágeis e aconselha ponderação na imunização. |
| **Bombeiros dizem que morreu uma pessoa numa ambulância nas filas dos hospitais**<br>O presidente da Liga dos Bombeiros Portugueses denunciou hoje que há doentes transportados para os hospitais a passar "horas nas macas das ambulâncias", tendo sido já registada a morte de um paciente dentro da ambulância sem entrar na unidade hospitalar. | **Nunca se morreu tanto em Portugal** Não há registos de tantos dias seguidos com mais de 500 mortes. Frio pode passar, mas não se prevê o pico na covid-19. |
| **Portugueses em casa, mas pouco. Apenas 39,5% estão mesmo confinados** Os dados da mobilidade medidos pela consultora PSE, indicam que os portugueses cumpriram "de forma muito ligeira" a orientação para "ficar em casa". Na verdade, o confinamento em casa de março de 2020 foi muitíssimo superior (61% vs. 39,5%) ao nível atingido desta vez, pelo menos neste primeiro dia. | **Procurador europeu na agenda de Bruxelas** Plenário debate no próximo dia 21 o caso, horas depois de Portugal apresentar o seu plano de atividades para o semestre na União Europeia. |
| **Marcelo volta a testar negativo à covid-19** Presidente da República realizou mais um teste este sábado de manhã. | **Livraria Campos Trindade fecha as portas** Encerramento deve-se à dificuldade de comportar uma renda elevada numa altura de fortes quebras devido à pandemia. Proprietário anunciou intenção de reabrir noutro espaço. |

Table A.5: News feeds for user study - Pair 5.

| Expresso | TSF |
|---|---|
| **Plano de recuperação português "precisa de reforço nas reformas", diz vice da Comissão Europeia**<br>Valdis Dombrovskis diz que Portugal está "muito ativo" a negociar o plano de recuperação com Bruxelas e quanto mais rápido foram aprovados os planos, mais rápido será o dinheiro distribuído. | **EDP antecipa pagamento a mais de mil fornecedores em Portugal e Espanha**<br>A EDP explica que esta decisão de avançar com os pagamentos a pronto "pretende ser um apoio no atual contexto de pandemia para garantir que essas empresas têm liquidez para pagar salários e mantêm a sua atividade". |
| **Alemanha. Aliado de Merkel eleito presidente dos democratas-cristãos**<br>Calmo, simpático e "terra à terra", Armin Laschet é uma boa notícia para o Governo português. Ministro-presidente do estado mais populoso desde 2017, domina o aparelho do partido, tem experiência como vice-chanceler e uma sensibilidade reconhecida aos assuntos sociais. Será muito provavelmente o candidato a chanceler pela CDU nas legislativas de setembro próximo. | **Armin Laschet sucede a Angela Merkel à frente da CDU**<br>Armin Laschet bateu Friedrich Merz, candidato mais à direita que já tinha concorrido à liderança dos democratas-cristãos alemães em 2018. |
| **Bombeiros dizem que morreu uma pessoa numa ambulância nas filas dos hospitais**<br>O presidente da Liga dos Bombeiros Portugueses denunciou hoje que há doentes transportados para os hospitais a passar "horas nas macas das ambulâncias", tendo sido já registada a morte de um paciente dentro da ambulância sem entrar na unidade hospitalar. | **Noite difícil em Torres Vedras. Fila de espera de ambulâncias e urgência lotada** Em Torres Vedras, "a situação é mais preocupante por causa do surto ativo [com um total de 157 casos confirmados] dentro da unidade". |
| **Portugueses em casa, mas pouco. Apenas 39,5% estão mesmo confinados** Os dados da mobilidade medidos pela consultora PSE, indicam que os portugueses cumpriram “de forma muito ligeira” a orientação para “ficar em casa”. Na verdade, o confinamento em casa de março de 2020 foi muitíssimo superior (61% vs. 39,5%) ao nível atingido desta vez, pelo menos neste primeiro dia. | **No primeiro dia de confinamento, 60% das pessoas saíram à rua** No primeiro dia de confinamento, menos de 40% da população ficou em casa, e a empresa PSE, que faz estudos de mobilidade, não espera que os níveis de confinamento atinjam as percentagens de março ou abril. |
| **Marcelo volta a testar negativo à covid-19** Presidente da República realizou mais um teste este sábado de manhã. | **Marcelo volta a testar negativo ao coronavírus** Marcelo Rebelo de Sousa realizou nos últimos dias seis testes, dos quais apenas um resultou em positivo para a Covid-19. |

# Appendix B

# Interface and Instructions for User Study



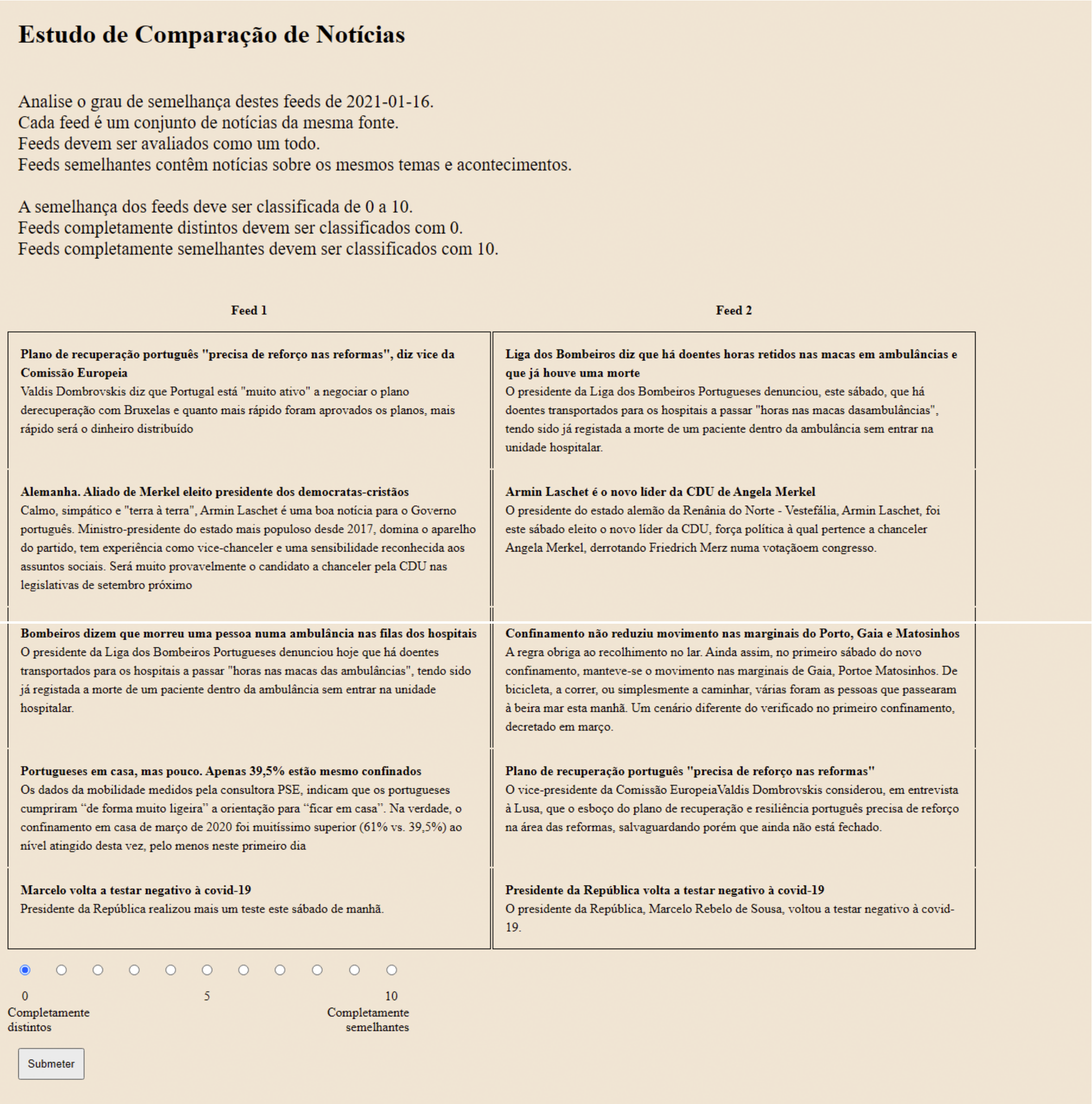
**Estudo de Comparação de Notícias**

Vão ser apresentados dois feeds de notícias.
O objetivo é comparar o grau de semelhança do feeds.
Cada feed é um conjunto de notícias da mesma fonte.
Feeds devem ser avaliados como um todo.
Feeds semelhantes contêm notícias sobre os mesmos temas e acontecimentos.

Clicar em "Começar" para iniciar o questionário.
Quando os feeds tiverem sido avaliados, clicar em "Submeter".
Obrigado pela participação!

*Trabalho realizado no âmbito da tese de mestrado:*
*Similarity Measures for Comparing and Measuring Diversity of News Feeds*
*Luís Diogo dos Santos Teixeira da Silva*
*Orientador: Sérgio Nunes*

Figure B.1: User study welcoming page.



### Estudo de Comparação de Notícias

Analise o grau de semelhança destes feeds de 2021-01-16.
Cada feed é um conjunto de notícias da mesma fonte.
Feeds devem ser avaliados como um todo.
Feeds semelhantes contêm notícias sobre os mesmos temas e acontecimentos.

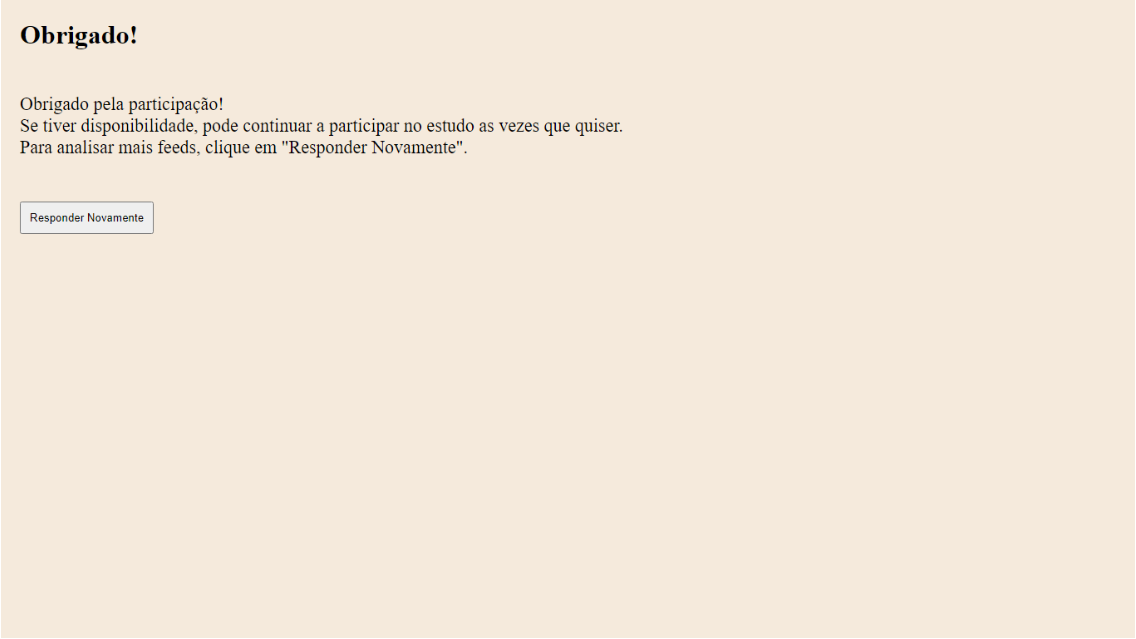
A semelhança dos feeds deve ser classificada de 0 a 10.
Feeds completamente distintos devem ser classificados com 0.
Feeds completamente semelhantes devem ser classificados com 10.

| Feed 1 | Feed 2 |
|---|---|
| **Plano de recuperação português "precisa de reforço nas reformas", diz vice da Comissão Europeia**<br>Valdis Dombrovskis diz que Portugal está "muito ativo" a negociar o plano derecuperação com Bruxelas e quanto mais rápido foram aprovados os planos, mais rápido será o dinheiro distribuído | **Liga dos Bombeiros diz que há doentes horas retidos nas macas em ambulâncias e que já houve uma morte**<br>O presidente da Liga dos Bombeiros Portugueses denunciou, este sábado, que há doentes transportados para os hospitais a passar "horas nas macas dasambulâncias", tendo sido já registada a morte de um paciente dentro da ambulância sem entrar na unidade hospitalar. |
| **Alemanha. Aliado de Merkel eleito presidente dos democratas-cristãos**<br>Calmo, simpático e "terra à terra", Armin Laschet é uma boa notícia para o Governo português. Ministro-presidente do estado mais populoso desde 2017, domina o aparelho do partido, tem experiência como vice-chanceler e uma sensibilidade reconhecida aos assuntos sociais. Será muito provavelmente o candidato a chanceler pela CDU nas legislativas de setembro próximo | **Armin Laschet é o novo líder da CDU de Angela Merkel**<br>O presidente do estado alemão da Renânia do Norte - Vestefália, Armin Laschet, foi este sábado eleito o novo líder da CDU, força política à qual pertence a chanceler Angela Merkel, derrotando Friedrich Merz numa votaçãoem congresso. |
| **Bombeiros dizem que morreu uma pessoa numa ambulância nas filas dos hospitais**<br>O presidente da Liga dos Bombeiros Portugueses denunciou hoje que há doentes transportados para os hospitais a passar "horas nas macas das ambulâncias", tendo sido já registada a morte de um paciente dentro da ambulância sem entrar na unidade hospitalar. | **Confinamento não reduziu movimento nas marginais do Porto, Gaia e Matosinhos**<br>A regra obriga ao recolhimento no lar. Ainda assim, no primeiro sábado do novo confinamento, manteve-se o movimento nas marginais de Gaia, Portoe Matosinhos. De bicicleta, a correr, ou simplesmente a caminhar, várias foram as pessoas que passearam à beira mar esta manhã. Um cenário diferente do verificado no primeiro confinamento, decretado em março. |
| **Portugueses em casa, mas pouco. Apenas 39,5% estão mesmo confinados**<br>Os dados da mobilidade medidos pela consultora PSE, indicam que os portugueses cumpriram "de forma muito ligeira" a orientação para "ficar em casa". Na verdade, o confinamento em casa de março de 2020 foi muitíssimo superior (61% vs. 39,5%) ao nível atingido desta vez, pelo menos neste primeiro dia | **Plano de recuperação português "precisa de reforço nas reformas"**<br>O vice-presidente da Comissão EuropeiaValdis Dombrovskis considerou, em entrevista à Lusa, que o esboço do plano de recuperação e resiliência português precisa de reforço na área das reformas, salvaguardando porém que ainda não está fechado. |
| **Marcelo volta a testar negativo à covid-19**<br>Presidente da República realizou mais um teste este sábado de manhã. | **Presidente da República volta a testar negativo à covid-19**<br>O presidente da República, Marcelo Rebelo de Sousa, voltou a testar negativo à covid-19. |

● ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

0 5 10
Completamente distintos Completamente semelhantes

Figure B.2: User study news feed page.



**Obrigado!**

Obrigado pela participação!
Se tiver disponibilidade, pode continuar a participar no estudo as vezes que quiser.
Para analisar mais feeds, clique em "Responder Novamente".

Figure B.3: User study end page.

# References

[1] Facebook API. Accessed Dec. 5, 2020. URL: https://developers.facebook.com/docs/graph-api/.

[2] Google News. Accessed Dec. 12, 2020. URL: http://news.google.com.

[3] Media Cloud. Accessed Dec. 3, 2020. URL: https://mediacloud.org/.

[4] Media Cloud API. Accessed Dec. 3, 2020. URL: https://github.com/mediacloud/backend/blob/master/doc/api_2_0_spec/api_2_0_spec.md.

[5] Natural Language Toolkit. Accessed Mar. 8, 2021. URL: https://www.nltk.org/.

[6] News Media Lists: Directory of world news & digital media. Accessed Dec. 8, 2020. URL: https://www.newsmedialists.com/.

[7] RSS Specification. Accessed Jan. 10, 2021. URL: https://www.rssboard.org/rss-specification.

[8] The Atom Syndication Format. Accessed Jan. 27, 2021. URL: https://tools.ietf.org/html/rfc4287.

[9] Twitter API. Accessed Dec. 5, 2020. URL: https://developer.twitter.com/en/docs/twitter-api.

[10] What is RSS? RSS explained. Accessed Dec. 3, 2020. URL: http://www.whatisrss.com/.

[11] Shikha Agarwal, Archana Singhal, and Punam Bedi. Classification of RSS feed news items using ontology. In *2012 12th International Conference on Intelligent Systems Design and Applications (ISDA)*, pages 491–496, 2012. doi:10.1109/ISDA.2012.6416587.

[12] Julie Kane Ahkter and Steven Soria. Sentiment analysis: Facebook status messages. *Master's thesis, Stanford, CA*, 2010.

[13] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. *WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web*, 01 2012. doi:10.1145/2187836.2187907.

[14] Satanjeev Banerjee and Ted Pedersen. An adapted Lesk algorithm for word sense disambiguation using WordNet. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 136–145, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

[15] Alberto Barrón-Cedeno, Paolo Rosso, Eneko Agirre, and Gorka Labaka. Plagiarism detection across distant language pairs. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 37–45, 2010.

[16] Anja Bechmann and Kristoffer L. Nielbo. Are We Exposed to the Same "News" in the News Feed? An empirical analysis of filter bubbles as information similarity for Danish Facebook users. *Digital journalism*, 6(8):990–1002, 2018.

[17] Kathleen Beckers, Andrea Masini, Julie Sevenans, Miriam van der Burg, Julie De Smedt, Hilde Van den Bulck, and Stefaan Walgrave. Are newspapers' news stories becoming more alike? Media content diversity in Belgium, 1983–2013. *Journalism*, 20(12):1665–1683, 2019. doi:10.1177/1464884917706860.

[18] Yochai Benkler, Hal Roberts, Robert Faris, Alicia Solow-Niederman, and Bruce Etling. Social mobilization and the networked public sphere: Mapping the SOPA-PIPA debate. *Political Communication*, 32(4):594–624, 2015. doi:10.1080/10584609.2014.986349.

[19] Sonia Bergamaschi, Francesco Guerra, Mirko Orsini, Claudio Sartori, and Maurizio Vincini. Relevant News: a semantic news feed aggregator. In *Semantic Web Applications and Perspectives*, volume 314, pages 150–159. Giovanni Semeraro, Eugenio Di Sciascio, Christian Morbidoni, Heiko Stoemer, 2007.

[20] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[21] Andrei Z. Broder, Moses Charikar, Alan M. Frieze, and Michael Mitzenmacher. Min-wise independent permutations. *Journal of Computer and System Sciences*, 60(3):630–659, 2000.

[22] Kate Bulkley. The rise of citizen journalism. *The Guardian*, Jun 2012. Accessed Dec. 1, 2020. URL: https://www.theguardian.com/media/2012/jun/11/rise-of-citizen-journalism.

[23] Adam Burkepile and Perry Fizzano. Classifying RSS feeds with an artificial immune system. In *2010 Second International Conference on Information, Process, and Knowledge Management*, pages 43–47, 2010. doi:10.1109/eKNOW.2010.19.

[24] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. Emerging topic detection on Twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*. Association for Computing Machinery, 2010. doi:10.1145/1814245.1814249.

[25] Dhivya Chandrasekaran and Vijay Mago. Evolution of semantic similarity - A survey. *CoRR*, abs/2004.13820, 2020.

[26] Tom Chatfield. *How to Thrive in the Digital Age*. Pan Macmillan, 2012.

[27] Jason Chuang, Sands Fish, David Larochelle, William P. Li, and Rebecca Weiss. Large-scale topical analysis of multiple online news sources with Media Cloud. *NewsKDD: Data Science for News Publishing, at KDD*, 2014.

[28] Hsiang Iris Chyi and Dominic Lasorsa. Access, use and preferences for online newspapers. *Newspaper Research Journal*, 20(4):2–13, 1999.

[29] R. L. Cilibrasi and P. M. B. Vitanyi. The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383, 2007. doi:10.1109/TKDE.2007.48.

[30] Abhinandan S. Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. Google news personalization: Scalable online collaborative filtering. In *Proceedings of the 16th International Conference on World Wide Web*, page 271–280. Association for Computing Machinery, 2007. doi:10.1145/1242572.1242610.

[31] Derek Davis, Gerardo Figueroa, and Yi-Shin Chen. SociRank: Identifying and Ranking Prevalent News Topics Using Social Media Factors. *IEEE Trans. Syst. Man Cybern. Syst.*, 47(6):979–994, 2017. doi:10.1109/TSMC.2016.2523932.

[32] José Luís Devezas and Sérgio Nunes. Social media and information consumption diversity. In Dyaa Albakour, David Corney, Julio Gonzalo, Miguel Martinez-Alvarez, Barbara Poblete, and Andreas Valochas, editors, *Proceedings of the Second International Workshop on Recent Trends in News Information Retrieval co-located with 40th European Conference on Information Retrieval (ECIR 2018), Grenoble, France, March 26, 2018*, volume 2079 of *CEUR Workshop Proceedings*, pages 18–23. CEUR-WS.org, 2018. URL: http://ceur-ws.org/Vol-2079/paper5.pdf.

[33] Tiago Devezas, Sérgio Nunes, and María Teresa Rodríguez. MediaViz: An Interactive Visualization Platform for Online Media Studies. In *Proceedings of the 2015 International Workshop on Human-Centric Independent Computing*, HIC '15, page 7–11, New York, NY, USA, 2015. Association for Computing Machinery. doi:10.1145/2808469.2808474.

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

[35] Peter Duffy and Axel Bruns. The use of blogs, wikis and RSS in education: A conversation of possibilities. In *Learning on the Move: Proceedings of the Online Learning and Teaching Conference 2006*, pages 31–38. Queensland University of Technology, 2006.

[36] Bruce Etling, Hal Roberts, and Robert Faris. Blogs as an alternative public sphere: The role of blogs, mainstream media, and TV in Russia's media ecology. *Berkman Center Research Publication*, (2014-8), 2014.

[37] Maurice Fabre. *Histoire de la Communication*. Edito Service, 1967.

[38] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *International Joint Conferences on Artificial Intelligence Organization*, volume 7, pages 1606–1611, 2007.

[39] Wael H. Gomaa and Aly A. Fahmy. A survey of text similarity approaches. *International Journal of Computer Applications*, 68(13):13–18, 2013.

[40] Erhardt Graeff, Matt Stempeck, and Ethan Zuckerman. The battle for 'Trayvon Martin': Mapping a media controversy online and off-line. *First Monday*, 19(2), 2014. doi:10.5210/fm.v19i2.4947.

[41] J. Grossnickle, T. Board, B. Pickens, and M Bellmont. RSS – Crossing into the Mainstream. *Yahoo! White Paper*, Oct 2005. Accessed Jan. 27, 2021. URL: https://content.marketingsherpa.com/heap/cs/rsscharts/7.pdf.

[42] Patrick A. V. Hall and Geoff R. Dowling. Approximate string matching. *ACM Comput. Surv.*, 12(4):381–402, December 1980. doi:10.1145/356827.356830.

[43] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques, 3rd edition*. Morgan Kaufmann, 2011.

[44] Mengting Han, Xuan Zhang, Xin Yuan, Jiahao Jiang, Wei Yun, and Chen Gao. A survey on the techniques, applications, and performance of short text semantic similarity. *Concurrency and Computation: Practice and Experience*, page e5971, 2020.

[45] Graeme Hirst and David St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An Electronic Lexical Database*, 305, 10 1995.

[46] Thomas Hofmann. Latent semantic models for collaborative filtering. *ACM Trans. Inf. Syst.*, 22(1):89–115, 2004. doi:10.1145/963770.963774.

[47] Robert W. Irving and Campbell B. Fraser. Two algorithms for the longest common subsequence of three (or more) strings. In Alberto Apostolico, Maxime Crochemore, Zvi Galil, and Udi Manber, editors, *Combinatorial Pattern Matching*, pages 214–229, Berlin, Heidelberg, 1992. Springer Berlin Heidelberg.

[48] Paul Jaccard. Etude de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 37:547–579, 01 1901. doi:10.5169/seals-266450.

[49] Matthew A. Jaro. Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84(406):414–420, 1989. doi:10.1080/01621459.1989.10478785.

[50] Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In Keh-Jiann Chen, Chu-Ren Huang, and Richard Sproat, editors, *Proceedings of the 10th Research on Computational Linguistics International Conference, ROCLING 1997, Taipei, Taiwan, August 1997*, pages 19–33. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), 1997.

[51] David Kansas and Todd Gitlin. What's the Rush: An e-epistolary debate on the 24 hour news clock. In Robert H. Giles and Robert W. Snyder, editors, *What's Next?: The Problems and Prospects of Journalism*, page 87. Transaction Publishers, 2001.

[52] Pankaj Karpe, Vijay Bhor, and Chetana Agarwal. News feed processing and analysis using Hadoop framework. *IJCA Proceedings on National Conference on Advances in Computing, Communication and Networking*, ACCNET 2016(1):16–18, June 2016.

[53] Thomas K. Landauer and Susan T. Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211, 1997.

[54] Claudia Leacock and Martin Chodorow. *Combining Local Context and WordNet Similarity for Word Sense Identification*, volume 49, pages 265–. 01 1998.

[55] Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International*

*Conference on Systems Documentation*, SIGDOC '86, page 24–26, New York, NY, USA, 1986. Association for Computing Machinery. doi:10.1145/318723.318728.

[56] Daniel J. Levitin. Why the modern world is bad for your brain. *The Guardian*, Jan 2015. Accessed Jan. 8, 2021. URL: https://www.theguardian.com/science/2015/jan/18/modern-world-bad-for-brain-daniel-j-levitin-organized-mind-information-overload.

[57] Ming Li, Jonathan H. Badger, Xin Chen, Sam Kwong, Paul Kearney, and Haoyong Zhang. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, 17(2):149–154, 02 2001. doi:10.1093/bioinformatics/17.2.149.

[58] Ye Liang, Ningning Guoa, Chunxiao Xing, Yong Zhang, and Chao Li. Multilingual information retrieval and smart news feed based on big data. In *2015 12th Web Information System and Application Conference (WISA)*, pages 85–88, 2015. doi:10.1109/WISA.2015.44.

[59] Dekang Lin. Extracting collocations from text corpora. In *Workshop on Computational Terminology*, pages 57–63, Montreal, Canada, 1998.

[60] Kevin Lund and Curt Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers*, 28(2):203–208, 1996.

[61] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[62] Panyamee Manusnanth and Somjit Arj-in. Document clustering results on the semantic web search. In *Proceedings of The 5th National Conference on Computing and Information Technology*, 2009.

[63] Andrew McDowall and Bibi van der Zee. 'Local media are simply disappearing': how financial pressures are killing independent media. *The Guardian*, Nov 2017. Accessed Dec. 1, 2020. URL: https://www.theguardian.com/media/2017/nov/30/closure-of-nepszabadsag-hungarian-daily-highlights-threat-to-independent-media.

[64] Rada Mihalcea and Paul Tarau. TextRank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004.

[65] George A. Miller. WordNet: A lexical database for English. *Commun. ACM*, 38(11):39–41, 1995. doi:10.1145/219717.219748.

[66] A. Niewiadomski. Appliance of fuzzy relations for text documents comparing. In *Proceedings of the 5th Conference on Neural Networks and Soft Computing, Zakopane, Poland*, pages 347–352, 2000.

[67] Viviane Moreira Orengo and Christian R. Huyck. A stemming algorithmm for the portuguese language. In Gonzalo Navarro, editor, *Eighth International Symposium on String Processing and Information Retrieval, SPIRE 2001, Laguna de San Rafael, Chile, November 13-15, 2001*, pages 186–193. IEEE Computer Society, 2001. doi:10.1109/SPIRE.2001.989755.

[68] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[69] Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. Using measures of semantic relatedness for word sense disambiguation. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 241–257, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[70] Michael Paul and Mark Dredze. You are what your tweet: Analyzing Twitter for public health. *Artificial Intelligence*, 38:265–272, 01 2011.

[71] Michael Paul and Mark Dredze. A model for mining public health topics from Twitter. *Health*, 11(1), 2012.

[72] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. WordNet:: Similarity-measuring the relatedness of concepts. In *Association for the Advancement of Artificial Intelligence*, volume 4, pages 25–29, 2004.

[73] James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of LIWC2015. Technical report, 2015.

[74] Andrew Perrin. Social Media Usage: 2005-2015. Washington, DC, 2015. Pew Res Cent.

[75] Tamara Polajnar, Richard Glassey, and Leif Azzopardi. Detection of news feeds items appropriate for children. In Ricardo Baeza-Yates, Arjen P. de Vries, Hugo Zaragoza, B. Barla Cambazoglu, Vanessa Murdock, Ronny Lempel, and Fabrizio Silvestri, editors, *Advances in Information Retrieval*, pages 63–72. Springer Berlin Heidelberg, 2012.

[76] Aurora Pons-Porrata, Rafael Berlanga-Llavori, and José Ruiz-Shulcloper. Topic discovery based on text mining techniques. *Information Processing & Management*, 43(3):752 – 768, 2007. Special Issue on Heterogeneous and Distributed IR. doi:https://doi.org/10.1016/j.ipm.2006.06.001.

[77] Martin F. Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137, 1980. doi:10.1108/eb046814.

[78] Martin Potthast, Benno Stein, and Maik Anderka. A wikipedia-based multilingual retrieval model. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White, editors, *Advances in Information Retrieval*, pages 522–530, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[79] Nitesh Pradhan, Manasi Gyanchandani, and Rajesh Wadhvani. A review on text similarity technique used in IR and its application. *International Journal of Computer Applications*, 120(9), 2015.

[80] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95, Montréal Québec, Canada, August 20-25 1995, 2 Volumes*, pages 448–453. Morgan Kaufmann, 1995.

[81] David J. Rogers and Taffee T. Tanimoto. A computer program for classifying plants. *Science*, 132(3434):1115–1118, 1960.

[82] Paul F. Russell and T. Ramachandra Rao. On habitat and association of species of anopheline larvae in south-eastern madras. *Journal of the Malaria Institute of India*, 3(1), 1940.

[83] Avani Sakhapara, Dipti Pawade, Hardik Chapanera, Harshal Jani, and Darpan Ramgaonkar. Segregation of similar and dissimilar live RSS news feeds based on similarity measures. In *Data Management, Analytics and Innovation*, pages 333–344. Springer, 2019.

[84] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. volume 24, pages 513–523. Elsevier, 1988.

[85] Shankar Setty, Rajendra Jadi, Sabya Shaikh, Chandan Mattikalli, and Uma Mudenagudi. Classification of Facebook news feeds and sentiment analysis. In *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 18–23, 2014. doi:10.1109/ICACCI.2014.6968447.

[86] Akash Shrivastava and Bhasker Pant. Opinion extraction and classification of real time Facebook Status. *Global Journal of Computer Science and Technology*, 2012.

[87] Helle Sjøvaag and Truls André Pedersen. The effect of direct press support on the diversity of news content in Norway. *Journal of Media Business Studies*, 15(4):300–316, 2018. doi:10.1080/16522354.2018.1546089.

[88] Jacob Soll. *The information master: Jean-Baptiste Colbert's secret state intelligence system*. University of Michigan Press, 2009.

[89] Miriam Steiner, Melanie Magin, and Birgit Stark. Uneasy Bedfellows: Comparing the diversity of German public service news on television and on Facebook. *Digital Journalism*, 7(1):100–123, 2019. doi:10.1080/21670811.2017.1412800.

[90] Mihai Tabara, Mihai Dascalu, and Stefan Trausan-Matu. Building a semantic recommendation engine for news feeds based on emerging topics from tweets. In *2016 15th RoEduNet Conference: Networking in Education and Research*, pages 1–5, 2016. doi:10.1109/RoEduNet.2016.7753209.

[91] Mike Thelwall and Rudy Prabowo. Identifying and characterizing public science-related fears from RSS feeds. *Journal of the American Society for Information Science and Technology*, 58(3):379–390, 2007. doi:https://doi.org/10.1002/asi.20504.

[92] Nava Tintarev, Emily Sullivan, Dror Guldin, Sihang Qiu, and Daan Odjik. Same, Same, but Different: Algorithmic diversification of viewpoints in news. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*, page 7–13. Association for Computing Machinery, 2018. doi:10.1145/3213586.3226203.

[93] Peter D. Turney. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In Luc De Raedt and Peter Flach, editors, *Machine Learning: ECML 2001*, pages 491–502, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

[94] Paul M. B. Vitányi and Rudi Cilibrasi. Normalized web distance and word similarity. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*, pages 293–314. Chapman and Hall/CRC, 2010.

[95] Ellen M. Voorhees. The TREC-8 question answering track report. In Ellen M. Voorhees and Donna K. Harman, editors, *Proceedings of The Eighth Text REtrieval Conference, TREC*

*1999, Gaithersburg, Maryland, USA, November 17-19, 1999*, volume 500-246 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 1999.

[96] Tom Waterman. Web scraping is now legal. Jan 2020. Accessed Jan. 27, 2021. URL: https://medium.com/@tjwaterman99/web-scraping-is-now-legal-6bf0e5730a78.

[97] Katarzyna Wegrzyn-Wolska and Piotr S. Szczepaniak. Classification of RSS-formatted documents using full text similarity measures. In David Lowe and Martin Gaedke, editors, *Web Engineering*, pages 400–405, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[98] William Winkler. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods*, 01 1990.

[99] Nic Newman with Richard Fletcher, Anne Schulz, Simge Andı, and Rasmus Kleis Nielsen. Reuters Institute Digital News Report 2020. *Reuters Institute for the Study of Journalism*, 2020.

[100] Zhibiao Wu and Martha Palmer. Verbs Semantics and Lexical Selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, ACL '94, page 133–138, USA, 1994. Association for Computational Linguistics. doi:10.3115/981732.981751.

[101] Bei Yu and Linchi Kwok. Classifying business marketing messages on Facebook. *Proceedings of the Association for Computing Machinery Special Interest Group on Information Retrieval, Bejing, China*, pages 24–28, 2011.

[102] Mingwei Yuan, Ping Jiang, Jin Zhu, and Xiaonian Wan. Sensing Semantics of RSS Feeds by Fuzzy Matchmaking. *Intelligent Information Management*, 2(2):110–119, 2010. doi:10.4236/iim.2010.22014.

[103] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing Twitter and traditional media using topic models. In Paul Clough, Colum Foley, Cathal Gurrin, Gareth J. F. Jones, Wessel Kraaij, Hyowon Lee, and Vanessa Mudoch, editors, *Advances in Information Retrieval*, pages 338–349, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.