FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Evaluation of Text Diversity over time for Automatically Generated Texts in Sports Journalism

José David Souto Rocha

Mestrado em Engenharia Informática e Computação

Supervisor: Prof. Sérgio Sobral Nunes

July 27, 2023

Evaluation of Text Diversity over time for Automatically Generated Texts in Sports Journalism

José David Souto Rocha

Mestrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

President: Prof. João Correia Lopes

Referee: Prof. Nuno Escudeiro

Supervisor: Prof. Sérgio Sobral Nunes

July 27, 2023

Resumo

A Geração de Linguagem Natural (GLN) é uma área de investigação derivada da Inteligência Artificial (IA), que se insere no domínio do Processamento de Linguagem Natural (PLN). A GLN tem um vasto leque de casos de uso, nomeadamente no jornalismo desportivo, permitindo a geração autónoma de notícias. A ZOS é uma empresa portuguesa especializada na criação de conteúdos desportivos para a plataforma online *zerozero.pt*. No decorrer dos últimos anos, a ZOS tem desenvolvido o Prosebot, um sistema GLN capaz de gerar resumos textuais de jogos de futebol com base em dados armazenados pela ZOS. Atualmente, devido à falta de recursos humanos, apenas uma fração dos jogos recebem artigos escritos por jornalistas, deixando uma lacuna na cobertura dos restantes. Os sistemas GLN podem desempenhar um papel importante para colmatar esta lacuna, especialmente em ligas amadoras, que também têm um um público dedicado dentro da comunidade futebolística. Nesta dissertação, introduzimos uma nova métrica que avalia a diversidade textual em *feeds* de notícias gerados por sistemas GLN ao longo do tempo. A nossa métrica utiliza o algoritmo de similaridade do cosseno em textos pré-processados, que incorpora *tokenização*, reconhecimento de entidades (REN) e *stemming*. Para validar a sua eficácia e robustez, a métrica foi integrada no Prosebot, atuando como módulo externo que tinha como objetivo identificar conteúdos repetitivos nos *feeds* de notícias do Prosebot. Efetuámos análises comparativas dos *feeds* de notícias antes e depois deste processo para avaliar o impacto da diversidade textual nas notícias do Prosebot. Além disso, realizámos três estudos, nos quais comparámos a diversidade textual das notícias geradas pelo Prosebot com as de notícias escritas por jornalistas. O objetivo destes estudos era avaliar o desempenho da nossa métrica em diferentes cenários, onde já esperávamos resultados específicos. Também utilizámos avaliação humana através de um formulário de avaliação com base na experiência de utilizadores. Os resultados estavam de acordo com o que era esperado, fornecendo provas empíricas da eficácia e fiabilidade da métrica. Além disso, realizámos um estudo centrado nas contribuições dos utilizadores para a pós-edição de texto gerado pelo Prosebot, que revelou que os utilizadores são mais mais propensos a alterar frases mais curtas e com mais significado.

Keywords: Geração de linguagem natural, Jornalismo desportivo, Geração de dados para texto, Diversificação Textual

Abstract

Natural Language Generation (NLG) is an area of research derived from Artificial Intelligence (AI), set on the topic of Natural Language Processing (NLP). NLG has a vast range of applications, particularly in journalism, where it enables the autonomous generation of news pieces. ZOS is a Portuguese company specialized in sports content creation for the online platform *zerozero.pt*. Over the last few years, ZOS has developed Prosebot, an NLG system which generates textual summaries of football matches based on data stored in ZOS' databases. Currently, due to a lack of human resources, only a fraction of matches receive articles written by journalists, leaving a gap in coverage for the remaining ones. NLG systems can play an important role in filling this gap, particularly in lower-level football leagues, which also have a dedicated audience within the overall football community. In this dissertation, we present a novel metric that evaluates text diversity in news feeds generated by NLG systems over time. Our metric uses the cosine similarity algorithm on preprocessed texts, which incorporates tokenization, named entity recognition (NER), and stemming. We integrated the metric into Prosebot, serving as an external module which identified repetitive content in Prosebot's news feeds. We performed comparative analyses of news feeds before and after this process to evaluate the impact of text diversity on Prosebot's output. Additionally, we carried out three separate studies, where we compared the text diversity of Prosebot-generated news with news written by journalists. The purpose of these studies was to assess the performance of our metric under different scenarios, where we already expected specific outcomes. We also employed human evaluation through a user-based evaluation form. The results were in alignment with what was expected, providing empirical evidence of the metric's effectiveness and reliability. Additionally, we conducted a parallel study focusing on users' contributions in post-editing Prosebot-generated text, which showed that users are more prone to changing shorter and more meaningful sentences.

Keywords: Natural Language Generation, Natural Language Processing, Sports Journalism, Data-to-text generation, Text Diversification

Acknowledgements

Words cannot describe the experience these past 5 years have been. The 18-year-old version of myself could have never envisioned the incredible journey that would unfold before me. I grew as an individual, I forged unforgettable memories and I uncovered aspects about myself that I didn't know of. I would like to thank everyone that accompanied me throughout this defining journey of my life.

First and foremost, I would like to thank my supervisor, Sérgio Nunes, for their expertise and guidance throughout this work. Even when I lacked clear direction, their guidance helped me find the right path. I am grateful for providing me with this opportunity and for always pushing me to produce the best work possible.

I would also like to thank all members of ZOS. A special appreciation goes to Marco Sousa and Pedro Dias, for their support and for providing everything necessary to carry out this work.

To my friends, who I have been lucky enough to share this journey with, whether in Portugal or during my time in Poland (specifically Kitchen 2A), I thank you. I specifically want to acknowledge and thank the "Auntdulce" group, with whom I shared countless memories and formed an invaluable bond that I treasure dearly. I will never forget the countless Discord calls, the laughter, the inside jokes, the exam struggles, and all the adventures we lived. It was a heck of a ride, which was nothing short of incredible. I am sure this was only the beginning of the story we began writing in 2018.

To my girlfriend, Rita, I thank you. I have been lucky enough to have crossed paths with you and to have found someone who has always shown me their full support and love. I hope to make more memories alongside you for years to come.

To my family, I want to express my utmost gratitude. To the ones who I am fortunate to now share and celebrate this moment - my parents António and Renata, my sister Teresa and my aunt Teresa - I thank you. To the ones who are no longer among us, but I'm sure are proudly watching from above - Licínio, Rosélia and Isabel - I thank you. You are my biggest pillar, my inspiration and my driving force. You have given everything a son, brother, nephew and grandson could have wished for and more. I can only hope one day to be able to repay you in the same way. I thank you for everything you have done for me, for the education you provided me with, and for shaping me into the person I am today. In what is the greatest exponent of my academic journey, I want to especially extend my gratitude to my grandmother Rosélia, who played an invaluable role in my life. Your presence continues to be felt in my heart each and every day, and I am forever indebted to you for the person I have become. Even during tough times, you always showered me with your unconditional love, support, company and wisdom. I am confident that if I can provide my future grandchildren with even a fraction of the love and support you have given me, I will have fulfilled an incredible role as a grandparent. No matter where life may lead me, know that I will forever be "my grandmother's boy". To you, I dedicate this work.

José Rocha

“Fisicamente, habitamos um espaço, mas, sentimentalmente,
somos habitados por uma memória.”

José Saramago

Contents

List of Figures

List of Tables

Abbreviations

NLG	Natural Language Generation
NLP	Natural Language Processing
NER	Named Entity Recognition

Chapter 1

Introduction

The Natural Language Generation (NLG) field relates to the use of Artificial Intelligence (AI) for several tasks, including producing textual content that mimics human language. In a world of rapid technological and economic evolution, NLG has established itself as a commercial software category, gathering frequent attention from industry media as well as the mainstream press, with businesses willing to invest substantial sums of money to benefit from this technology [7]. Therefore, it is only natural that companies have gained a particular interest in developing their tools to be up-to-date with the requirements and needs of several industries, such as the journalism industry.

1.1 Context

ZOS is a Portuguese company that creates sports content for the online website *zerozero.pt* and serves as a platform provider for several media outlets, ranging from television and newspapers to betting companies. Throughout the last few years, ZOS has developed Prosebot, a natural language system that generates textual football match summaries based on data and statistics stored on ZOS's databases. This system is already used by journalists as a starting point to create news content. Additionally, advanced users of *zerozero.pt* can use Prosebot to automatically generate match summaries for football games, which span from amateur to professional leagues, featuring both professional players and those from youth football academies [34].

1.2 Motivation

NLG systems, such as Prosebot, have the potential to revolutionize the way sports media outlets report on matches. According to Shao et al. [44], one of the key challenges in text generation is achieving a good balance between word diversity and content accuracy, which becomes even more significant in the context of sports reporting, where repetitive writing can diminish reader engagement and interest.

Currently, due to a lack of human resources, only a fraction of matches (usually the most relevant ones) receive articles written by journalists, leaving a gap in coverage for the remaining ones. NLG systems can play an important role in filling this gap, particularly in lower-level and amateur football leagues, which also have a dedicated and interested audience within the overall football community. In spite of this, there are still some issues regarding generating text that appears natural and human-like. These systems often struggle to produce varied and dynamic content, which is essential for captivating readers and maintaining their interest while reading pieces of news. Addressing this challenge and developing systems capable of generating diverse and engaging news is crucial for sports media outlets to keep up with the fast-paced nature of sports reporting. This improvement can lead to numerous benefits, not only for media outlets seeking to maintain reader engagement but also for sports fans who rely on up-to-date information.

1.3 Problem Identification

The current state of Prosebot presents some challenges related to the text diversity of the generated news feeds, particularly in their news titles, where there is a noticeable repetition, as evidenced by the striking similarities between them. Consequently, the news feeds produced by Prosebot become highly redundant, specifically in domains like French or Spanish, which have access to a smaller pool of available templates in comparison to other domains, such as Portuguese. Thus, it becomes imperative to adjust the system to enhance text diversity evaluation, thereby ultimately improving the overall reading experience.

1.4 Hypothesis

Based on the previously identified issues of text diversity within the current Prosebot system, this work proposes the following hypothesis:

The development and implementation of a metric designed to assess text diversity will result in a measurable improvement in evaluating the overall diversity of generated content.

By investigating this hypothesis, this study aims to provide empirical evidence on the effectiveness of a novel metric specifically designed to evaluate text diversity in news feeds. The proposed metric will serve as a tool for assessing and monitoring the level of diversity in the generated content, enabling quality control measures to be implemented.

1.5 Objectives

Given the proposed hypothesis, the main objective of this work is to develop a metric that can evaluate text diversity over time. This metric will incorporate linguistic and NLP techniques to analyze various aspects of the generated text. Furthermore, it will be tailored to accurately measure the level of diversity in Prosebot's news feeds.

Another objective is to integrate the developed text diversity evaluation metric within the Prosebot system. The integration process will ensure that the metric can automatically evaluate the diversity of generated content in real time and will enable continuous monitoring and improvement of text diversity within Prosebot's news feeds.

The third objective focuses on evaluating the quality of the implemented metric in identifying text diversity in news feeds. To achieve this, we will perform a series of studies which aim to test and compare the metric's perception of text diversity to what is already expected. Additionally, feedback will be collected through surveys, allowing insights into users' experiences with news feeds. This user feedback will serve as a benchmark to assess the effectiveness of the implemented metric in detecting repetitive content.

1.6 Document Structure

This document consists of 7 chapters, each focusing on a specific aspect of this work. Chapter 2 serves as an introduction to the Natural Language Generation (NLG) field and provides an overview of NLG systems employed in journalism. Chapter 3 discusses the problem of text diversity in NLG, text similarity algorithms used to compare texts, and their applicability in comparing articles from news feeds. Chapter 4 delves into a better explanation of the problem under investigation, providing a clearer understanding of its significance. The implementation of the proposed text diversity metric is illustrated in Chapter 5, including a detailed description of its integration within Prosebot. Chapter 6 showcases a series of studies conducted to validate the effectiveness of the proposed metric, offering insights into its reliability. Finally, Chapter 7 concludes this dissertation, summarises the main findings and contributions, and outlines possible avenues for future research and development. Additionally, a parallel study related to user contributions regarding post-editing Prosebot text can be found in Appendix A.

Chapter 2

Overview of Natural Language Generation

The following chapter presents an overview of the field of Natural Language Generation (NLG) along with its definition, applications and tasks. It also discusses the history of NLG, as well as its origin and its subsequential growth as an area of research and development within the artificial intelligence community. Furthermore, it will explore the different approaches commonly used in NLG as well as the several tasks involved and the challenges and limitations associated with it. Lastly, we will describe the applicability of NLG in real-world scenarios for both journalism and sports journalism.

2.1 What is Natural Language Generation?

Reiter and Dale [39] defined the concept of Natural Language Generation as *"the subfield of artificial intelligence and computational linguistics that is concerned with the construction of computer systems that can produce understandable texts in English or other human languages from some underlying non-linguistic representation of information"*. While this definition accurately describes the purpose of data-to-text generation, which involves converting information from structured data such as tables and knowledge graphs into natural language, it is essential to acknowledge other areas of NLG research. Text-to-text is one of the alternative approaches, which utilizes linguistic content rather than structured data to generate new texts. Text-to-text methods find applications in several domains, such as text summarization and machine translation, as evidenced in the biomedical field [27].

Despite becoming an ever more active area of research and development in the last couple of years, Natural Language Generation is a field that has existed for a long time, dating back to the 1950s. Although works from earlier periods can be found, it was in this era that Alan Turing proposed what is now known as the Turing Test [51], the first criterion of intelligence for a machine.

Since then, the applications of NLG have been growing day by day [17], due to the increasing need to understand and derive meaning from a language with its numerous ambiguities and varied structure [47]. Furthermore, Panetta [11] has stated in the past that *"natural-language generation will be a standard feature of 90 per cent of modern BI and Analytics platforms"*. Similarly, Dale [7] also believes in the expansion of NLG as a field of interest and from a commercial point of view. From a practical point of view, and as Reiter and Dale evidenced in 1997, NLG technologies have a wide range of applications. They can be used in several daily scenarios, such as:

- **Healthcare** – facilitating patient review, shifting handover, and care transitions [42].
- **Weather forecasts** – selecting suitable terminology for conveying numerical weather information [31].
- **Automatic Route Description** – generating natural route descriptions based on input obtained from a commercially available way-finding system [8].
- **Sports Journalism** – generating textual news concerning Finnish ice hockey games [16].

2.2 NLG Tasks

Natural Language Generation systems typically involve several tasks organized into an overall architecture, which can vary depending on the system or application being developed. The definition of this set of tasks was proposed by Reiter and Dale [40] in 2000. Reiter refined this proposal in 2007, stating that an effective NLG system should handle a wide range of data sources and generate coherent, accurate, and appropriate text for the target audience. This architecture took raw data as input instead of knowledge bases and was designed to be flexible and modular, allowing different components to be customized or replaced to meet the specific needs of others [38]. Years later, in 2018, Gatt and Krahmer described NLG tasks as *"the problem of converting input data into output text"* by *"splitting it up into several subproblems"* [12]. Moreover, they enumerated the tasks proposed by Reiter and Dale in the following manner:

1. **Content determination**: choosing the information to be included in the text;
2. **Text structuring**: deciding in which order the order in which the information will be presented in the text;
3. **Sentence aggregation**: determining which information should be included in each sentence;
4. **Lexicalisation**: selecting the appropriate words and phrases to convey the information;
5. **Referring expression generation**: choosing the words and phrases to label domain objects;
6. **Linguistic realization**: assembling all the words and phrases into grammatically correct sentences.

2.2.1 Content determination

The starting point for generating text involves selecting the appropriate content to be included in the end product based on the input data and the purpose of the text. Furthermore, it consists in deciding which pieces of information are most relevant and essential to include and how they should be presented.

Content determination is a critical task in NLG, as it determines the relevance and coherence of the generated text. For instance, and in the case of Prosebot, the generated text should only contain the focal points of interest for a specific match and not information concerning smaller events, such as fouls, corner kicks and throw-ins, for example. Despite some of these events being relevant to the match (i.e. a player committing a foul and conceding a penalty or a goal resulting from a corner kick, for example), most of them are not relevant for the overall view of the event and thus should not be reported.

2.2.2 Text structuring

The second step for generating text involves determining the text's overall structure and deciding how the content should be organized and divided into smaller units, such as paragraphs or sections. Text structuring is essential for generating coherent and understandable text, as it helps to ensure that the text flows smoothly and makes sense to the reader.

2.2.3 Sentence aggregation

Sentence aggregation combines smaller units of content, such as clauses or phrases, into larger units, such as paragraphs or sections. It involves deciding how to group the content and how to connect it using linking words and phrases, aiming to ensure that the text is coherent and flows smoothly. For instance, considering the football domain, the following scenario could occur:

1. Lionel Messi had an incredible World Cup.

2. Lionel Messi was considered the best player in the tournament.

These sentences exhibit redundancy and lack coherence, resulting in a less enjoyable reading experience. Therefore, a more suitable alternative could be, for example:

1. Lionel Messi had an incredible World Cup *and* was considered the best player in the tournament.

This technique was defined by Reiter and Dale as conjunction. Additionally, they introduced other techniques for sentence aggregation, such as pronominalization and discourse markers. Pronominalization involves replacing a noun, which is mentioned on multiple occasions, with a pronoun to avoid repetition. Similarly, the following sentences

1. Lionel Messi just won the World Cup.

2. Lionel Messi was over the moon.

could be converted to:

1. Lionel Messi just won the world cup. *He* was over the moon.

Instead of mentioning *"Lionel Messi"* multiple times, we can use the pronoun *"he"* to refer back to him, which not only reduces redundancy but also improves the overall flow of the text. Discourse markers, on the other hand, are words or phrases that connect sentences or ideas within a text, providing coherence and guiding the reader through the content. They can include words which indicate relationships between different parts of the text and make it more readable and comprehensible. By using appropriate discourse markers, writers can create a logical progression and smooth transitions between sentences, enhancing the overall clarity and cohesion of the text. Likewise, the following sentence

1. Cristiano Ronaldo is completely off form, the manager should replace him with another player.

could be converted into:

1. Cristiano Ronaldo is completely off form, *thus* the manager should replace him with another player.

2.2.4 Lexicalisation

The process of selecting the appropriate words and phrases to use in the generated text is described as lexicalisation. Moreover, this phase involves deciding which words and phrases best convey the meaning of the content and how they should be used in the context of the text. It is influenced by factors such as the style and tone of the text, the intended audience, and the context in which the text will be used.

Regarding sports journalism, which encompasses a distinct and well-defined domain, it proves advantageous to connect phrases to specific events. Associating multiple phrases with each event and randomly selecting one for each instance within the input enables the generated text to be more diverse, fluid, and easily understandable. For example, when referring to a particular occurrence, such as a player's remarkable performance, there are various expressions like *"scored three goals"*, *"scored a hat-trick"*, and *"put three past the keeper"* that all pertain to the same event. The incorporation of such variation allows the generated text to be more engaging and captures the essence of the event in a human-readable manner.

2.2.5 Referring expression generation

Similar to lexicalisation, referring expression generation (REG) involves generating phrases or expressions, with the exception that it refers to specific entities or concepts in the input data. In other words, an NLG system might generate a referring expression to refer to a particular person, place, or thing. This step is important for ensuring that the text is coherent and understandable, as it helps to disambiguate references and establish the relationships between different entities in the text.

Gatt and Krahmer [12] state that the manner in which a system refers to an entity can be influenced by various factors, including whether the entity has been previously mentioned in the text. In cases where this is true, the system has the option of using a pronoun as a means of reference. Additionally, it may also be necessary to distinguish the entity from other similar entities by including specific attributes, such as the time of the event. For example, in a sports article mentioning Lionel Messi multiple times, the system might choose to employ alternative pronouns or aliases, such as *"Argentina's captain"* or *"seven-time Ballon d'Or winner"*, to differentiate him from other players and enhance clarity and readability.

2.2.6 Linguistic realisation

This is the process of generating the final text from the structured content, ensuring that it is grammatically correct and follows the appropriate style and conventions. It is the final step in the NLG process and involves two sub-tasks: surface realization, which involves generating the final text, and style realization, which involves ensuring that the text is appropriate for the intended audience and context. Several approaches, which were discussed by Gatt and Krahmer [12], can be used to support linguistic realization in NLG systems, including rule-based systems, machine-learning techniques, and hybrid approaches.

2.2.6.1 Human-crafted Templates

Human-crafted templates are predefined structures or frameworks that are created by human experts to guide the generation of the final text. They typically include placeholders for the content to be included in the text and specify the overall structure and organization of the text. When the scope of the application is limited, and changes are expected to be minimal, creating the output is a straightforward task. Templates can be used to specify the output, as seen in studies by Reiter et al. [41] and McRoy et al. [24].

An example of a human-crafted template for a football report system would be:

$player scored for *$team* in the *$minute* minute.

In this template, the placeholders in brackets would be filled in with the appropriate content based on the input data. For example, if the input data indicated that Lionel Messi had scored a goal in the 23rd minute for Argentina, the template could be filled in as follows:

Lionel Messi scored for *Argentina* in the *23rd* minute.

On the one hand, using human-crafted templates provides a clear and consistent structure for the generated text, which can help to ensure it is coherent and understandable. Moreover, they can also be customized to meet the specific needs and goals of the intended audience. However, one disadvantage of using human-crafted templates is that they require significant effort and expertise to create, and they may not be able to handle all possible combinations of input data and contexts. In addition to this, they may not be able to capture the nuances and complexities of natural language, resulting in text that is less natural and less engaging for the reader [39]. In spite of this, some authors have claimed that this is not necessarily true. For example, van Deetmer et al. [52] state that a template approach does not necessarily produce a worse quality of output in comparison to other approaches.

2.2.6.2 Hand-coded grammar-based systems

Hand-coded grammar-based systems use a set of rules and guidelines that have been manually coded by human experts to guide the generation of the final text. These rules are based on the grammar and syntax of the target language and specify how different elements of the text should be combined and structured. Furthermore, they provide a clear and consistent structure for the generated text, which can help to ensure that the text is coherent and easy to follow. Similarly to a template-based approach, they can also be customized to meet the specific needs and goals of the intended audience [25], as Mernik et al. evidenced. The authors also state that one of the difficulties of hand-coded grammar-based systems is that they require significant effort and expertise to create and maintain, and they may not be able to handle all possible combinations of input data and contexts. In other words, it may be challenging to design rules that have the appropriate sensitivity to context.

Furthermore, Gatt and Krahmer discussed that designing hand-crafted rules with the appropriate sensitivity to context and input is challenging for grammar-based systems, particularly when it comes to making choices among related options. An example of this is, for instance, choosing amongst the following:

1. Lionel Messi scored for Argentina in the 23rd minute.
2. For Argentina, Lionel Messi scored in minute 23.
3. Argentina's player Lionel Messi scored after 23 minutes.

2.2.6.3 Statistical approaches

As the name indicates, statistical approaches use statistical models [1] and machine learning techniques to learn patterns and relationships in large amounts of text data and to generate new text that follows these patterns. These approaches typically involve training the model on a large dataset of text, such as news articles or football reports, and then using the model to generate new text based on a set of input data [3]. For example, a statistical model is able to generate the following sentence:

Lionel Messi is a key player in Argentina's build-up play.

To generate this phrase, the statistical model would have been trained on a dataset of texts about football, including descriptions of players, teams, and tactics. The model would have learned patterns and relationships between words and phrases such as *"Lionel Messi"*, *"key player"*, *"Argentina"*, and *"build-up play"*. When given the input *"Lionel Messi"* and *"Argentina"*, the model would use these patterns and relationships to generate the output mentioned previously. In comparison to template-based approaches and hand-coded grammar-based systems, statistical approaches are prone to be more flexible and adaptable, as they can learn from a wide range of examples [37] and can handle a wide range of input data and contexts. They can also generate text that is more natural and engaging for the reader, as they can capture the nuances and complexities of natural language.

However, statistical approaches can be more complex and resource-intensive to implement and maintain, and they may not always produce a grammatically correct or coherent text. In addition, they may be less transparent and interpretable than template-based approaches, making it more difficult to understand how the generated text was produced.

2.3 NLG in Journalism

Natural Language Generation (NLG) systems in journalism have been increasingly used to automate the production of news articles [29], particularly in sports journalism and financial reporting. These systems typically take structured data as input and generate coherent, accurate, and appropriate text for the target audience.

One advantage of using NLG systems in journalism is that they can save time and effort in the production process, allowing journalists to focus on more complex tasks such as analysis and interpretation. They can also provide coverage of events that would be difficult or impossible for human journalists to attend, such as multiple sports games happening simultaneously [21]. However, NLG systems also face several challenges and limitations in journalism. One challenge is ensuring that the generated text is accurate and reliable, as errors or misinformation in the generated text could have serious consequences. Another challenge is ensuring that the generated text is engaging and reader-friendly, as poorly written or boring text may not attract and retain readers. To address these challenges, NLG systems have made use of a range of approaches, including rule-based systems, machine learning-based systems, and hybrid systems. These systems have also been evaluated using a variety of methods, including human evaluation, readability metrics, and quality measures such as informativeness and fluency.

Overall, the state-of-the-art in NLG systems for journalism is rapidly evolving, with ongoing research and development aimed at improving the accuracy, reliability, and engagement of the generated text. We will now examine a number of known and widespread NLG systems used in the field of journalism. These systems were chosen due to their state-of-the-art nature, as well as their widespread usage and popularity in various fields, including journalism. The result of this selection can be seen in Table 2.1. With this in mind, we will provide an analysis of the current landscape of NLG technology based on the most commonly employed tools available.

Table 2.1: Example of NLG systems used for journalism (among other fields).

Type	**Name**	**Year**	**Linguistic Realisation**
Commercial NLG System	Quill	2010	Templates + Machine Learning
	Wordsmith	2011	Templates + Statistical Approach
	Arria NLG Studio	2016	Grammar-Based
Autoregressive transformer language models	GPT-1	2018	Statistical Approach
	GPT-2	2019	Statistical Approach
	GPT-3	2020	Statistical Approach
	GPT-3.5	2022	Statistical Approach
	GPT-4	2023	Statistical Approach

Wordsmith[1] is a commercial system developed by Automated Insights and first released in 2011. It is used for various applications, including generating news articles, sports reports, and financial reports. This system uses a combination of templates and statistical approaches to generate written content and includes a set of pre-built templates that can be customized and used to generate text. When it comes to evaluating generated text, a combination of human evaluation and automatic algorithm checks is used. The quality and accuracy of the generated content are considered by humans, who review and edit the content as needed. In contrast, several built-in quality checks and metrics ensure the generated content meets specific standards, such as readability and grammar.

Quill[2] is a system developed by Narrative Science and was first introduced in 2010. This system serves as a way to automatically generate written content based on structured data inputs and is designed to work with a wide range of data sources. Furthermore, it is commonly used to create news articles, sports reports, and financial reports while also being able to be customized to meet the specific needs of different organizations and industries. In terms of algorithms deployed, Quill uses a combination of templates and machine learning algorithms to generate written content and, similar to the previous system, it uses both human and automated evaluation.

Arria NLG Studio[3] was developed by Arria NLG and first released in 2016. In terms of its functionality, Arria NLG Studio allows users to create custom NLG templates using a drag-and-drop interface. These templates define the structure and content of the generated text and can be tailored to the user's specific needs. The platform also includes a range of pre-built templates and templates for specific industries and uses a combination of hand-coded grammar-based systems and statistical approaches to generate text. It is evaluated through human and automated methods, including algorithms that measure the quality and relevance of the generated text.

Over the last few years, OpenAI[4] has developed the GPT series. In spite of the intent of these systems being generating human-like text (similar to an NLG system), they are categorized

[1] https://automatedinsights.com/wordsmith/

[2] https://www.discovercloud.com/products/quill

[3] https://www.arria.com/nlg-studio/

[4] https://openai.com/

as autoregressive transformer language models. The GPT systems use neural networks trained to generate text one word at a time based on the words that have come before it, emplouying a transformer architecture that allows the system to handle long-term dependencies in the text, generating output text more efficiently and effectively than previous language models.

GPT-1 [36] (Generative Pre-trained Transformer) was the first model of this series. Developed in 2018, it is primarily used for tasks such as generating news articles, stories, and prompt responses. The model was pre-trained on a large dataset, which means that it had already learned specific patterns and features of language before it was fine-tuned for particular tasks. It relies on statistical approaches to understand the structure and patterns of language from the data it was trained on rather than using templates or hand-coded grammar-based systems. This allows it to generate new text similar in style and content to the training data. GPT-1 can be evaluated automatically, using automated metrics, and by humans, through subjective evaluations of the generated text. OpenAI then released **GPT-2** [37], in 2019, and **GPT-3** [5], in 2020. These models improved upon GPT-1 by having more parameters, being able to perform tasks without fine-tuning, and having better performance on natural language processing tasks.

More recently, **ChatGPT**[5] (2022) was introduced, which employed the underlying **GPT-3.5** [55] architecture, constituting an upgrade over the previous version. This version was trained on vast amounts of conversational data to generate human-like text in response to prompts, which were a result of a training process involving a combination of supervised learning and reinforcement learning techniques. Additionally, the system underwent fine-tuning through human evaluation, where human trainers ranked the model-generated responses from previous conversations. These rankings were used to create "reward models," which further refined the model using multiple iterations of the Proximal Policy Optimization (PPO) algorithm. PPO serves as a computationally efficient alternative to Trust Region Policy optimization algorithms, which are reinforcement learning algorithms, reducing the need for computationally expensive operations while maintaining faster performance. Moreover, in 2023, the incorporation of **GPT-4** [30] took place in ChatGPT, further enhancing its capabilities. GPT-4 introduced the capability to process both images and text as inputs, expanding the range of data modalities it can handle. Additionally, GPT-4 increased the input size, enabling better performance in tasks that involve processing longer documents or mixed text and image data.

2.4 NLG in Sports Journalism

Similar to what was done in the previous section, we will now examine several known and widespread NLG systems. The selection of these systems was based on a methodology that considered their relevance for sports journalism and their discussion among the scientific community. With this in mind, we included systems that have been recognized for their effectiveness in generating sports journalism content. The result of this methodology can be seen in Table 2.2.

[5] https://openai.com/blog/chatgpt/

Table 2.2: NLG systems used for sports journalism.

Name	Year	Linguistic Realisation	Evaluation
Chen et al. [6]	2008	Machine-Learning	Human
PASS [54]	2017	Template-based	Human
Kanerva et al. [16]	2019	Machine-Learning	Automatic
Taniguchi et al. [48]	2019	Encoder-Decoder	Automatic + Human

PASS (Personalized Automated Soccer texts System) [54] is a template-based system that generates Dutch sports reports in real-time. It utilizes a modular and customizable approach to generate tailored reports for specific audiences, with its key feature being the ability to generate reports with different tones of voice based on the targeted audience's emotional connection to the teams involved in a match. For example, if the audience's favourite team loses, the generated report will reflect disappointment or frustration, whereas if the team wins, the tone will be more upbeat. Hence, the goal is to replicate the emotional language expected from human-written reports while maintaining a professional writing style. Due to this aspect, the authors argued that the system was capable of producing text with a high level of variation, similar to that of Goal-Getter [50], a data-to-speech football reporting system. This system generates spoken reports of football matches based on input in table format and a small database. The input text is stored on a Teletext page and contains data on one or more football matches, while the database holds information about the teams and their players. In terms of data collection, PASS scrapes soccer match data from goal.com[6], which provides information about teams, players, match events, and statistics. The system focuses on generating reports that are similar to those written by human journalists and aims to capture the emotional tone present in reports from the MeMoFC corpus [4], which contains match reports published by soccer clubs for their supporters. The report generation process involves designing templates based on sentences from the previously mentioned corpus, which is achieved by replacing specific match-related information with gaps to be filled with data from goal.com.

Kanerva et al. [16] proposed, in October 2019, a news generation system for generating Finnish ice hockey news articles based on structured data. To train the system, they compiled a corpus of information based on more than 2000 game reports spanning over 20 years of ice hockey data. The authors argue that one of the challenges in using real journalistic material is that the articles contain a mixture of information directly available in the statistics and information inferred from the statistics, such as background knowledge, game insight, and player interviews. Directly using the limited amount of actual news articles for end-to-end system training can lead to the generation model hallucinating facts that are not present in the statistics. To address this issue, the news corpus was manually cleaned by rephrasing the text and removing portions of the text that were not directly supported by the available game statistics. The authors concluded that

[6] https://goal.com/

the resulting system was able to generate text that was relatively close to what was considered a viable product by human journalists, taking into account the word error rate. Most of these errors fell into a small number of categories, such as copying names from the input, types of events, and time references, which would need to be addressed in future work.

Chen et al. [6] introduced, in 2008, a commentator system that learned vocabulary from sportscasts of simulated football matches, in a similar fashion to how children acquire language through exposure to linguistic input and perceptual experience. This system utilizes a simulated environment that mimics a dynamic world with multiple agents and actions, similar to a real soccer game, but without the complexities of robotics and vision. Furthermore, the RoboCup simulator was used to provide a detailed physical simulation of robot soccer. In order to train and test the model, symbolic representations of game events were automatically extracted from the RoboCup simulator traces, while human commentators provided the natural language commentaries by typing them into a text box. These game events and human commentaries were then paired, and the system learned to interpret and generate language by observing them. However, it is worth noting that, unlike Prosebot, this system highlights every event during a match instead of only reporting the relevant ones.

Taniguchi et al. [48] proposed, in 2019, a data-to-text system that generated commentary for English Premier League games. Furthermore, they used an encoder-decoder approach, which is a framework that uses an encoder to process input and produce a context vector, then used by a decoder to generate an output sequence. Additionally, they used an attention mechanism, which focused on the relevant events regarding the matches, and a placeholder reconstruction, enabling the system to copy appropriate player and team names from the input data. An automatic evaluation was conducted to assess the system's performance, with BLEU as the metric for automatic evaluation, alongside human evaluations, which aimed to supplement the possibly inaccurate BLEU scores. These scores were divided according to the length of the evaluated text (10 words or fewer, 15 words or fewer, and 20 words or fewer). Moreover, ten subjects rated the generated texts in terms of their grammar correctness and informativeness on a scale from 1 to 3, with the overall results showing that the scores decreased as the text length increased.

Chapter 3

Text Diversification Techniques

The following chapter discusses the problem of text diversification regarding automatic text generation and its relevance in NLG systems. Moreover, we will depict commonly used metrics for text similarity evaluation as well as their applicability in article comparison of news feeds.

3.1 The Problem of Diversification

While NLG technologies have come a long way since first being introduced, generating text that can be interpreted as diverse remains challenging. Regarding NLG, improving diversity in the generated text is a synonym for creating a more varied set of output options for a given input or task. Furthermore, this improvement can include generating text that uses different words, phrases, and structures, as well as styles that express different emotions, tones, or perspectives. The goal is to make the generated text more varied and engaging while reducing the likelihood that the same or similar text will be generated multiple times. This is an issue that has been recurrently discussed by multiple authors. Gao et al. [10] stated, in 2019, that NLG models could often generate bland and generic text in spite of the great potential they displayed. Likewise, Montahaei et al. [28] observed that, while text generation models had shown progress in generating high-quality sentences, there was still limited diversity in the generated texts. They claimed that the existing metrics for evaluating NLG systems did not adequately address the diversity aspect, since the ones that existed did not account for both sentence quality and diversity at the same time. Shao et al. [44] also commented that one of the key challenges in text generation was finding a balance between word diversity and accuracy. On the other hand, Tevet and Berantto [49] later demonstrated, in 2021, that while these models could already produce diverse outputs due to technological advances, they still lacked principled methods for evaluating the diversity of NLG systems.

One of the main reasons for this issue is the intrinsic complexity of human languages, which are highly context-dependent and can vary significantly across different cultures, dialects and styles. Moreover, NLG systems often rely on large amounts of data to learn the patterns and structures of human language, and the quality and diversity of this data can significantly impact the variety of the generated text. Thus, NLG systems often struggle to capture the nuances and subtleties of human languages, such as irony, sarcasm, and humour, which can limit the diversity of the generated text. Although the generated texts may be considered acceptable, they can be redundant or lack coherence when multiple different outputs are required. This issue is caused by how the training of these models is done. As a result, they become overfitted and cannot utilize the diverse vocabulary they can access [22]. With this in mind, we will go over the similarity algorithms used for evaluating sets of texts in terms of their similarity.

3.2 Text Similarity Evaluation

The evaluation of text similarity involves the use of text similarity algorithms, which provide a numeric value that indicates how closely two texts resemble each other. This value reflects the degree of similarity between the texts, with higher values indicating greater similarity and vice versa [56]. However, before delving into how these algorithms work and their purpose, it is important to understand the concept of the vector space model.

The vector space model (VSM), initially introduced by Gerard Salton in 1989, is a mathematical representation that enables us to express text documents as numerical vectors in a multi-dimensional space. In this model, each dimension of the vector corresponds to a specific term or feature, and the value of each dimension represents the importance or presence of that term in the document. The vector space model finds numerous applications, including the measurement of similarity between two text documents. This analysis involves calculating the similarity between the vector representations of these documents, which is accomplished by measuring the angle or distance between the document vectors. Through this process, we can determine the degree of similarity between texts and identify documents that are similar to each other.

3.2.1 Vector representation of text documents

In the vector space model, there are several techniques for converting text documents into vector representations. **TF-IDF** [15] (Term Frequency-Inverse Document Frequency) is the most commonly used method in NLP for converting text documents into a matrix representation of vectors [19]. It takes into account both the term frequency (TF), which measures the importance of a term within a document and the inverse document frequency (IDF), which measures the rarity or uniqueness of a term in the entire document collection. TF-IDF assigns higher weights to terms that are more frequent within a document but less frequent across the entire collection, capturing their discriminative power.

Another approach to creating vector representation for text documents relates to the usage of word embeddings. Word embeddings represent words as dense vector representations in a continuous space, capturing semantic and syntactic relationships between words. **GloVe** [32] uses word counts to build a co-occurrence matrix, where each matrix's row corresponds to a specific word and each column represents the different contexts in which the word can be found. The scores assigned by GloVe indicate the frequency of co-occurrence between words. **Word2Vec** [26] takes into account the context of each word, aiming to assign similar numerical representations to similar words. This predictive model learns vector representations by minimizing the loss of predicting target words from given context words. Over the course of the years, variations of this model have risen, such as **SentenceToVec** [18] and **Doc2Vec** [20]. SentenceToVec learns feature representations at the sentence or document level instead of individual words by averaging the vector representations of all words in a sentence, while Doc2Vec extends this procedure to capture document-level embeddings, including sentences as part of the document structure. While these word embeddings are widely used and capture syntactic and semantic information, they have limitations in handling more complex NLP tasks, which require context-independent embeddings with greater capabilities [19].

ELMo [33] (Embeddings from Language Model) is a bidirectional Language Model (biLM) that uses a large corpus to generate multi-layered word embeddings. ELMo's pretrained vectors are capable of extracting conceptualized word representations that encompass syntax, semantics, and word sense disambiguation (WSD). By integrating ELMo with existing deep learning approaches, it becomes possible to enhance the performance of diverse complex NLP tasks through the creation of supervisory models. **BERT** [9] (Bidirectional Encoder Representations from Transformers), builds upon the bidirectional concept introduced by ELMo but adopts a transformer architecture. BERT's training involves learning bidirectional representations by considering the context of the corpus in both directions across all layers. These pretrained vectors can be applied to complex NLP tasks, surpassing existing benchmarks by adding just one additional layer to the output.

3.2.2 Similarity algorithms

After the text documents are converted into vector representations, which can be plotted on the vector space model, we can compare them in order to assess their level of similarity. In the context of text similarity evaluation, various algorithms have been developed to quantify the resemblance between two texts. In this section, we will discuss some of them, which stand as the most common ones used. The choice of these algorithms was based on the literature review carried out by Zhang [56].

The **inner product similarity** algorithm considers the weights of features shared by the compared vectors, excluding features possessed by only one vector. This makes it limited to some extent since it does not consider the global pool of all features in the two vectors, leading to biased calculations. To illustrate, we will consider two scenarios: in the first scenario, two vectors contain 5 keywords each, all present in both vectors. In the second scenario, two vectors have 10 keywords each, but only 5 are shared. Ideally, the first scenario should yield a higher similarity value due to the 100% keyword match, but this is not the case with the inner product similarity. To address this issue, alternative similarity measures such as the **Dice similarity** are employed. The Dice similarity not only considers the shared features between both vectors but also includes features possessed by either vector. This measure works similarly to the inner product measure, except it adds a denominator to the formula, the sum of weights of both vectors, as a way of normalizing the result. The normalization avoids the problem of unfair calculations previously described. A similar algorithm to the Dice algorithm is the **Overlap Coefficient similarity measure**, which normalizes the result by taking the minimum sum weights from the two vectors as its denominator. Another popular algorithm is the **Jaccard similarity**, which measures the similarity between two sets by calculating the ratio of the intersection to the union of the sets. In the context of text comparison, the sets represent the unique words present in each text. **Herdan's C algorithm** is a measure used to assess the vocabulary richness and diversity of a text. It takes into account the number of unique words and their frequencies in a given text to calculate a diversity score. This algorithm calculates the logarithmic ratio of the total number of unique word types to the total number of word tokens in a given text. A higher Herdan's C score indicates greater lexical diversity and textual richness, suggesting a more diverse and varied composition of words in the text.

In addition to these measures, distance-based similarity measures are also utilized to compare two vectors in terms of similarity, such as the Manhattan distance and the Euclidean distance. The **Manhattan distance** between two vectors is the sum of the absolute differences between their corresponding coordinates, i.e. the total distance from one vector to another, only considering horizontal and vertical movements. On the other hand, the **Euclidean distance** represents the distance in a straight line from one vector to the other.

Another approach to calculating text similarity relates to the **Cosine similarity**, which measures the cosine of the angle between the compared vectors, determining the degree of overlap or dissimilarity between the texts. Lower cosine similarity values indicate greater diversity, while higher values suggest similarity or redundancy.

On the whole, these similarity algorithms provide different perspectives for assessing the similarity between vectors, which represent text documents in this case. While some focus on shared features, others consider individual features or use distance metrics. Choosing the most appropriate measure depends on the specific context and requirements of the problem at hand.

3.3 Text Similarity algorithms in article comparison of news feeds

Several authors have explored the usage of text similarity algorithms for comparing news articles in news feeds, with a lot of the carried out research on this domain regarding news clustering and grouping articles based on their topics. For instance, Bergamaschi et al. [2] introduced a web feed reader that used the Jaccard similarity algorithm to group news articles from different newspapers and published on different days. Their tool successfully grouped over 700 news articles published in 30 different newspapers over a span of four days. Similarly, Pons-Porrata et al. [35] proposed a hierarchical clustering algorithm that employed cosine similarity to uncover implicit knowledge within news streams. Sakhapara et al. [43] also presented a system that clustered news based on their similarity, utilizing measures such as Jaccard Similarity and Cosine Similarity to group news feeds from different sources.

However, despite the existing body of work that utilizes text similarity algorithms for comparing news articles, there is a significant gap when it comes to addressing the specific problem we aim to tackle. While these articles mention the usage and effectiveness of text similarity algorithms in comparing news articles, as described in the previous paragraph, there is a lack of research concerning the evaluation of text diversity within a news feed. We were unable to find any directly relevant entries that explore this aspect and, consequently, we concluded that research in this particular field is scarce, as there are currently no existing metrics that can serve as benchmarks for comparing with our proposed metric. Thus, we came to the conclusion that the evaluation of text diversity within a news feed remains an unexplored area.

Chapter 4

Problem Statement

The following chapter details the problem regarding the Prosebot system and describes the proposed solution in order to mitigate it. Moreover, we will go over the research questions, requirements, scope and steps in order to carry out the implementation of our solution.

4.1 Problem

The primary issue regarding the current state of Prosebot lies in its repetitive use of similar or near-identical templates during the process of generating news. Consequently, the resulting news feeds become highly redundant, which gives the content an artificial tone and degrades the reading experience for visitors of ZOS' online platforms. To address this problem effectively, it becomes mandatory to develop a mechanism capable of detecting common patterns and repetitiveness in the generated news feeds. The implementation of such a solution can not only mitigate the issue of detecting repetitiveness but also contribute to improving Prosebot's evaluation of its generated content.

4.2 Research Questions

The problem identified in the previous chapter, which relates to the lack of text diversity in Prosebot's generated content, can be further explored and addressed through the following research questions:

RQ1: What factors contribute to the limited text diversity in Prosebot's current state? Identifying the underlying reasons for the lack of diversity in the generated text is crucial to understand the challenges and opportunities for improvement.

RQ2: How can we effectively measure and evaluate text diversity in the context of Prosebot? Developing appropriate metrics and evaluation methods specific to Prosebot's generated football match summaries will enable an accurate assessment of text diversity and facilitate meaningful comparisons.

4.3 Solution Requirements

In order to address the problem outlined in Section 4.1 and effectively answer the research questions presented, the proposed solution must fulfil the following requirements:

D1: **Develop a text diversity evaluation metric**. The solution needs to introduce a robust evaluation metric specifically designed to assess and monitor text diversity over time. In other words, we should take into account that two titles with similar content closer in time should yield higher similarity scores compared to a similar pair of titles that are more distantly separated temporally. The metric should consider various aspects of diversity, such as lexical variation, syntactic structures and semantic expressions. Additionally, it should accurately depict text diversity in news feeds.

D2: **Incorporate linguistic and NLP techniques**. The solution should employ linguistic and NLP techniques to analyze the diversity of the generated text. These techniques may involve measuring vocabulary richness, analyzing sentence structures, detecting repeated patterns, and assessing stylistic variations. The combination of linguistic and NLP approaches aims to ensure the solution provides a more accurate evaluation of text diversity.

Through fulfilling these solution requirements, we can benefit from an effective and comprehensive evaluation of text diversity, leading to improved insights, refinements, and ultimately, more diverse and engaging generated content.

4.4 Scope

The scope of this study is to develop a text evaluation metric that addresses the issue of text diversity within news feeds, specifically focusing on the syntactic aspect of text diversity. The aim is to detect excessive repetition within a group of news and track text diversity variation across time, which has the potential to be useful for mitigating limited text diversity.

To ensure the practicality and applicability of the metric, this study will use data provided by ZOS and the available resources within the Prosebot system. Using these resources, the evaluation process will involve analyzing the generated content to identify and measure text diversity patterns. It is worth noting that the scope of this study does not involve significant modifications to the infrastructure or core functionalities of NLG systems. Instead, it seeks to enhance a system's evaluation capabilities, thus enabling real-time monitoring of text diversity across time. In spite of this, the developed evaluation metric and its findings have the potential to serve as a foundation for further research and advancements in the field of natural language generation and text diversity evaluation.

4.5 Proposed solution

This section presents the proposed solution to address the identified challenges related to text diversity evaluation.

4.5.1 Syntactic Diversity Measurement

The proposed metric will make use of several measures, such as sentence length, word order, part-of-speech variations and syntactic dependencies to assess syntactic diversity, in order to quantify the level of diversity exhibited for each news within the news feed.

4.5.2 Time-based Analysis and Visualization

The metric development process will incorporate time-based analysis techniques to capture changes in syntactic diversity over a given period. Drawing from research on time series analysis and visualization methods, the metric will analyze how syntactic diversity evolves and fluctuates over time. The results of this analysis will then be visually represented on a graph, allowing users to observe variations, trends, and patterns in text diversity. The visual representation of syntactic diversity across time aims to facilitate a better understanding of how diversity changes over different periods of time.

4.5.3 Validation and Fine-tuning

To ensure the accuracy and effectiveness of the metric, extensive testing and validation will be conducted. To do so, representative datasets of Prosebot-generated titles, covering significant time spans, will be selected for evaluation purposes. The metric's outputs will be compared to human judgments of text diversity, enabling fine-tuning and refinement of the metric.

By validating the metric against human judgments, the study aims to ensure that the metric aligns with general perceptions of text diversity. This process will enhance the reliability and relevance of the metric's results, providing a robust evaluation tool for measuring text diversity across time.

4.5.4 Integration and Real-time Analysis

In order to test the usability and effectiveness of this metric in a real-world context, the text diversity metric will be integrated within Prosebot, allowing us to monitor text diversity in real-time. With this in mind, we aim to provide a valuable tool for assessing and monitoring the evolution of text diversity in news feeds, enabling researchers, developers, and users to gain deeper insights into text generation systems' performance and the generated content's diversity.

Chapter 5

Implementation of the Text Diversity Metric

The following chapter presents the implementation details of the text diversity evaluation metric. We will provide an overview of the metrics pipeline and highlight the key components, methodologies, and considerations involved. Additionally, we will outline the integration of this metric alongside the Prosebot system.

5.1 Program Design and Architecture

The metric follows a structured pipeline to evaluate the text diversity of news feeds. Figure 5.1 provides an overview of its design and architecture, which consists of several interconnected components that work together to assess text diversity. Furthermore, the architecture is designed to be modular, allowing for flexibility, scalability, and easy maintenance. Each component operates independently but collaboratively to execute the evaluation process, ensuring a smooth and efficient functioning of the metric.

The process of evaluating text diversity starts by taking a group of news pieces as input. These news pieces undergo preprocessing, which includes several steps which will be detailed in the following sections. Next, the metric compares the news pieces amongst each other – using these

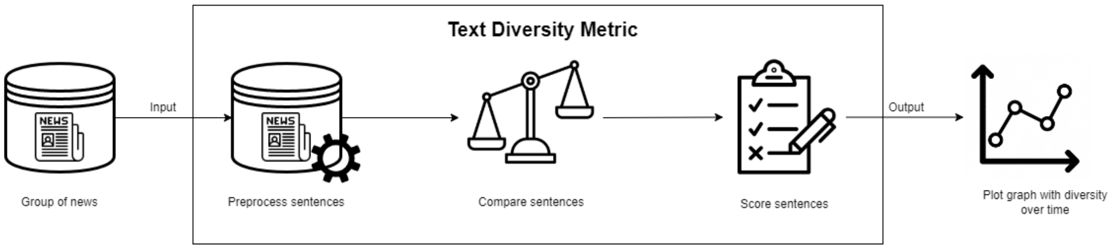


Figure 5.1: Program Design and Architecture Diagram for the assessment of text diversity.

comparisons, the metric assigns a score to each news piece. For each news piece, the corresponding score represents the degree of similarity between the specific news piece and its previous n news pieces, ordered chronologically. The formula for achieving this score and the amount of n news used for generating a score will be detailed further ahead. Finally, we generate a chart that visually represents the variation of scores, ordered chronologically across the news feed, providing an overview of the text diversity trends across time. This process enables a quantitative assessment of text diversity and facilitates the identification of patterns and shifts in diversity within news feeds.

In the subsequent sections, we will delve into the specific implementation details of each module, explaining the algorithms, methodologies, and techniques employed. The aim is to provide a complete understanding of the metric's functioning and its ability to assess text diversity. Moreover, we will detail the applicability of this system within Prosebot's news feeds.

5.2 Input Processing

The input processing module outlines the data preprocessing steps performed on each piece of news before comparing it with other pieces of news. The steps involved in preprocessing are the following:

1. **Named Entity Recognition (NER)**: the first step is to perform NER on the text to identify entities such as football clubs, players, or other relevant entities. Identifying and removing such entities will allow us to focus on analyzing the remaining text for its assessment of text diversity.

2. **Tokenization**: after performing NER, the sentence is tokenized by splitting it into individual tokens. This process allows for the analysis of each word separately. Additionally, during tokenization, any punctuation marks are removed.

3. **Stemming and lowercasing**: The remaining words in the title, after removing entities, undergo stemming. Stemming reduces words to their base or root form, removing prefixes or suffixes. Additionally, all words are converted to lowercase to standardize the vocabulary.

The aim behind these preprocessing steps was to create a standardized representation that reduces the original text to its relevant content, ensuring consistency and promoting a more accurate analysis of the generated news. Furthermore, our intent regarding our metric's text diversity perception lies in the syntactic phrasal constructions used during the generation of the sentence. We considered the presence of specific entities should not influence the text diversity evaluation since multiple news articles may discuss the same entity, leading to potential bias in the calculated score. To illustrate, let's consider the news feeds presented in Table 5.1. In Feed 1, there is a noticeable repetition of the phrasal construction *"... beat ... in a thrilling encounter"*. Similarly, Feed 2 consistently mentions the topic of *"FC Porto"* in every news title. However, it is evident that Feed 2 demonstrates greater text diversity compared to Feed 1, despite Feed 2 referring to the same entity in all of its news. This discrepancy arises from Feed 2's use of distinct verbs, nouns, and adjectives to report results, in contrast to Feed 1, which relies on a limited set of words. Therefore, we believed that removing entities and focusing on the words responsible for conveying sentences' meaning would be advantageous for comparing news titles within a feed.

Likewise, another important aspect to address is our reason behind not removing stopwords during the preprocessing of news. Stopwords are common words that contribute little to the overall meaning of the text. The usual practice is to eliminate them to only retain content-rich words in the preprocessed representation of a sentence. However, since news titles tend to be shorter sentences, we believed removing stopwords would result in excessively brief representations. For example, the sentences

1. Valencia too good for Real Madrid
2. Cancelo and Guardiola: all good as of now

would be reduced to only ["good"], failing to accurately capture the original sentence's meaning. Since entities are already removed from the original sentences, removing stopwords as well would result in an oversimplified version, potentially leading to incorrect interpretations or outcomes. Moreover, in this case, these sentences would be a perfect match, despite pertaining to completely different topics.

It is important to mention that the above-mentioned factors constitute limitations to our work. With these thought processes, we intend to justify our decisions regarding these limitations, aiming to provide transparency and clarity regarding our approach to the preprocessing phase. Table 5.2 displays the preprocessing of an English news feed to better understand this process's result.

Table 5.1: Example of two news feeds.

Feed 1	**Feed 2**
FC Porto beats Benfica in a thrilling encounter	FC Porto dominates Tondela
Real Madrid beats Barcelona	5th win in a row for FC Porto
Paris SG beats AS Monaco in a thrilling encounter	A night to remember for FC Porto
Manchester City beats Arsenal in a thrilling encounter	FC Porto defeats SC Braga by a narrow margin
Napoli beats AC Milan in thrilling encounter	FC Porto seals important win against Benfica

Table 5.2: Preprocessed titles derived from a news feed.

Original Title	Preprocessed Title
"Manchester United get the better of Manchester City"	[’get’, ’the’, ’better’, ’of’]
"Valencia too good for Real Madrid"	[’too’, ’good’, ’for’]
"Spoils shared between Sporting and Benfica"	[’spoil’, ’share’, ’between’, ’and’]
"Real Betis put to the sword at home by Sevilla"	[’put’, ’to’, ’the’, ’sword’, ’at’, ’home’, ’by’]
"Auxerre lose out to Paris SG"	[’lose’, ’out’, ’to’]
"Lazio beat Udinese in close encounter"	[’beat’, ’in’, ’close’, ’encount’]
"Nothing to separate Santos and Palmeiras"	[’noth’, ’to’, ’separ’, ’and’]
"Newcastle and Leicester City go head-to-head this Monday"	[’and’, ’go’, ’head’, ’to’, ’head’, ’thi’, ’Monday’]
"FC Augsburg handed lesson by Borussia Dortmund"	[’handed’, ’lesson’, ’by’]
"Atlético Madrid thrash Osasuna"	[’thrash’]
"West Ham defeat Leeds United"	[’defeat’]

Table 5.3: Models used for preprocessing.

Language	NER Model	Tokenizer Model	Stop Words
English	Stanford NLP [23]	dslim/bert-base-NER[1]	Spacy[2]
Portuguese		pierreguillou/ner-bert-large-cased-pt-lenerbr [14]	
Spanish		mrm8488/bert-spanish-cased-finetuned-ner[3]	
French		flair/ner-french[4]	
Italian	Spacy[5]	spacy/it_core_news_lg[6]	

Additionally, while developing the metric, one of our goals was to ensure its applicability across multiple languages, making it language-independent. To validate its effectiveness, we conducted tests using English, Portuguese, Spanish, French, and Italian. However, it is important to highlight that, as long as a language has a corresponding tokenizer available, the metric can be adapted accordingly. For reference, Table 5.3 presents the models and tokenizers utilized for each of the aforementioned languages.

[1] https://stanfordnlp.github.io/CoreNLP/
[2] https://huggingface.co/dslim/bert-base-NER
[3] https://spacy.io/
[4] https://huggingface.co/pierreguillou/ner-bert-large-cased-pt-lenerbr
[5] https://huggingface.co/mrm8488/bert-spanish-cased-finetuned-ner
[6] https://huggingface.co/flair/ner-french
[7] https://spacy.io/models/it

5.3 Comparing Methodology

In order to assess the text diversity between pairs of sentences, we started by converting each text to a vector representation, using the TF-IDF technique described in Section 3.2.1. Furthermore, to ensure these representations were comparable and did not favour longer sentences, we applied unit normalization, which rescaled vector components to unit length while preserving their direction. This approach guaranteed that the length of a text did not have an impact on its comparison of text diversity. Once the vector representations were obtained for a pair of texts, we used the cosine similarity algorithm between the vectors and calculated the resulting similarity score between the two texts. Higher similarity scores indicated a higher level of similarity and, therefore, less text diversity, while lower scores suggested greater text diversity between the sentences. The reason behind the choice of using cosine similarity will be explained in the next section.

As mentioned earlier, the score assigned to an entry in a news feed is determined by a formula that takes into account the previous *n* entries. The details of this formula will be elaborated upon in the next section.

5.4 Scoring mechanism

The developed scoring mechanism assesses the text diversity of each sentence by comparing it with the previous *n* sentences within the news feeds, using the methodology described in the previous section. If there are fewer than *n* previous sentences available, the comparison is made with the maximum number of preceding sentences up until *n*. The value of *n* can be adjusted depending on the system's needs. For this work, we used *n*=5. The choice of this value relates to the depth to which a human can detect text similarity between news titles. For instance, if two similar titles were positioned consecutively, they would catch the reader's attention much more easily compared to if they were separated by ten other news pieces. We will delve into this concept further in this work when we discuss human evaluation regarding this metric.

To obtain the final cosine similarity score, a weighted average of all *n* calculated cosine similarities is computed, where the weights of each score depend on the proximity sentences to the reference sentence. In order to give more emphasis to closer ones, the weight distribution follows an exponential decay function, ensuring that sentences closer to the reference carry more significance in determining the final score. Moreover, formula 5.1 represents how this calculation is made.

$$\text{final\_score} = \frac{\sum_{i=0}^{n-1}(\text{scores}[i] \cdot e^{-i})}{\sum_{i=0}^{n-1} e^{-i}} \tag{5.1}$$

The numerator of this formula represents the weighted sum of the *n* cosine similarity scores and is computed by taking the sum of the products of each score and its corresponding weight. As stated, the weight for each score is determined by the exponential decay function, where the weight decreases exponentially as the index increases. Specifically, the weight is calculated as e^{-i}, where *e* is Euler's number and *i* is the index of the score. Likewise, the denominator represents the total weight, which is computed by taking the sum of the weights for all indices in the list. This sum serves as a normalization factor, ensuring that the weighted average is appropriately scaled. Finally, the formula divides the weighted sum by the total weight to obtain the weighted average of the cosine similarity scores.

The usage of cosine similarity as our metric's text similarity algorithm was based on its extensive discussion and usage within the scientific community. It has been widely employed for comparing news articles, as referenced in Section 3.3. However, it is important to note that the essence of our scoring mechanism does not primarily rely on the used text similarity algorithm. Rather, it centres around our approach to preprocessing sentences and understanding the relationships between article distance and perceived diversity. Besides that, and as previously mentioned, given that we are working with news titles, which tend to be short in terms of sentence length, we believed it would be advantageous to utilize a commonly adopted text algorithm. If we were comparing longer texts, additional factors would need to be considered, and a more detailed evaluation of the most suitable algorithm would be necessary. Similarly to what was stated in Section 5.2, this pertains to a limitation to our work. Our intent was, once more, to justify our decisions regarding these limitations.

An example of what the output of the scoring mechanism for a news feed would look like can be seen in Table 5.4. It is worth mentioning that, for this example, each news piece was assigned an ID, ranging from 1 to 20. This ID reflects the news' chronological order, with 20 being the most recent and 1 being the oldest. The analysis of the scores allows us to conclude that, for this scenario, the metric identified sentences with ID 9, 13, 16, 17 and 18 as exhibiting higher degrees of similarity among the feed. These similarities arise from the presence of nearly identical phrasal structures, such as the recurring use of the verb *"overcome"* (sentences 15 and 18), the phrasal construction *"with a tough task against"* (sentences 14 and 16), and the expressions *"expected to"* and *"widely expected to"* (sentences 11, 13 and 17).

While identifying repetitive patterns is relatively straightforward in this particular scenario, since the feed size is relatively small (20 news pieces), it can become challenging to visualize and monitor text diversity when dealing with larger input sizes. To address this issue, we have implemented a graph display to output the final scores, which facilitates better visualization and interpretation of the results.

Table 5.4: Example of a news feed scoring.

ID	News Titles	Score
20	Little to separate Argentina and France	0
19	Croatia and Morocco will both fancy their chances	0
18	Argentina overcome Croatia	0.09
17	France expected to beat Morocco	0.04
16	Argentina with a tough task against Croatia	0.23
15	Portugal overcome Switzerland	0
14	Portugal with a tough task against Switzerland	0
13	Spain widely expected to beat Morocco	0.23
12	Brazil crush South Korea’s World Cup dreams	0
11	Brazil widely expected to beat South Korea	0
10	Japan and Croatia play later on at the Al Janoub Stadium	0
9	England move on to the next round of World Cup	0.09
8	England tipped to sweep Senegal aside	0
7	Can minnows Poland humble France?	0
6	Argentina move on to the next round of World Cup	0
5	Argentina and Australia face each other in mouth-watering clash	0
4	Netherlands need to bring A-game against United States of America	0
3	Cameroon emerge victorious against Brazil	0
2	Switzerland defeat Serbia	0
1	Uruguay run out winners versus Ghana	0

5.5 Graph Visualization and Interpretation

After obtaining the scores and predictions for each sentence, a line graph is plotted to visualize the text diversity trends. The line graph represents the variation in scores over time, with the x-axis denoting the chronological order of the news pieces, and the y-axis representing the assigned scores. This visual representation provides a clear indication of the positions of repetitive sentences within the news feed and allows for easy identification of patterns and clusters. Figure 5.2 displays an example of the graph obtained from the scores and predictions showcased in Table 5.4, presenting the text diversity dynamics within the news feed.

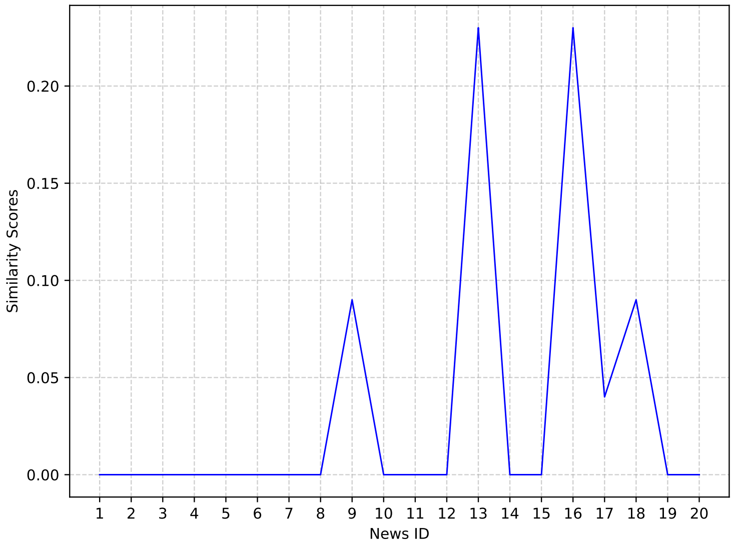


Figure 5.2: Graph obtained from the values of Table 5.4.

The overall trend of the graph, such as the presence of peaks and valleys, indicates the presence of diverse or repetitive content over time. Sharp declines or spikes in the scores highlight instances of significant shifts in text diversity within the news feed. It is important to note that the presented graph is an introductory example, and further experimentation and analysis is needed to assess the accuracy of the metric in representing text diversity. In order to do so, we will showcase a series of experiences, which will provide a better understanding of how these line graphs function and how to interpret them. These studies and their findings will be detailed further in this work.

5.6 Integration of the Text Diversity Metric within Prosebot

Our next task involved integrating the metric as an external module into Prosebot, enabling it to monitor text diversity across various domains in the content it generates. Furthermore, we developed two additional modules that identified repetitive titles and then automatically rephrased them, in an attempt to promote diversity in the generated news feeds. In the upcoming sections, we will provide an overview of the implementation and functionality of these modules.

5.6.1 Architecture and System Integration

The architecture was designed to facilitate the detection and replacement of repetitive titles within Prosebot domains. With this in mind, the architecture consists of three main modules:

1. **Data acquisition module**: leverages ZOS's APIs to scrape the most recent news articles, extracting their titles for further analysis;

2. **Text diversity module**: utilizes the text diversity metric to assess the similarity between titles and identify repetitive patterns;

3. **Title rewriting module**: interacts with the ChatGPT API[7] to generate alternative paraphrases.

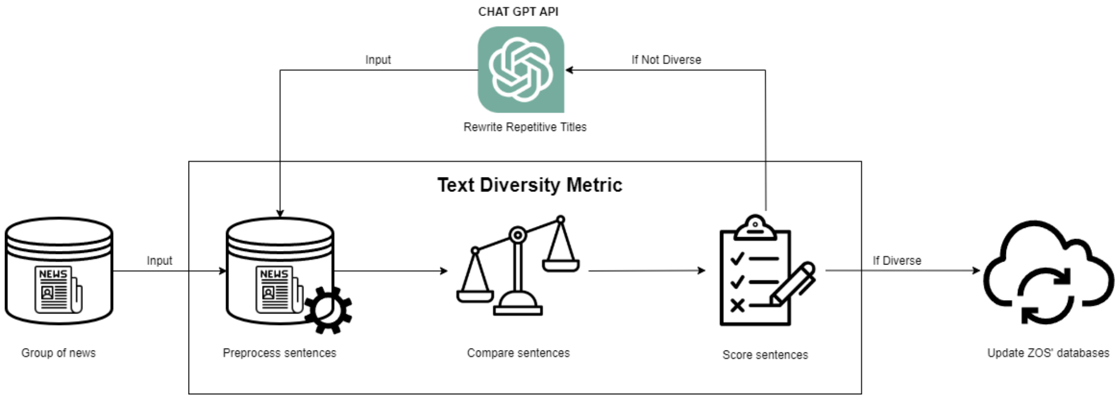


Figure 5.3: Program Design and Architecture Diagram for detecting and rewriting repetitive titles.

The implemented architecture operates in an iterative manner, repeating the detection and rewriting process until no repetitive titles are identified or the maximum number of attempts is reached. The maximum number of attempts can be customized. For this work, we set it to 5 tries. Figure 5.3 further illustrates the rewriting architecture pipeline.

5.6.2 Detection of repetitive titles

Our approach to detecting repetitive titles involved training a Linear Regression model using the scores obtained from the text diversity metric. In order to do so, we constructed a dataset comprising 200 news titles generated by Prosebot. Each title was then classified as repetitive (assigned a value of 1) or non-repetitive (assigned a value of 0), which was done based on our human judgement alone. Moreover, we established a minimum precision threshold of 0.9 to determine the model's satisfactory performance. The results from training showed that the model met our expectations by achieving a precision of 1.0, higher than the minimum established. In a practical scenario, when applied to the news feed depicted in Figure 5.4, the model accurately identified titles with the IDs 9, 13, 16, 17, and 18 as repetitive, aligning with our initial expectations. In order to make sure the model didn't suffer from overfitting, which happens when models succeed in predicting training data but fail to do so on new data, we monitored the metric's assessment of text diversity in Prosebot feeds during the course of a month and compared them to our own judgements, in order to check if any unexpected results occurred. This was not the case, which allowed us to carry on with our investigation.

[7] https://chat.openai.com/

5.6.3 API Request Structure and Rewriting

To enable effective title rewriting using the external API, a structure was implemented for sending requests and extracting their outputs. This section provides a detailed explanation of the API request structure employed in the title rewriting process, along with optimization techniques to enhance the reliability and completeness of responses.

As stated previously, we used ChatGPT as a way of rewriting the repetitive titles. Furthermore, our idea was to construct an input for each request, which contained the flagged content, and extract the output from the API in the simplest way possible. For the case of Table 5.4, the input would be the following:

"Rephrase the following sports titles to make them more appealing to readers while maintaining the original capitalization and providing accurate titles. Use different verbs and synonyms related to sporting terms. Output the revised titles as a Python list with the original capitalization while maintaining accuracy and not using subjective terms.

- Argentina overcome Croatia
- France expected to beat Morocco
- Argentina with a tough task against Croatia
- Spain widely expected to beat Morocco
- England move on to the next round of World Cup

Output: Python list"

Upon sending the request to the external API, the response contained the revised titles in the form of a Python list. To extract the rewritten titles, a parsing process was implemented, ensuring the information was organized into a valid format for further analysis. For the mentioned scenario, a possible API output could be the following:

["Argentina outshine Croatia", "France poised to defeat Morocco", "Argentina face a challenge against Croatia", "Spain strongly favoured to triumph over Morocco", "England advance to the next stage of World Cup"]

In cases where the API response did not meet the desired criteria, i.e. either the rewritten titles did not differ from their original versions or the responses were incomplete, a try timeout of 10 attempts for exchanging requests and responses was set to handle these cases. The adoption of the structured API request format and the optimization techniques employed played a pivotal role in achieving diverse and engaging title rewriting within the Prosebot domains. Although there were other options besides using an external API, such as employing a pre-trained paraphrasing model, we considered this to be the best option since this work's focus isn't on rewriting repetitive content but rather on monitoring and detecting it. Training a model to do such a thing would constitute a difficult task and would deviate from our main objective, which was to integrate this metric into Prosebot as a way to measure its applicability in NLG systems, as previously stated. The subsequent section will present and analyze the results obtained from using this approach.

5.7 Results and Evaluation

To evaluate the performance of this architecture, we decided to plot line graphs similar to the ones in Figure 5.4.

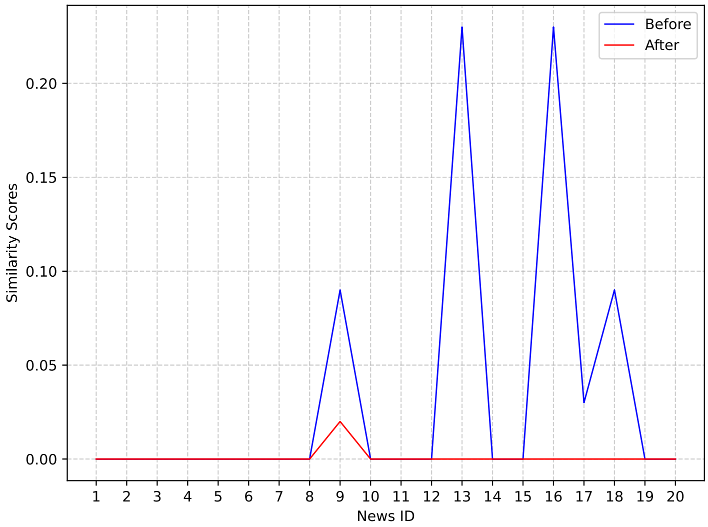


Figure 5.4: Output graph for the news feed from Table 5.4 before and after being rewritten.

The graph illustrates that the curve depicting the post-rewriting process demonstrated reduced peaks in comparison to the previous scenario. The results indicate that the curve representing the rewritten news feed exhibits fewer peaks and fluctuations in text diversity compared to the curve representing the original news feed. Additionally, there is a decrease in the proportion of repetitive titles within the news feeds after identifying and rewriting repetitive content. While the original news feed contained five repetitive titles, the metric did not identify any repetitive titles in the rewritten feed. Although further investigation would be required to effectively say the rewriting architecture had been successful in its tasks, the results suggested that the text diversity metric performed well in identifying the repetitive content. However, in order to provide empirical evidence for this hypothesis, we needed to subject our metric to a series of studies and evaluations to substantiate this claim.

Chapter 6

Diversity Evaluation Studies

This chapter provides an overview of a series of evaluation studies designed to analyze the degree of diversity within news feeds generated by Prosebot. These studies explore various aspects pertaining to textual diversity and establish a comparative analysis between Prosebot's outputs and those of human journalists, considering different scenarios. The primary objective of conducting these evaluations is to gain insights into the current text-generation capabilities of Prosebot and to experiment with the aforementioned comparing methodology outlined in Section 5.3. The anticipated outcome of this research is an alignment of results with our initial expectations, supporting the effectiveness and reliability of the methodology in quantitatively evaluating and comparing the text diversity of news feeds. We will outline the objectives and research questions guiding these studies, describe the methodologies employed, present the results obtained and discuss their findings.

Furthermore, we will also provide details about the methodologies and details used to conduct a form-based evaluation, which involved gathering participants' perceptions of text diversity. The responses obtained from this evaluation served as a benchmark for our metric, further validating its assessments and outputs.

6.1 Comparison of Text Diversity in Portuguese Football League News

6.1.1 Objectives and Research Questions

The aim of this study was to examine text diversity in news feeds derived from 100 Liga Portugal football games, comparing the output generated by human journalists and Prosebot. The research questions guiding this investigation were as follows:

1. What is the extent of text diversity in news feeds generated by Prosebot for Liga Portugal's football games?
2. How does the text diversity of Prosebot-generated news compare to news articles produced by human journalists?

3. Are there significant differences in text diversity between Prosebot and human-generated news articles for Liga Portugal football games?

6.1.2 Methodology

In order to carry out this study, we used a dataset comprising 100 pieces of news related to Liga Portugal football games. For each football game, the dataset included a title generated by human journalists and a title generated by Prosebot. To assess and compare the text diversity between these two sources, the comparing methodology outlined in Section 5.3 was employed. The results were then visualized and analyzed through a line graph, in the same fashion described in Section 5.5.

6.1.3 Results

The initial hypothesis put forth in this study predicted that news articles produced by human journalists would exhibit a higher level of text diversity compared to those generated by Prosebot. However, upon analyzing the resulting graph, visible in Figure 6.1, it became apparent that the graph's non-linear patterns and numerous peaks posed challenges in its interpretation. Even though we could already conclude that news generated by Prosebot exhibited less text diversity, due to the more frequent presence of bigger peaks and valleys pertaining to the Prosebot news line, the graph would become more complex in a scenario where we compared multiple sets of news, which would result in a graph with multiple lines.

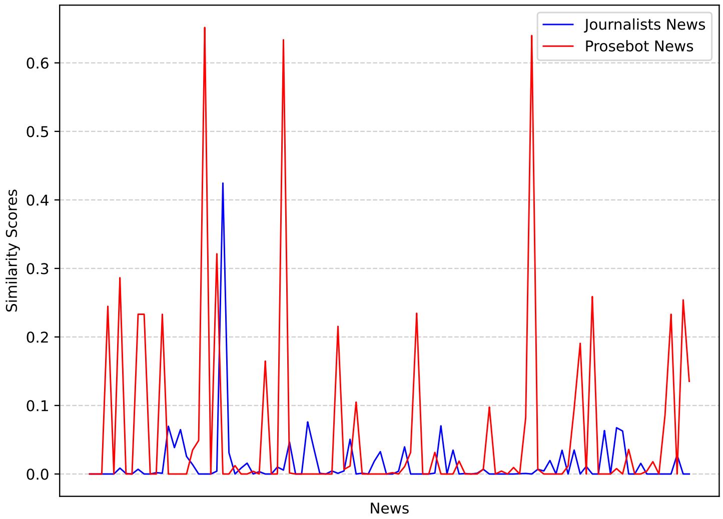


Figure 6.1: Original text diversity graph for 100 Liga Portugal's games generated by journalists and Prosebot (without moving average).

To address the complexity and non-linear nature of the original graph, we employed a moving average technique, which was achieved by taking the average value of a specified number of neighbouring data points, known as the window size. For each score, we calculated a window, excluding the edges where a complete window cannot be formed, and then calculated the average by summing the values within the window and dividing by its size. Since this process smoothens out the data, which may introduce a delay in capturing trends and changes (opposite of our goal), we experimented with several values in order to achieve a balance between easy interpretation and analysis of trends. In the end, we used a window size of 10, which accomplished the desired result. The resulting graph can be seen in Figure 5.4.

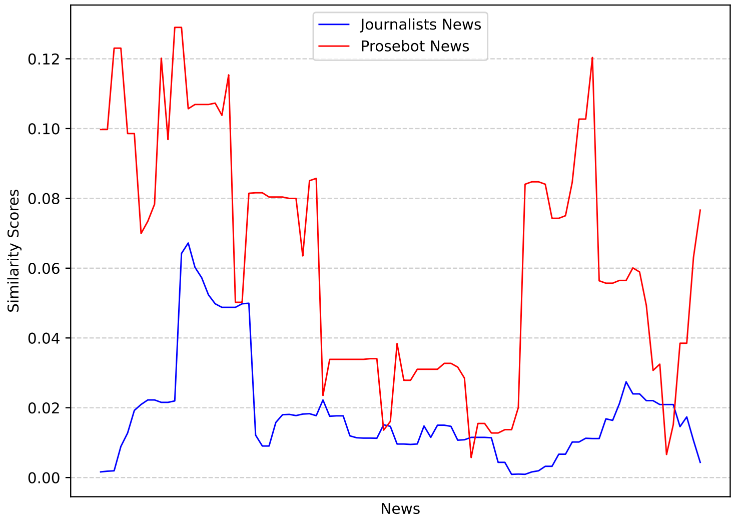


Figure 6.2: Text diversity for 100 Liga Portugal's games generated by journalists and Prosebot (with moving average).

Upon analyzing the smoothed graph, it became even more evident that the metric yielded lower levels of text diversity for Prosebot's outputs compared to those produced by human journalists. Analyzing both curves, we could see the one related to Prosebot news tended to yield higher similarity scores compared to the curve pertaining to journalist news. Furthermore, the metric's output was in alignment with what was expected and supported the assumption regarding the superior text diversity achieved by human journalists, serving as a foundation for the subsequent studies.

It is also worth mentioning that statistical studies were also conducted to further support these findings. The datasets were subjected to tests for normality to determine if there was a significant difference between the sets. The Shapiro-Wilk test, which is used to assess whether a given data sample is drawn from a normally distributed population, was employed, with the results revealing that neither set of news data followed a normal distribution, as evidenced by their p-values close to 0 (0.005 and 0, respectively). Subsequently, the Wilcoxon rank-sum test was utilized to compare the two sets of news data and determine whether there existed a significant difference between sets. This test yielded a *p*-value of 0, confirming a significant difference between the two sets, which provided further evidence to support our findings.

6.2 Comparison of Text Diversity in News Feeds between Clubs

6.2.1 Objectives and Research Questions

This study builds upon the previous study's findings on text diversity in Liga Portugal's football news and focuses on comparing the text diversity in news feeds from a specific football club. Specifically, we examine FC Porto, a club within Liga Portugal. By restricting the news domain to a single club, we intended to study whether a more restricted domain of content contributes to lower diversity in the generated content. With this in mind, the main objective of this study was to individually analyze and compare the level of text diversity between FC Porto's news feeds written by human journalists and Prosebot. Regarding this study, the research questions guiding this investigation are as follows:

1. What are the differences in text diversity between the outputs of Prosebot and human journalists for FC Porto's news?
2. Does restricting the domain influence the text diversity in the generated content?

6.2.2 Methodology

The methodology employed in this study was similar to the study outlined in the previous section. We collected a dataset comprising news written by human journalists and Prosebot for the last 100 FC Porto's football matches. Similar to the previous study, the text diversity of these news feeds was assessed and compared using the comparing methodology outlined in Section 5.3. We then plotted the line graphs and drew conclusions.

6.2.3 Results

Figure 6.3 displays the plotted line graph for this study. Similar to the findings from the previous study, the analysis of text diversity in the news feeds revealed that Prosebot-generated news exhibited lower diversity compared to news articles produced by human journalists, thus indicating that the restriction of the domain did not significantly alter the level of diversity in the generated content. Since we aimed for our metric to only take into account the syntactic constructions of the texts, we expected it to disregard the domain size, i.e. it should be indifferent whether we analyzed a set of news from several football clubs or just one, since the matter here is only how the phrase is constructed and not the meaning it carries. Once again, our metric's output was in alignment with what was expected.

Similarly to the previous study, we conducted additional statistical tests, with the Shapiro-Wilk test indicating both datasets did not follow a normal distribution, as evidenced by their p-values of 0. Moreover, the Wilcoxon rank-sum test revealed a significant difference between the two sets, with a p-value of 0.

Based on these findings, and considering that Prosebot operates as a template-based system, where the content generation relies on pre-existing templates, the limited diversity in the generated text could be attributed to the structure and design of the underlying template-based approach. With this in mind, it became evident that exploring the relationship between template availability and text diversity was a natural extension of this study.

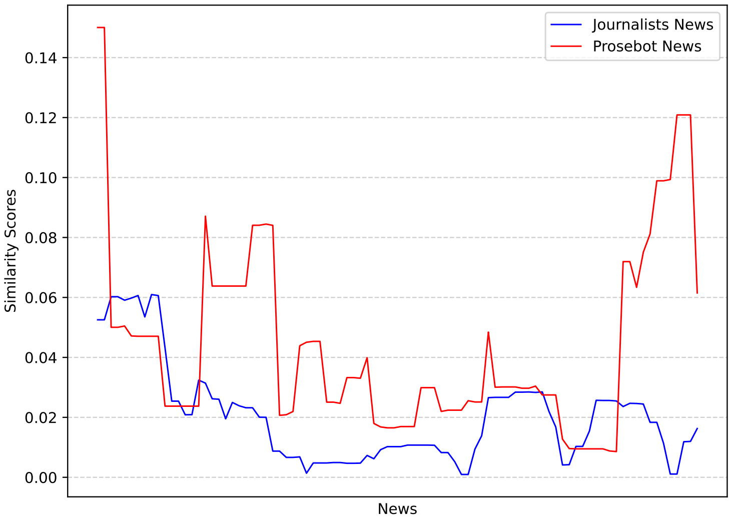


Figure 6.3: Comparison of text diversity in news feeds between FC Porto written by journalists and Prosebot.

6.3 Analysis of Text Diversity Based on Template Availability

6.3.1 Objectives and Research Questions

In the following study, we investigated the hypothesis presented at the end of Section 6.2.3 by examining the diversity of generated text based on the number of available templates. The objective of this study was to explore the relationship between template availability and text diversity in news feeds generated by Prosebot. By varying the number of available templates, we aimed to assess how template availability influences the level of text diversity in the generated content. We expected that the higher the number of available templates, the more likely it would be to exist diversity among them and, consequently, in the generated content. The research questions guiding this investigation are as follows:

Table 6.1: Mean, Median and Standard Deviation for each template availability scenario.

Template Availability	Mean	Median	Standard Deviation
20%	0.39	0.36	0.09
40%	0.18	0.17	0.05
60%	0.09	0.09	0.03
80%	0.08	0.08	0.05
100%	0.05	0.03	0.03

1. How does increasing the number of available templates affect the text diversity in news feeds generated by Prosebot?

2. To what degree does the text diversity increase when Prosebot has access to a larger pool of templates?

3. Is there a significant correlation between the number of templates and the level of text diversity in the generated news content?

6.3.2 Methodology

For this study, we used the same dataset of news feeds pertaining to Liga Portugal's football games, employed previously in the study described in Section 6.1. Within this dataset, we had access to the match IDs for each game and, leveraging the capabilities of Prosebot, which can generate news content based solely on a match ID, we collected rewritten news titles by providing Prosebot with them.

To investigate the impact of template availability on text diversity, we created five separate datasets. Each dataset was constructed by modifying the template availability settings for Prosebot during the text generation process. We created datasets with 20%, 40%, 60%, 80%, and 100% of Prosebot's original templates, which, in absolute terms, accounted for 15, 28, 41, 54 and 67 templates, respectively. For each dataset, we collected the rewritten news titles generated by Prosebot using the specific template availability setting. These news titles were then subjected to the comparing methodology outlined in Section 5.3 to evaluate and quantify the level of text diversity. In the same fashion as the previous studies, we plotted a line graph to visualize and analyze the results.

6.3.3 Results

In addition to the graphical representation of the diversity scores, visible in Figure 6.4, we also present the values for the median, mean, and standard deviations for each template availability scenario. These values can be seen in Table 6.1. This table offers a quantitative summary of the diversity scores, allowing for a more detailed comparison between the different cases.

The analysis of the graph present in Figure 6.4 shows that a higher number of available templates leads to an increase in text diversity. These findings are supported by the distribution of

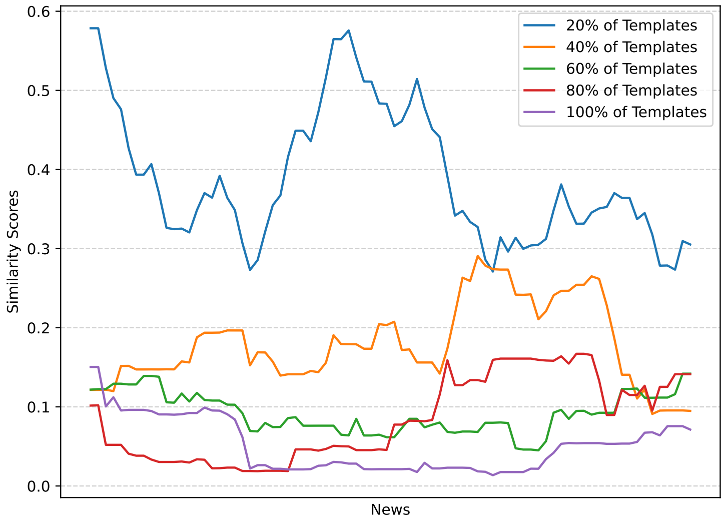


Figure 6.4: Comparison of text diversity in relation to template availability.

values presents in Table 6.1. Examining the mean values, we observe a decreasing trend in diversity scores as the template availability increases. This suggests that a higher number of available templates leads to a decrease in text similarity scores. Notably, the scenario with 100% template availability exhibits the lowest mean diversity score, indicating a higher level of similarity in the generated texts.

Consistent with the mean values, the median values also exhibit a decreasing trend as the template availability increases. The scenario with 100% template availability again shows the lowest median diversity score, reinforcing the finding that a higher number of templates results in a lower level of diversity. These findings support our initial hypothesis between template availability and text diversity. The results tend to indicate that a higher number of available templates leads to a decrease in diversity scores, signifying increased similarity in the generated texts.

Moreover, upon validating that none of the datasets followed a normal distribution, with corresponding p-values close to 0 (0, 0.005, 0.02, 0, and 0) for the Shapiro-Wilk tests, we proceeded to utilize the Kruskal-Wallis test, which is used for comparing the distribution of three or more independent groups in terms of significant difference. This test confirmed the presence of a significant difference among all datasets, yielding a p-value of 0.

These findings underscore the influence of template availability on text diversity in the generated news feeds. The results highlight the importance of considering the impact of template availability when aiming to achieve diverse and varied text generation with Prosebot or similar NLG systems.

Having performed three studies, where we evaluated the metric's performance regarding predefined assumptions about text diversity, we decided it would be beneficial to compare its output with human perception.

6.4 User-based Evaluation of Text Diversity

Human evaluation is widely recognized as the gold standard for assessing the quality and effectiveness of natural language generation (NLG) systems [53]. The subjective judgments provided by human evaluators offer valuable insights into the nuances and complexities of text generation, surpassing the limitations of automated metrics. While our metric incorporates various objective measures to evaluate text diversity, we recognize the importance of human judgment in capturing the intricacies and subtleties that automated metrics may overlook. To further enhance the reliability and robustness of our metric's evaluation, we introduced a complementary approach by employing a feedback form to solicit human feedback. Our goal was to gather information regarding human perception of text diversity and subsequently compare users' responses with the outputs of our metric, thereby assessing the degree of alignment between them.

This user-based evaluation serves as an essential tool in verifying and augmenting our metric's assessments. This additional layer of human evaluation not only serves the purpose of validating our metric's assessment but also provides insights into the alignment between automated analysis and human perception. Through this structured evaluation process, we seek to compare the judgments of human evaluators with the metric's automated assessments, facilitating the understanding of its accuracy in capturing text diversity in news feeds.

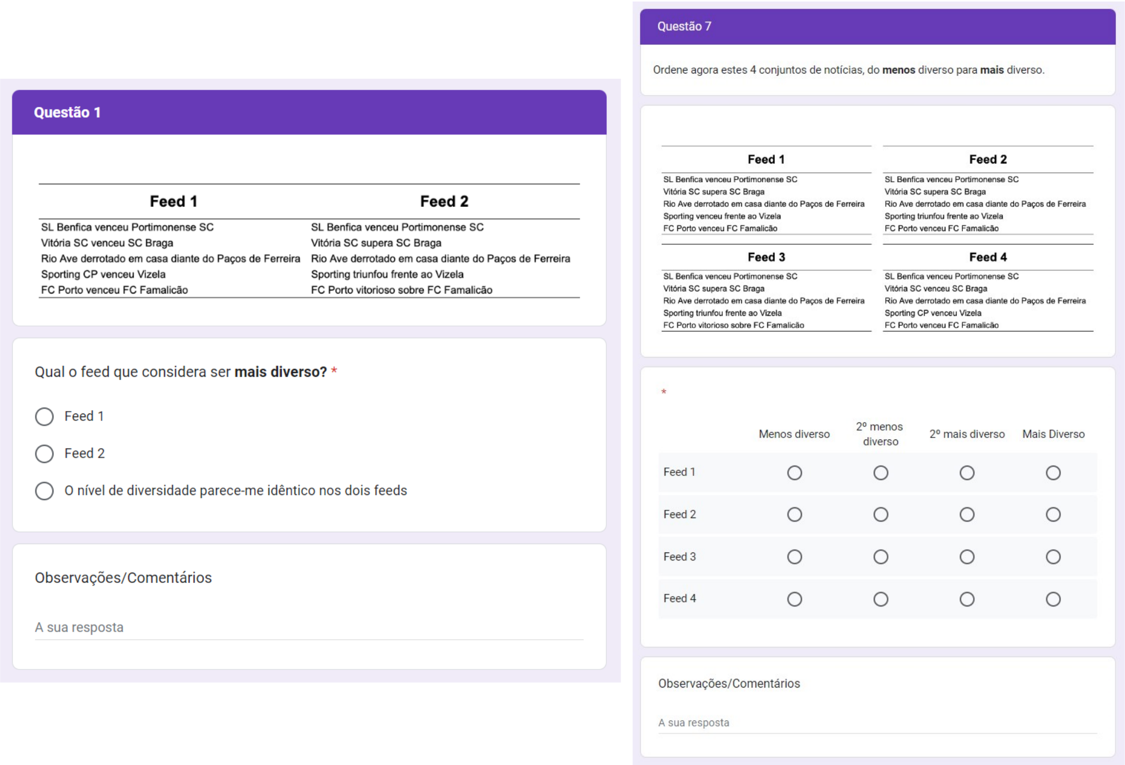
6.5 Methodology

In this section, we will outline the methodology employed to assess the alignment between human evaluation and the metric's assessment of text diversity.

6.5.1 Form Design

The form was designed to capture participants' perceptions of text diversity. To create the feeds for evaluation, we started by generating four news feeds using the rewriting architecture presented in Chapter 5.6. Moreover, each used feed represents the output from one of the iterations pertaining to the rewriting process, with each subsequent iteration exhibiting a higher level of text diversity compared to its preceding iteration. Accordingly, we expected participants to perceive feeds from earlier iterations to be less diverse than feeds from later iterations, which would be in alignment with our metric's assessment.

Concerning the form's questions, each question contained two different feeds, and participants were asked to evaluate and select which feed appeared to be more diverse: Feed 1, Feed 2 or whether they both seemed similar. Additionally, a seventh question required participants to rank the four different generated feeds from less to most diverse based on their perceived text diversity. Both types of questions can be seen in Figure 6.5.



Questão 1

Feed 1	Feed 2
SL Benfica venceu Portimonense SC Vitória SC venceu SC Braga Rio Ave derrotado em casa diante do Paços de Ferreira Sporting CP venceu Vizela FC Porto venceu FC Famalicão	SL Benfica venceu Portimonense SC Vitória SC supera SC Braga Rio Ave derrotado em casa diante do Paços de Ferreira Sporting triunfou frente ao Vizela FC Porto vitorioso sobre FC Famalicão

Qual o feed que considera ser **mais diverso?** *

○ Feed 1

○ Feed 2

○ O nível de diversidade parece-me idêntico nos dois feeds

Observações/Comentários

A sua resposta

Questão 7

Ordene agora estes 4 conjuntos de notícias, do **menos** diverso para **mais** diverso.

Feed 1	Feed 2
SL Benfica venceu Portimonense SC Vitória SC supera SC Braga Rio Ave derrotado em casa diante do Paços de Ferreira Sporting venceu frente ao Vizela FC Porto venceu FC Famalicão	SL Benfica venceu Portimonense SC Vitória SC supera SC Braga Rio Ave derrotado em casa diante do Paços de Ferreira Sporting triunfou frente ao Vizela FC Porto venceu FC Famalicão

Feed 3	Feed 4
SL Benfica venceu Portimonense SC Vitória SC supera SC Braga Rio Ave derrotado em casa diante do Paços de Ferreira Sporting triunfou frente ao Vizela FC Porto vitorioso sobre FC Famalicão	SL Benfica venceu Portimonense SC Vitória SC venceu SC Braga Rio Ave derrotado em casa diante do Paços de Ferreira Sporting CP venceu Vizela FC Porto venceu FC Famalicão

*

	Menos diverso	2º menos diverso	2º mais diverso	Mais Diverso
Feed 1	○	○	○	○
Feed 2	○	○	○	○
Feed 3	○	○	○	○
Feed 4	○	○	○	○

Observações/Comentários

A sua resposta

Figure 6.5: Single choice question example and ordering question.

6.5.2 Participant Selection

To ensure a representative sample for the data analysis phase, the form was distributed through ZOS' employees. Since the matter of discussion is related to news feeds, participants were categorized into two distinct groups: journalists and non-journalists. This categorization aimed to assess any potential variations in the evaluation based on participants' professional backgrounds.

6.6 Data Analysis

The collected data from the form was analyzed to evaluate the alignment between human evaluations and the program's assessment of text diversity. This analysis aims to provide an overview of the participant's responses and the distribution of their evaluations regarding text diversity. Additionally, we examined the composition of the participant sample, which was initially done without distinguishing between journalists and non-journalists. After this analysis, we checked to see any discrepancies in terms of answers for both sets.

6.6.1 Participants

A total of 25 participants completed the evaluation form. Among these participants, 10 identified themselves as journalists, while 15 were not affiliated with journalism or media-related professions. This distribution allowed for a comparison between the perspectives of individuals with expertise in the field and those with different backgrounds.

6.6.2 Quantitative Analysis

The quantitative analysis of the results began by examining the participant's responses to the multiple-choice questions, where they were asked to indicate which feed appeared more diverse (Feed 1, Feed 2, or both seemed similar). Figure 6.6 displays the summary of responses.

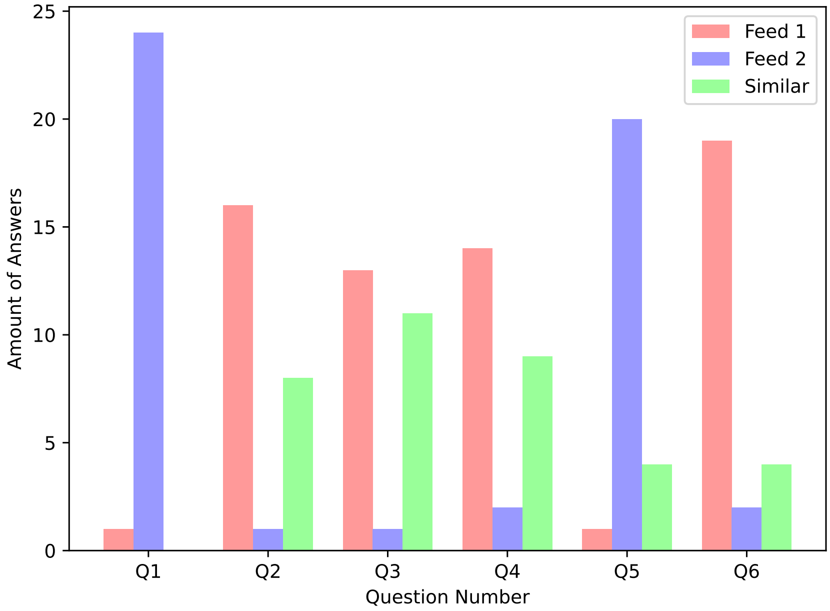


Figure 6.6: Distribution of participants' answers from questions 1 to 6.

The analysis of the participants' responses showed that there was a common consensus regarding text diversity perception, despite some questions having the participants more in agreement about which feed seemed more similar. For instance, question 1 had the majority of users (96%) depicting Feed 2 as more diverse, while, for example in question 3, there were more split opinions, where users were divided in terms of considering Feed 1 as more similar (52%) or considering both feeds as similar (44%). Taking a look at journalists and non-journalists answers individually, there were some discrepancies. Regarding the first two questions, both journalists and non-journalists exhibited a similar distribution of percentages for each choice. However, when examining the responses to the remaining questions, some differences emerged. Journalists tended to identify a higher percentage of feeds as similar compared to non-journalists. For example, for questions 3 and 4, 70% of journalists identified the feeds as being similar, while for non-journalists the percentage was only 33% and 20%, respectively. Likewise, non-journalists displayed a more distinct preference towards either Feed 1 or Feed 2, with a larger proportion of their responses favouring one feed over the other.

In order to interpret the results from the last question, where we asked participants to rank the feeds from less to most diverse, we converted each answer to a numeric representation. For example, if one of the answers was ’Feed 1 – Feed 2 – Feed 3 – Feed 4’, the matching representation would be ’1234’. The overall trends regarding the most common order chosen by participants can be seen in Figure 6.9, with the answer ’Feed 2 – Feed 3 – Feed 4 – Feed 1’ being the most common (80% of answers).

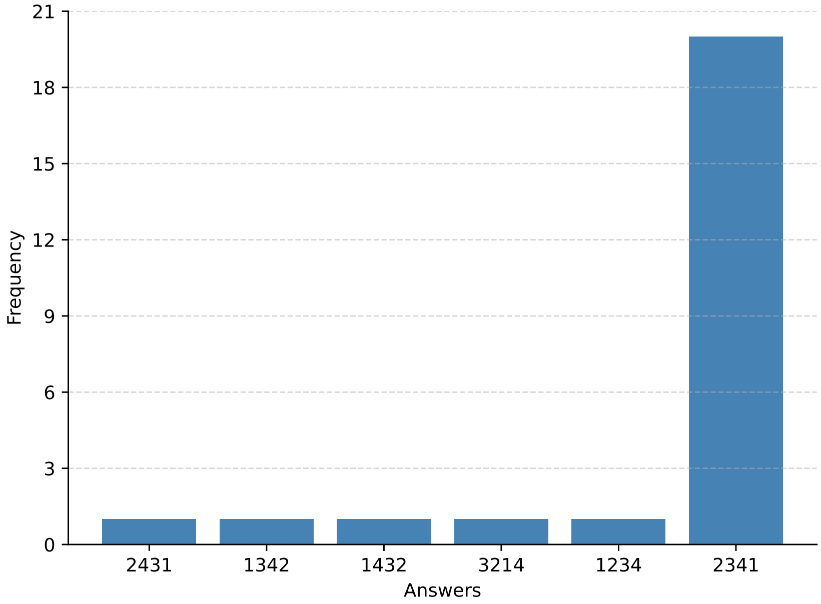


Figure 6.7: Question 7’s answers.

6.6.3 Comparison with Program’s Assessment

To compare the participants’ evaluations with the program’s assessment of text diversity, we conducted an analysis of the responses to each question. Figure 6.8 displays the results regarding the average of responses for each question alongside the metric’s perception for each question.

The analysis of the plotted line graphs revealed patterns that allowed us to observe a similarity regarding the overall shape of each line. This indicated an alignment between the participants’ evaluations and the metric’s assessment. The consistent equal shifts in terms of trends for both lines supported the idea that the metric had good accuracy in identifying text diversity, as perceived by the participants.

For Question 7, which involved ordering the feeds based on diversity, we used Kendall’s coefficient of concordance to compare participants’ answers and the metric’s assessment for this specific scenario. This measure is used to assess the agreement between two ranked sets of data. It quantifies the concordance or discordance between rankings, ranging from -1 to +1. A value of +1 indicates perfect agreement, -1 indicates perfect disagreement, and 0 indicates no significant agreement or disagreement. Using this coefficient, we computed the degree of agreement among

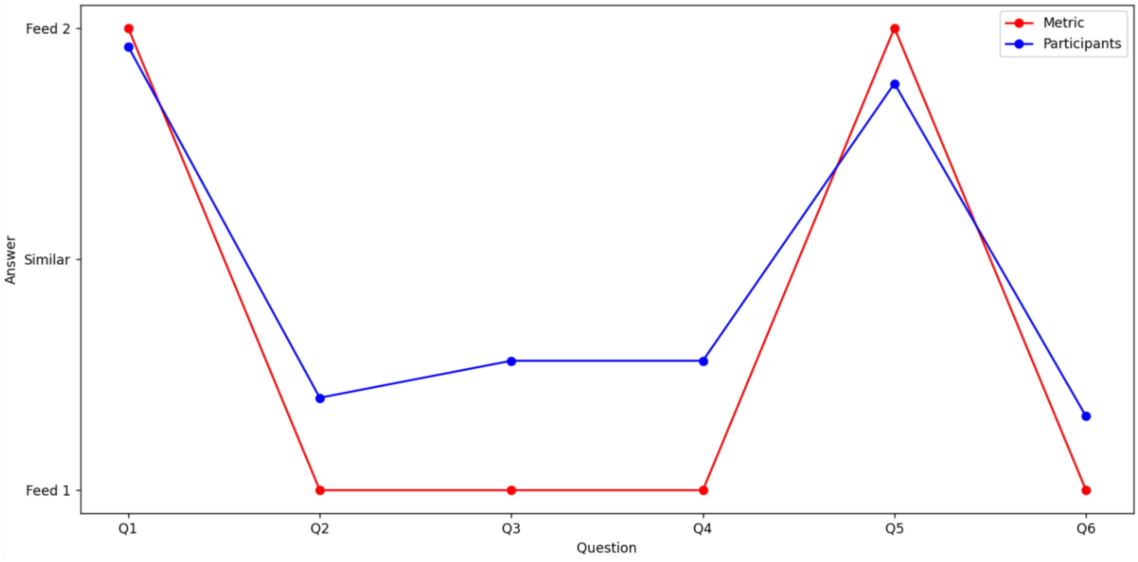


Figure 6.8: Comparison of participants' answers with the metric's assessment for questions 1 to 6.

the participants' rankings and the metric's assessment. The coefficient yielded a value of 0.83, indicating a high level of agreement, and further supporting the alignment between the participants' evaluations and the metric's assessment of text diversity. The high degree of agreement suggests that the metric employed in the program effectively captures the perceived diversity of the feeds, as reflected in the participants' rankings. Moreover, this alignment is visible in Figure 6.8, which depicts a correlation matrix regarding participants' answers and the metric's answer. This matrix exhibits a pronounced diagonal, indicating a positive correlation between the participants' answers and the metric's answer.

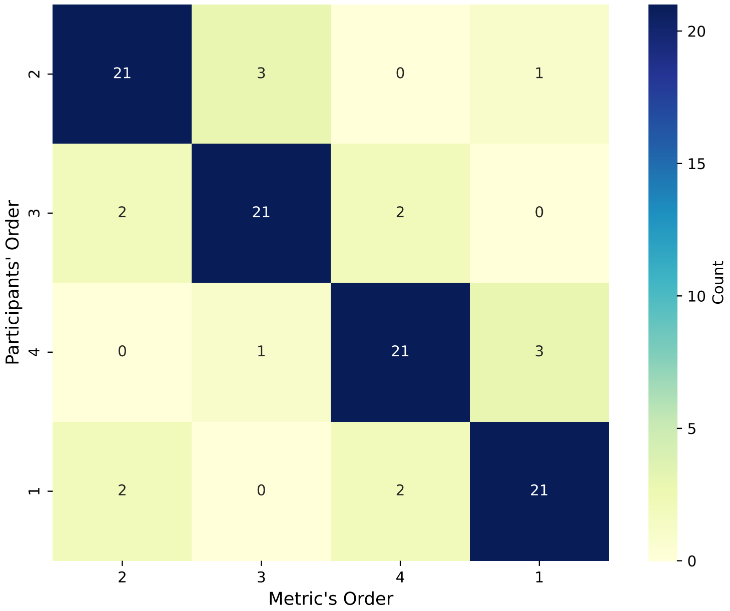


Figure 6.9: Comparison of participants' answers with the metric's assessment for question 7.

6.7 Conclusions

In this chapter, our objective was to measure and validate the accuracy of our metric in assessing text diversity. We conducted a series of studies to compare the metric's evaluation of text diversity with the expected results. The studies confirmed that the metric successfully provided a representation that aligned with our expectations, with the results from all of them being the expected ones.

Additionally, we also conducted a form-based evaluation to gather participants' perceptions of text diversity and further compare them with the metric's automated assessment. One important regard that should be taken into account is the fact that, for every question, there wasn't a case where the majority of answers weren't aligned with the metric's assessment, i.e. the majority of users identifying a feed as more diverse which hadn't been identified by the metric as such. Furthermore, it is worth noting that while there may be more disagreement when evaluating the diversity of individual feeds in isolation, there is a higher level of agreement when considering the ordering of the feeds as a whole. Regarding questions individually, journalists and non-journalists exhibited similar distributions of responses for the majority of questions. In spite of this, there were some small differences, especially in questions 2, 3 and 4, where journalists tended to perceive a higher percentage of feeds as similar, whereas non-journalists displayed a stronger preference for one feed over the other. This suggests that journalists may have a more nuanced perspective on text diversity, while non-journalists exhibit a clearer differentiation in their evaluations. Regarding the last question, which is related to ranking the feeds, we registered a high level of agreement between human judgment and the metric's perception of text diversity, based on Kendall's coefficient of 0.83. The consistency observed in the rankings supports the effectiveness of the program's metric in capturing the perceived diversity of the feeds.

Overall, the human evaluation process validated the metric's assessment of text diversity and provided insights into the alignment between automated analysis and human perception. The incorporation of human judgment enhanced the reliability of our evaluation, with its findings further strengthening the confidence in the metric's ability to accurately assess text diversity in news feeds.

Chapter 7

Conclusions

7.1 Final Remarks

In recent years, Natural Language Generation (NLG) systems have made remarkable strides, enabling computers to produce text that resembles human writing for a wide range of applications. These systems have become essential in our daily lives, serving as chatbots, virtual assistants, content creators, and data analyzers. However, a persistent challenge for researchers and developers is ensuring diversity in the generated text. Although NLG systems excel at producing coherent and contextually relevant content, achieving a sense of diversity remains an ongoing task. To tackle this issue, it is crucial to develop techniques and metrics that consider text diversity in NLG systems. By improving the evaluation methods for measuring diversity in automatically generated texts, we can gain a better understanding of the performance of each system and consequently use them to enhance the quality of automated content.

In this dissertation, we discussed the concept of text diversity and its significance in natural language generation. Through research and analysis, we successfully developed an automated metric that measures text diversity and seems to be well aligned with human judgments, as demonstrated in the conducted studies. We conducted a literature review regarding text diversity techniques and, additionally, defined clear objectives and goals for this study, which we can now say have been successfully accomplished. The next step in our work was related to the implementation of the proposed metric and its seamless integration within Prosebot, which we described in detail. We also documented a series of studies investigating text diversity in Prosebot, which not only shed light on the current text generation capabilities of the system but also validated and enhanced the robustness of the developed metric. Finally, we conducted a parallel study regarding user contribution in post-editing Prosebot-generated text.

Based on our research, we can now provide answers to the proposed research questions:

Q1: What factors contribute to the limited text diversity in Prosebot's current state? Taking into account that Prosebot is a template-based system, the most important aspect in determining text diversity within Prosebot-generated news pertains to the number of available templates, as well as the syntactical diversity between them. This became ever more obvious once we conducted the study which analyzed the relationship between template availability and text diversity in news feeds.

Q2: How can we effectively measure and evaluate text diversity in the context of Prosebot? While there might be more than a way of evaluating text diversity within Prosebot, the development of an automated metric capable of monitoring text diversity constitutes a useful, reliable and robust way of measuring text diversity within Prosebot, which is supported by the conducted studies throughout this work.

Having answered the research questions, we can now validate or refute our initial hypothesis, which stated that the *"development and implementation of a metric designed to assess text diversity will result in a measurable improvement in evaluating the overall diversity of generated content"*. The performed studies supported the idea that the developed metric offered a robust and accurate way to monitor and measure text diversity over time, proving its usefulness in evaluating NLG systems. Thus, the proposed hypothesis is validated.

7.2 Future Work

While our work represents progress in the field of text diversity evaluation, there are still areas that require further exploration and development. The following directions are suggested for future research:

1. **Refining the Automated Metric**: Although our automated metric shows promising results, there is room for improvement. Future work should focus on incorporating additional linguistic and contextual features to enhance the metric's accuracy and reliability. Exploring the effectiveness of other text similarity algorithms or machine learning techniques, such as deep learning models, could also contribute to refining the metric.

2. **Understanding Subjective Perceptions**: While our metric correlates well with human judgments, further investigation is needed into the subjective perceptions of text diversity. Future research should aim to understand how individual preferences, cultural factors, and personal backgrounds influence perceptions of diversity. This knowledge can help fine-tune the automated metric to better align with human evaluations.

3. **Real-World Applications**: The practical application of text diversity evaluation holds immense potential. Collaborating with industry partners and stakeholders can help integrate the automated metric into commercial natural language generation systems. This integration can enable the generation of more diverse, personalized, and engaging content, benefiting industries such as journalism, marketing, and creative content generation.

4. **Ethical Considerations**: As text generation technology advances, it is crucial to address ethical concerns. Future work should continue to investigate potential biases, unintended consequences, and the responsible use of text generation systems. Developing guidelines

and best practices for ethical text generation can ensure that these systems are employed in a fair, transparent, and accountable manner.

In conclusion, this dissertation contributes to the advancement of text diversity evaluation in the field of natural language generation. By considering text diversity as a fundamental criterion, we can promote the development of more diverse, inclusive, and engaging automated content generation systems. The future of natural language generation relies on leveraging the insights gained from this research to refine the automated metric, understand human perceptions, explore real-world applications, and address ethical considerations. Through these efforts, we can shape a future where text generation is not only accurate and fluent but also diverse, creative, and captivating.

References

[1] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A Neural Probabilistic Language Model. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000.

[2] Sonia Bergamaschi, Francesco Guerra, Mirko Orsini, Claudio Sartori, and Maurizio Vincini. Relevantnews: a semantic news feed aggregator. In Giovanni Semeraro, Eugenio Di Sciascio, Christian Morbidoni, and Heiko Stoermer, editors, *Proceedings of the 4th Italian Semantic Web Workshop, Dipartimento di Informatica - Universita' degli Studi di Bari - Italy, 18-20 December, 2007*, volume 314 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2007.

[3] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating Sentences from a Continuous Space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany, August 2016. Association for Computational Linguistics.

[4] Nadine Braun, Martijn Goudbeek, and Emiel Krahmer. The Multilingual Affective Soccer Corpus (MASC): Compiling a biased parallel corpus on soccer reportage in English, German and Dutch. pages 74–78, January 2016.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[6] David L. Chen and Raymond J. Mooney. Learning to sportscast: a test of grounded language acquisition. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 128–135, New York, NY, USA, July 2008. Association for Computing Machinery.

[7] Robert Dale. Natural language generation: The commercial state of the art in 2020. *Nat. Lang. Eng.*, 26(4):481–487, 2020.

[8] Robert Dale, Sabine Geldof, and Jean-Philippe Prost. Using natural language generation in automatic route description. *J. Res. Pract. Inf. Technol.*, 37(1), 2005.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran,

and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[10] Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. Jointly Optimizing Diversity and Relevance in Neural Response Generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1229–1238, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[11] Gartner. Neural networks and modern bi platforms will evolve data and analytics. Visited on 21-06-2023.

[12] Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Intell. Res.*, 61:65–170, 2018.

[13] Asghar Ghasemi and Saleh Zahediasl. Normality tests for statistical analysis: A guide for non-statisticians. *International journal of endocrinology and metabolism*, 10:486–489, 12 2012.

[14] Pierre Guillou. NLP | Modelos e Web App para Reconhecimento de Entidade Nomeada (NER) no domínio jurídico brasileiro. Medium [Visited on 21-06-2023].

[15] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *J. Documentation*, 60(5):493–502, 2004.

[16] Jenna Kanerva, Samuel Rönnqvist, Riina Kekki, Tapio Salakoski, and Filip Ginter. Template-free Data-to-Text Generation of Finnish Sports News. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 242–252, Turku, Finland, September 2019. Linköping University Electronic Press.

[17] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural language processing: state of the art, current trends and challenges. *Multim. Tools Appl.*, 82(3):3713–3744, 2023.

[18] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3294–3302, 2015.

[19] Anita Kumari and M. Shashi. Vectorization of text documents for identifying unifiable news articles. *International Journal of Advanced Computer Science and Applications*, 10:305, 08 2019.

[20] Quoc V. Le and Tomás Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1188–1196. JMLR.org, 2014.

[21] Leo Leppänen, Myriam Munezero, Mark Granroth-Wilding, and Hannu Toivonen. Data-Driven News Generation for Automated Journalism. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 188–197, Santiago de Compostela, Spain, September 2017. Association for Computational Linguistics.

[22] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California, June 2016. Association for Computational Linguistics.

[23] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, System Demonstrations*, pages 55–60. The Association for Computer Linguistics, 2014.

[24] Susan Mcroy, Songsak Channarukul, and Syed Ali. An augmented template-based approach to text realization. *Natural Language Engineering*, 9:381–420, December 2003.

[25] Marjan Mernik, Matej vCrepinvsek, Tomaž Kosar, and Damijan vZumer. Grammar-Based Systems: Definition and Examples. *Informatica*, 28:245–255, November 2004.

[26] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[27] Rashmi Mishra, Jiantao Bian, Marcelo Fiszman, Charlene R. Weir, Siddhartha Jonnalagadda, Javed Mostafa, and Guilherme Del Fiol. Text summarization in the biomedical domain: A systematic review of recent research. *Journal of Biomedical Informatics*, 52:457–467, December 2014.

[28] Ehsan Montahaei, Danial Alihosseini, and Mahdieh Soleymani Baghshah. Jointly measuring diversity and quality in text generation models. *CoRR*, abs/1904.03971, 2019.

[29] Changhoon Oh, Jinhan Choi, Sungwoo Lee, SoHyun Park, Daeryong Kim, Jungwoo Song, Dongwhan Kim, Joonhwan Lee, and Bongwon Suh. Understanding User Perception of Automated News Generation System. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Honolulu HI USA, April 2020. ACM.

[30] OpenAI. Gpt-4 technical report, 2023.

[31] Steffen Pauws, Albert Gatt, Emiel Krahmer, and Ehud Reiter. Making effective use of healthcare data using data-to-text technology. In Sergio Consoli, Diego Reforgiato Recupero, and Milan Petkovic, editors, *Data Science for Healthcare - Methodologies and Applications*, pages 119–145. Springer, 2019.

[32] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL, 2014.

[33] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics, 2018.

[34] Raquel Pires. Prosebot: o comentador de bancada baseado em inteligência artificial. `https://noticias.up.pt/prosebot-o-comentador-de-bancada-baseado-em-inteligencia-artificial/`, August 2021. Notícias U.Porto [Accessed June 26, 2023].

[35] Aurora Pons-Porrata, Rafael Berlanga Llavori, and José Ruiz-Shulcloper. Topic discovery based on text mining techniques. *Inf. Process. Manag.*, 43(3):752–768, 2007.

[36] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

[37] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[38] Ehud Reiter. An architecture for data-to-text systems. In Stephan Busemann, editor, *Proceedings of the Eleventh European Workshop on Natural Language Generation, ENLG 2007, Schloss Dagstuhl, Germany, June 17-20, 2007*, 2007.

[39] Ehud Reiter and Robert Dale. Building applied natural language generation systems. *Nat. Lang. Eng.*, 3(1):57–87, 1997.

[40] Ehud Reiter and Robert Dale. *Building Natural Language Generation Systems*. Studies in Natural Language Processing. Cambridge University Press, 2000.

[41] Ehud Reiter, Chris Mellish, and John Levine. Automatic Generation of Technical Documentation. *Applied Artificial Intelligence*, 9(3):259–287, May 1995. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/08839519508945476.

[42] Ehud Reiter, Somayajulu Sripada, Jim Hunter, Jin Yu, and Ian Davy. Choosing words in computer-generated weather forecasts. *Artif. Intell.*, 167(1-2):137–169, 2005.

[43] 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 Valentina Emilia Balas, Neha Sharma, and Amlan Chakrabarti, editors, *Data Management, Analytics and Innovation*, pages 333–344, Singapore, 2019. Springer Singapore.

[44] Huajie Shao, Jun Wang, Haohong Lin, Xuezhou Zhang, Aston Zhang, Heng Ji, and Tarek F. Abdelzaher. Controllable and diverse text generation in e-commerce. In Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia, editors, *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 2392–2401. ACM / IW3C2, 2021.

[45] Somayajulu Sripada, Ehud Reiter, and Lezan Hawizy. Evaluating an NLG system using post-editing. In Leslie Pack Kaelbling and Alessandro Saffiotti, editors, *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, July 30 - August 5, 2005*, pages 1700–1701. Professional Book Center, 2005.

[46] Somayajulu Sripada, Ehud Reiter, and Lezan Hawizy. Evaluation of an NLG System using Post-Edit Data: Lessons Learnt. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*, Aberdeen, Scotland, August 2005. Association for Computational Linguistics.

[47] Abhishek Sunnak. Evolution of natural language generation, Mar 2019.

[48] Yasufumi Taniguchi, Yukun Feng, Hiroya Takamura, and Manabu Okumura. Generating live soccer-match commentary from play data. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19, pages 7096–7103, Honolulu, Hawaii, USA, January 2019. AAAI Press.

[49] Guy Tevet and Jonathan Berant. Evaluating the Evaluation of Diversity in Natural Language Generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346, Online, April 2021. Association for Computational Linguistics.

[50] Mariët Theune, Esther Klabbers, Jan-Roelof de Pijper, Emiel Krahmer, and Jan Odijk. From data to speech: a general approach. *Nat. Lang. Eng.*, 7(1):47–86, 2001.

[51] Alan M. Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, 1950.

[52] Kees van Deemter, Mariët Theune, and Emiel Krahmer. Real versus template-based natural language generation: A false opposition? *Comput. Linguistics*, 31(1):15–24, 2005.

[53] Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Comput. Speech Lang.*, 67:101151, 2021.

[54] Chris van der Lee, Emiel Krahmer, and Sander Wubben. PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 95–104, Santiago de Compostela, Spain, September 2017. Association for Computational Linguistics.

[55] Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. A comprehensive capability analysis of GPT-3 and GPT-3.5 series models. *CoRR*, abs/2303.10420, 2023.

[56] Jin Zhang. *Visualization for Information Retrieval*, volume 23 of *The Information Retrieval Series*. Springer, 2008.

Appendix A

Analysis of user contributions in post-editing Prosebot text

While the use of NLG systems has been growing in the sports journalism industry, there is limited research on how end-users interact with these systems and the impact of post-editing on the overall quality of the generated text. In this study, the aim was to address this gap by analysing the post-editing behaviour of users who interact with Prosebot to generate news articles. Specifically, an analysis of the user changes to the automatically generated text and an investigation of the patterns and trends in the post-editing behaviour were performed. Text similarity algorithms were employed to measure the similarity between the original and edited texts. Moreover, part-of-speech (POS) tagging was used to understand which words were added or removed by users in the post-editing process. Finally, both outputs were compared in order to draw final conclusions.

By understanding how users interact with Prosebot, this study strives to provide valuable insights into how NLG systems can be optimised for sports journalism. This includes improvements to the content, style, and structure of the generated text. These insights can inform the development of more advanced NLG systems that can generate even more complex and nuanced content in the future.

A.1 Related Work

Previous research has investigated post-editing behaviour in data-to-text NLG systems. Sripada et al. [46] described a study in which human post-editing was used to evaluate the output of a natural language generation system that produced weather forecasts. The authors described their large-scale post-edit evaluation of the system, which involved generating draft forecasts and then having human forecasters post-edit them before releasing them to clients. The evaluation was conducted on 2728 forecasts collected over a three-month period (June to August 2003), and each forecast was approximately 400 words long. The authors argued that this approach could provide an accurate and reliable form of evaluating NLG systems, giving valuable insights into improving a system.

The evaluation process involved automatically breaking the forecasts into phrases and aligning corresponding phrases between the pre-edited (generated) and post-edited (edited) texts. The aligned phrases were then compared and labelled as matches, replacements, additions, or deletions. The authors analyzed a total of 73041 phrases, out of which 60% were perfect matches, 30% were mismatches and the remaining 10% couldn't be aligned by the procedure. The most common type of mismatch was ellipsis (word additions and deletions), with deletions being the majority of errors.

The article discussed the motivations behind using post-edit evaluation, including the belief that people would only edit things that were clearly wrong and the importance of post-editing as a metric for the system's usefulness to users. However, the authors discovered that post-editing was not solely for fixing mistakes but also involved refining and optimizing the texts, reflecting individual preferences, and addressing downstream consequences of earlier changes. They found the post-editing process useful for evaluating the system and gaining insights on how to improve it. Furthermore, the paper emphasized the significance of post-edit evaluation in assessing the practical usefulness of NLG systems, stating that the post-edit evaluation technique proved effective in evaluating the system and providing actionable feedback for future enhancements. In a follow-up paper [45], the authors provided a more detailed analysis of the results of their previous study, along with insights and lessons learned from the experience. They discussed the strengths and limitations of using post-editing as an evaluation method.

It was found that post-edit evaluations were not effective in identifying specific problems due to the presence of noisy post-edits that did not fix the generated text. The noise was attributed to the lack of post-editing support tools and the acceptance of post-edit data with noise. Integrating post-editing tools into natural language generation (NLG) systems was suggested to reduce noisy post-edits and focus on genuine corrections. The importance of understanding individual post-editing behaviour and conducting a pilot study to analyze noisy post-edit data was emphasized. The authors also mentioned the need for preparation in terms of developing post-edit tools and carrying out pilot studies. The overall cost of the post-edit evaluation was deemed reasonable compared to conducting end-user experiments on a large number of texts, but it noted that cost-effectiveness could vary if evaluators had to organize and pay for post-editing. The authors speculated that as more NLG systems are deployed, post-editing would become an accepted component of the automatic text generation process, similar to its role in machine translation (MT).

This study's approach is complementary to the previously referred studies in the sense that the presented methodology investigates the post-editing behaviour of users interacting with Prosebot to generate news articles. This study aimed to understand the impact of post-editing on the final output quality, which contributes to comprehending how the system can be improved. While Sripada et al. might have explored different aspects or employed alternative analysis techniques, this study employed cosine similarity combined with TF-IDF to measure the similarity between automatically generated and post-edited texts. Additionally, a part-of-speech tagging analysis was conducted to identify the most commonly modified words by users. By comparing the final output of the original and post-edited versions, this analysis sheds light on the text diversity achieved through the introduced changes. Its findings complement the existing research and contribute to a comprehensive understanding of the impact and effectiveness of post-editing in automated text generation. Additionally, since the target NLG system was Prosebot, this work was conducted within a real-world context, utilizing a system that is already in use and production for news creation. The analysis was done over 2,500 pairs of generated texts, which provides a comprehensive dataset for this study.

A.2 Document Collection

To conduct this study, the used dataset was provided by ZOS, which comprised 2,734 rows, where each entry represents a piece of news. Among these entries, the vast majority, totalling 23,320 entries (equivalent to 84.86% of the total), were related to football. Additionally, there were 272 entries (representing 9.95%) associated with 7-a-side football, while 137 entries (making up 5.01%) were focused on futsal. Lastly, a minimal proportion of just 5 entries (accounting for 0.18%) were specifically dedicated to U-9 football.

Each entry within the dataset consisted of a pair of titles, comprising the original title (generated by Prosebot alongside its corresponding post-edited version (by zerozero.pt's users). Likewise, a pair of summaries were provided, encompassing the original summary and its post-edited counterpart. The dataset also contained additional information, such as the author of each news article and the publication date. Spanning from January 2021 to February 2023, the news articles included in the dataset offered a diverse and current assortment of content for analysis. To ensure consistency and relevance of the data, preprocessing steps were applied to the text, including tokenization, removal of stopwords, stemming, and lemmatization. These preprocessing techniques were implemented with the objective of improving the quality of the textual data for subsequent analysis.

When examining the average length of titles, some discrepancies emerge between the original and post-edited versions. Regarding football entries, the average word count for the original titles stands at 7.90, whereas the post-edited titles exhibit a slightly higher average of 8.08 words. Similarly, for 7-a-side football entries, the average length of original titles amounts to 12.0 words, which experiences a slight increase to 12.12 words in the post-edited versions. In the case of futsal entries, there is also a noticeable increase, with the original titles averaging 7.77 words, while the post-edited titles reach an average of 8.42 words. Lastly, both the original and post-edited titles of U-9 football entries exhibit an average length of 12.6 words.

Turning to the summaries, it is possible to observe similar patterns. The original summaries for football entries have an average length of 166.49 words, while the post-edited versions show a slight increase to an average of 182.22 words. For 7-a-side football entries, the average length of the original summaries is 224.07 words, which rises to 228.34 words in the post-edited summaries. Futsal entries have an average of 211.29 words in the original summaries, increasing to 216.77 words in the post-edited versions. Lastly, the U-9 football entries maintain an average length of

Table A.1: Overview of the document collection.

Sport	**Percentage (%)**	**Avg. Word Count**		**Avg. Sentence Count**		**Avg. Number of Words per Sentence**	
		Original	Post-Edited	Original	Post-Edited	Original	Post-Edited
Football	84.84	7.99	8.08	6.67	8.74	26.35	22.02
7-a-side football	9.96	12.11	12.12	10.21	12.47	23.35	19.58
Futsal	5.02	7.82	8.42	10.48	11.23	21.39	20.53
U-9 football	0.18	12.6	12.6	8.0	8.0	23.65	23.65

180.2 words in both the original and post-edited summaries. Table A.1 provides further details and additional data on the subject.

These initial findings serve as a significant indicator for the forthcoming analysis. The disparities in average length offer valuable insights into the alterations made during the post-editing stage and their potential ramifications for readability and information delivery.

Furthermore, it is worth emphasizing that a substantial portion of the dataset has undergone post-editing, encompassing 1152 titles and 2215 summaries. This underscores the extensive scope of the post-editing process and its impact on the dataset's content. These early findings provide a strong foundation for further analysis and exploration of the dataset's characteristics and implications.

In addition to the information provided above, an analysis was conducted to examine the distribution of news articles edited by each user in the dataset. This analysis aimed to assess whether the distribution followed a tail pattern, where a few users contribute a large proportion of the articles while the majority of users contribute a relatively small number.

The distribution of news creation by users indeed followed a tail distribution, with a small number of users contributing a significant portion of the articles. The distribution shows that, out of the edited news articles in the dataset, approximately 75% were edited by the top 10% users, and out of this, 80% were related to the top 5 users, This finding highlights the presence of a skewed distribution, with a few highly active users responsible for generating a substantial amount of content. Furthermore, while analyzing each user's edited news, we came to the conclusion that they tended to pertain to the same competitions, which made sense since users will tend to focus on their personal interests. Figure A.1 displays the cumulative percentage of edited news articles against the cumulative percentage of users. This graph demonstrates a steep curve indicating the dominance of a small group of users in terms of news production. The curve gradually levels off as we move towards the majority of users who contribute a smaller number of articles.

A.3 Methodology

This study aimed to investigate the post-editing behaviour of users who interact with Prosebot to generate news articles and understand the impact of post-editing on the quality of the final output. To achieve our research objectives, several tasks were carried out that allowed for the comparison of the level of similarity between automatically generated and post-edited texts.

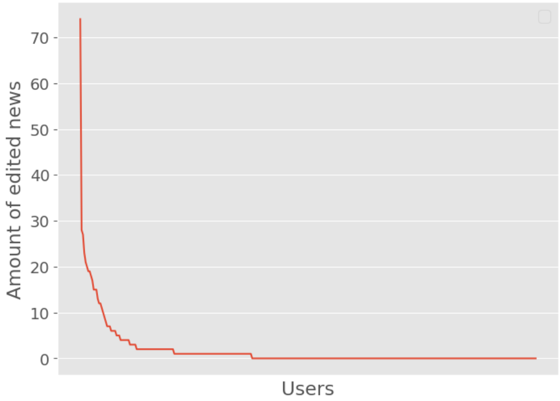


Figure A.1: Distribution of edited news per users.

The first step was to conduct a preliminary analysis of the similarity between both versions of the texts. To compare their similarity level, cosine similarity combined with TF-IDF was employed, which reflects a commonly used approach in natural language processing that is able to measure different aspects of text similarity. This analysis of the similarity scores aided in identifying patterns and trends in the types of post-edits made by users and factors that influenced the level of post-editing required.

The study's findings suggest that users tend to change the titles of news articles to a greater degree than their summaries. Most changes to titles involved either adding or removing words, while changes to summaries were mostly rewording or paraphrasing. These results suggest that users may be more concerned with generating attention and engagement for their posts rather than providing extra information on match reports. A more comprehensive analysis will be provided in subsequent sections of this document.

After conducting the preliminary analysis of the degree of similarity between pairs of generated titles and summaries and their post-edited versions, the goal was to gain further insight into the types of changes made by users. To do this, a part-of-speech (POS) tagging analysis was conducted. This involved analyzing the frequency of POS tags added or removed in the post-edited texts. By doing so, it was possible to identify which types of words were most commonly modified by users. In addition, the distribution of POS tags was computed in the original and post-edited texts to identify any patterns and trends. This allowed for a better understanding of the types of changes made by users and identified any areas where Prosebot may produce less-than-ideal output.

The final step of the analysis was to compare the final output between the original titles and summaries and their post-edited versions. This analysis was helpful in truly understanding whether the introduced changes had, in fact, produced a better output in terms of text diversity. To evaluate whether this was true, the average sentence length was compared with the average number of words in each text. Moreover, Herdan's C algorithm was applied to the obtained result, which allowed for the quantitative evaluation of the impact of post-editing on text diversity.

A.4 Results

Our initial hypothesis was that the titles generated by Prosebot would undergo more post-edits than the summaries and, therefore, would differ more from their original versions, as the titles are less time-consuming and require less effort to change. Since summaries contain more content and detail, users might feel altering these more extensive texts would be more complex and require more effort.

A.4.1 Preliminary similarity comparison

Table A.2: Distribution of cosine similarity of scores per bins of 0.1 for titles and summaries, respectively.

Intervals	Titles	Summaries
0.0-0.1	60	1
0.1-0.2	0	1
0.2-0.3	1	1
0.3-0.4	11	2
0.4-0.5	47	4
0.5-0.6	37	12
0.6-0.7	45	19
0.7-0.8	81	30
0.8-0.9	2303	105
0.9-1.0	79	2556

As stated in the previous chapter, the starting point for this analysis was to compare both sets of titles and summaries using popular state-of-the-art methods. Table A.2 contains the distribution of cosine similarity scores, categorized according to intervals of 0.1, which are also visible as bar charts in Figure A.3. Furthermore, since most of the cosine similarity scores are comprised between 0.80 and 1.00 (89% of scores regarding titles and 97.5% for summaries), this figure provides a visual depiction, in the form of line graphs, of the distribution of scores regarding lower values, to better differentiate both sets.

Based on the titles' distribution of cosine similarity scores, there is a greater proportion of lower scores compared to the distribution for the summaries (4.5% of titles scores exhibit a score below 0.5, while this distribution is only 0.3% for summaries), indicating that a higher number of changes were made to the text. These lower scores suggest substantial modifications and variations in the titles, whereas the summaries exhibit a larger proportion of scores in the higher range, implying relatively fewer textual alterations. Based on these results, it appears users seem to make more significant changes to titles than to summaries.

In terms of patterns, it is interesting to note that there are peaks in the distribution at certain intervals for both titles and summaries. In the case of titles, the peak at 0.00-0.05 indicates that

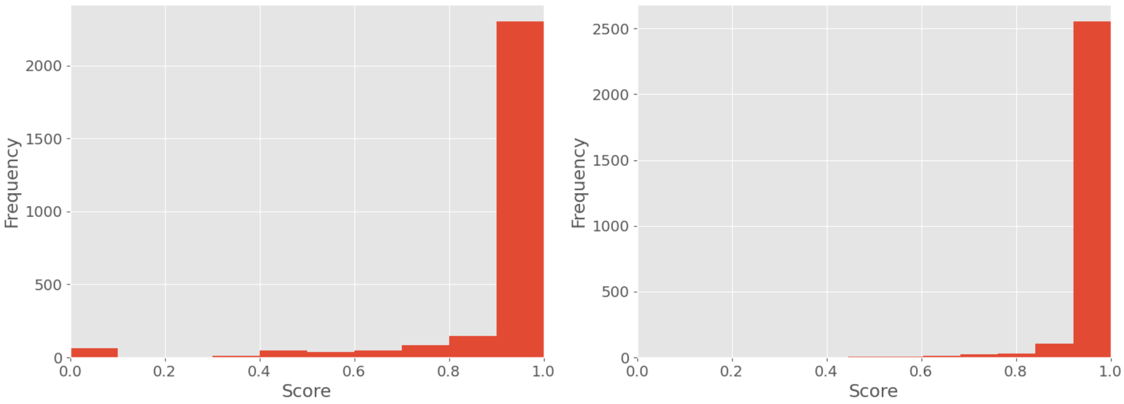


Figure A.2: Distribution of cosine similarity per bins of 0.1 for titles and summaries, respectively.

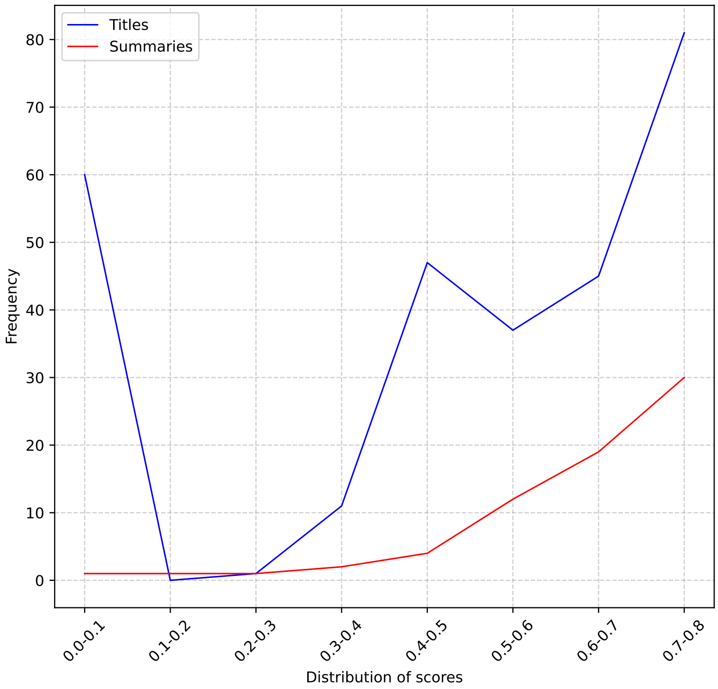


Figure A.3: Distribution of cosine similarity per bins of 0.1 for titles and summaries, respectively, considering an interval from 0.0-0.8.

some users are making significant changes to them, while the peaks at 0.40-0.45 and 0.95-1.00 suggest that some users are making only minor or no changes at all. As for the summaries, there is a clear peak at the highest score interval (0.95-1.00), indicating that most users are making only minor changes or no changes to the generated summaries. However, there are also smaller peaks at intervals of 0.80-0.90, 0.90-0.95, and 0.75-0.80, suggesting that some users are making some changes to the summaries.

Overall, it seems that users are generally satisfied with the generated summaries, as they are making relatively few changes to them. However, there is more variation in the changes made to the titles, with some users making significant changes while others making only minor or none. These findings suggest that Prosebot is a useful tool for generating match summaries but that some users may prefer to make their own editorial changes to the generated text to fit their needs or style preferences better.

Regarding the initial hypothesis, a possible explanation for the results can be that users might consider the title to be the most important part of the news article as it is the first thing that catches the reader's attention. Hence, users may spend more time and effort in making it more appealing and informative. Another possible explanation may be related to titles being typically shorter than summaries, making them easier to edit. Furthermore, users may feel overwhelmed by the length of summaries and not know where to start when making changes. While titles need to be clear and concise to effectively communicate the topic of the article, summaries need to provide more details, which can make users feel more confident in making changes to them rather than summaries. On the whole, users may feel more inclined to personalize the title to their own preferences, while the summary may be more fact-based and require less personalization.

It is worth noting that the results of our analysis may not be generalizable to other domains or applications. The needs and preferences of users may vary depending on the context, which could influence the extent and nature of the post-edits made on automatically generated texts. Further research is needed to explore these issues in different contexts.

A.4.2 Part-of-Speech (POS) Tagging

The results from the POS tagging analysis are presented in figure A.4, which display the distribution of added and removed keywords in titles and summaries. These results show that users tend to add and remove a large number of nouns and verbs, both in titles and summaries. This was already expected since nouns and verbs are vital for constructing meaningful sentences in natural language. They are the building blocks of language and are necessary for conveying information and communicating effectively. Based on this information, the focus here was to identify keywords that could have been more frequently added/removed than what would be expected. Furthermore, proper names related to teams were frequently removed, which could suggest that some users may prefer headlines that focus on the team rather than individual players. For instance, a user decided to alter a title from:

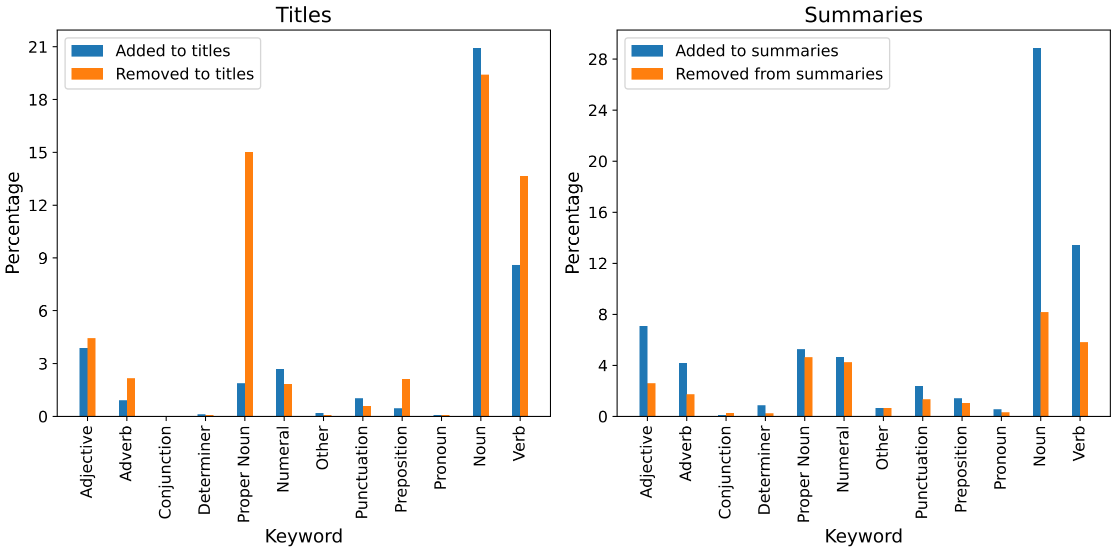


Figure A.4: Histogram of added and removed keywords to titles and summaries.

Juventude da Portelinha levou a melhor sobre o Santa Catarina (*Juventude da Portelinha got the best over Santa Catarina*)

To the following:

Tarde de gala em Gondomar (*Gala afternoon in Gondomar*)

This kind of change indicates that some users might be more interested in capturing the overall experience or atmosphere of a particular event.

One interesting finding is the high frequency of added numerals in titles (8.38%), which could suggest that users are adding specific statistics or scores to the title to make it more informative or attention-grabbing. For instance, a title that initially read:

Vila FC derrotado pelo Sport Canidelo (*Vila FC defeated by Sport Canidelo*)

Was transformed to :

SC Canidelo goleia Vila FC por 3-0 (*SC Canidelo thrashes Vila FC 3-0*)

Similarly, another entry originally titled:

Duelo entre Alta de Lisboa e Desportivo O. Moscavide não foi além de um empate (*The clash between Alta de Lisboa and Desportivo O. Moscavide didn't go beyond a draw*)

Was changed to:

6 minutos de compensação e golo do empate aos 7 minutos após os 90 (*6 minutes of stoppage time and the equalizing goal at the 97th minute, after the 90th minute*)

Additionally, the relatively high frequency of added adjectives in titles (10.52%) compared to removed adjectives (14.29%) is noteworthy. This could indicate that users are trying to make the titles more descriptive and engaging. Furthermore, the high frequency of removed proper names (16.03%) as opposed to their additions (3.34%) is obvious, which could suggest that users prefer headlines that focus on the team rather than individual players. This behaviour would explain why proper nouns related to players are frequently removed. However, it's also possible that users are removing player names because they're already mentioned elsewhere in the article, and the headline doesn't need to repeat them. It's also worth noting that while the frequency of added conjunctions and determinants is relatively low in titles, they were still added more frequently than removed. This could suggest that users are trying to improve the flow and coherence of the title by adding connecting words and articles.

When it comes to editing summaries, adjectives, numerals and adverbs tend to be the focus of additions, with 3527 adjectives (12.7%), 1909 numerals (6.87%) and 1518 (5.47%) adverbs added. The impression this gives is that users are making an effort to enhance the descriptive and vivid nature of the text by adding more descriptive words. Regarding removed keywords, numerals are removed significantly more than expected, with 2,120 numerals removed, which suggests that users are perhaps simplifying the text by removing numbers, which may be less relevant or less informative than other types of words.

Overall, based on the available evidence, it appears that *Prosebot* users prioritize the content of their news rather than the stylistic quality or fluency of the text. While this tendency is commonly observed in this particular domain, its generalizability to other domains remains uncertain. In the context of news pieces for lower league games, the concise and straightforward text is deemed more appropriate, as opposed to a detailed analysis that would be more suitable for high-profile matches like the Champions League. Additionally, it is probable that Prosebot users prefer not to invest significant time in writing match summaries, although exceptions may exist.

A.4.3 Output comparison

In order to deduce whether the differences between both sets of original and post-edited texts were statistically relevant and not due to randomness, the Shapiro test [13] was employed. This test is designed to assess the normality of a distribution, specifically, whether the data follows a normal distribution or not. By applying the Shapiro test to our dataset, the null hypothesis that the data is normally distributed against the alternative hypothesis that it is not was evaluated. The results of the Shapiro test indicated that the distributions were not normal, providing evidence that the data significantly deviated from a normal distribution. This finding justified the decision to employ the non-parametric Wilcoxon test [13].

The p-value obtained from the Shapiro test was 0, providing strong evidence against the null hypothesis and suggesting a departure from normality. Similarly, the p-value obtained from the Wilcoxon test was also 0, indicating a significant difference between the original and post-edited texts. These p-values reinforce our conclusion that the observed disparities are indeed statistically relevant and not merely the result of randomness. The Wilcoxon test, as a robust non-parametric alternative, strengthened the validity of our findings by accounting for the distributional assumptions.

The Wilcoxon test, also known as the Wilcoxon signed-rank test, is a non-parametric statistical test used to compare paired samples or matched data. Unlike parametric tests, the Wilcoxon test does not rely on the assumption of normality. Instead, it focuses on the ranks or relative order of the data values. By comparing the ranks of the paired differences between the original and post-edited texts, the Wilcoxon test determines whether there is a significant difference between the two sets.

In this analysis, the Wilcoxon test provided further confirmation of the statistical relevance of the observed differences between the sets of texts. With a p-value of 0, below our predetermined alpha level of 0.05, it is possible to conclude that the disparities between the original and post-edited texts are indeed statistically relevant and not merely the result of randomness. The Wilcoxon test, as a robust non-parametric alternative, strengthened the validity of these findings by accounting for the distributional assumptions.

The initial prediction was that users would introduce changes that would improve the overall diversity in the automatically generated text. Contrary to our previous hypothesis, these findings support this one, as is shown in Figure A.5. Although most of the original and post-edited versions of the news display similar scores for Herdan's C algorithm, there are some cases where scores differ vastly. The outliers on the graph demonstrate that some users are willing to spend more time and extra effort on personalizing their own texts. This depends entirely on the users' own needs as well as interest in reporting matches or available time to do so.

In order to assess the impact of the text size on the changes made during post-editing, we conducted an analysis by dividing the texts into ten distinct buckets based on their size. This division allowed us to explore the relationship between text size and the modifications made in the post-edited versions. Within each bucket, we computed the cosine similarity and the proportion of changes between the original and post-edited texts. The results of this analysis are summarized in Table 5.

Our analysis yielded interesting findings, shedding light on the relationship between text size and the extent of changes made during post-editing. The proportions of changes, indicating variations between the post-edited versions and the original texts, exhibited variations across different text size buckets. Notably, the smallest bucket, encompassing texts ranging from approximately 40 to 111 words, displayed the highest proportion of changes, reaching around 82-83%. However, the proportions of changes varied within the different bucket ranges, ranging from approximately 73% to 83%.

The average cosine similarity showed an increasing trend as text size increased. The smallest bucket, comprising texts of 40 to 111 words, exhibited the lowest similarity score, approximately 0.93, while the largest bucket, spanning texts of approximately 226 to 575 words, demonstrated the highest similarity score, around 0.98. These results indicate that the average similarity scores generally rise with increasing text size.

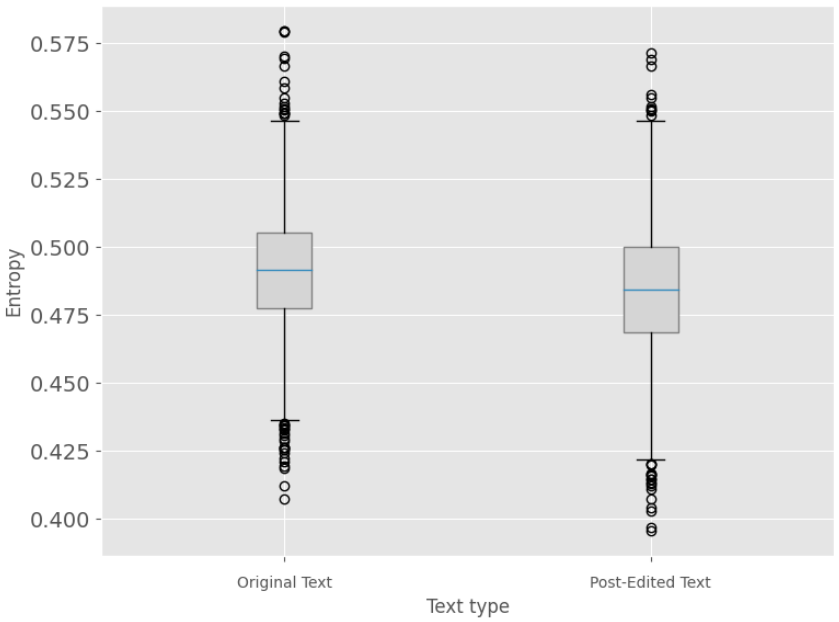


Figure A.5: Herdan's C scores for original and post-edited texts.

The outcomes of our analysis reveal a notable correlation between text size and both the proportion of changes made during post-editing and the similarity between the original and post-edited versions. Smaller texts tend to undergo a higher proportion of changes and exhibit lower similarity scores, whereas larger texts display lower proportions of changes and higher similarity scores.

These findings provide robust support for our initial hypothesis that larger texts are less susceptible to modifications during post-editing and exhibit higher similarity with the original texts.

A.5 Discussion and Conclusions

In this study, the impact of post-editing on automatically generated texts was investigated. Regarding the initial hypothesis, the results were unexpected. It had been anticipated that the larger the text, the more prone it would be for changes and the more users would alter it. However, the results unveiled patterns that challenge the initial assumptions and open new avenues for discussion.

Upon analyzing the data, a relationship between text size and the extent of changes made during post-editing was discovered. Contrary to the predictions, smaller texts consistently displayed higher proportions of changes, indicating a greater likelihood of modification. These findings suggest that users perceive smaller texts as more malleable and are more inclined to make adjustments to them. On the other hand, larger texts exhibited lower proportions of changes, implying a higher level of content satisfaction or a reduced need for substantial modifications. This unexpected pattern raises intriguing questions regarding the users' perception of text size and their decision-making processes during post-editing.

Furthermore, the investigation revealed a correlation between text size and the similarity between the original and post-edited versions. As text size increased, the average cosine similarity scores also increased, indicating a higher degree of resemblance between the two versions. This suggests that larger texts possess inherent qualities that align closely with user preferences, resulting in fewer alterations during the post-editing phase. These insights emphasize the importance of considering text size as a contributing factor in the post-editing process and its implications for the quality and fidelity of automatically generated texts.

The unexpected findings in this study prompt further exploration into the underlying mechanisms that influence post-editing decisions. Future research could delve into user preferences, motivations, and contextual factors that contribute to the observed patterns. Additionally, investigating the impact of text size on other aspects of post-editing, such as readability, coherence, and overall user satisfaction, could provide a more comprehensive understanding of the dynamics at play.

Overall, this study highlighted the nuanced relationship between text size and post-editing outcomes, challenging existing assumptions and paving the way for future investigations. These findings underscore the need for a deeper understanding of user behaviour and preferences in the context of post-editing, ultimately guiding the development of more effective and user-centric approaches to automatically generated texts.