A Community-Driven Data-to-Text Platform for Football Match Summaries

Pedro Fernandes[1], Luís António Santos[2], Sérgio Nunes[1]

[1] INESC TEC · FEUP, University of Porto
[2] ICS, University of Minho

INESCTEC

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Universidade do Minho
Instituto de Ciências Sociais

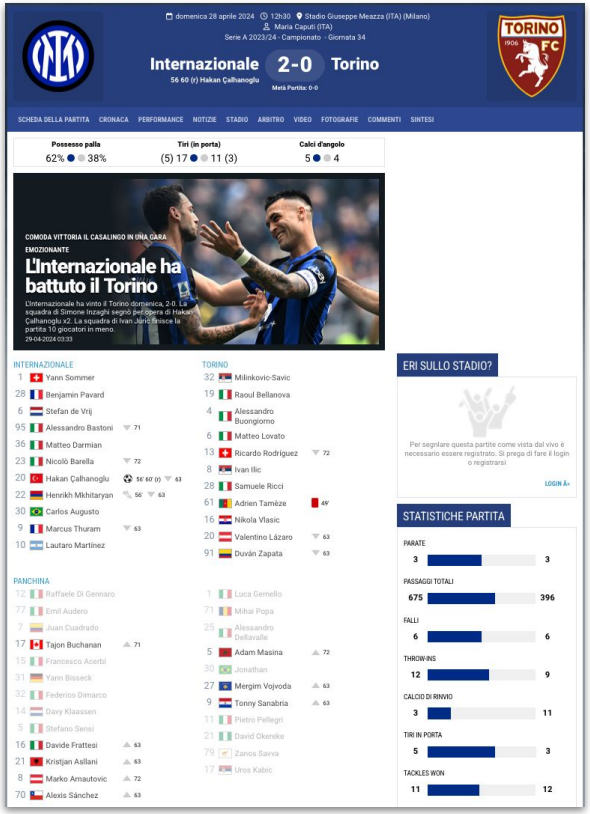
LREC-COLING 2024 | Turin, Italy

Introduction

This work describes the design and deployment of Prosebot, a community-driven data-to-text platform tailored for generating textual summaries of football matches derived from match statistics.

Prosebot uses a template-based Natural Language Generation module to produce initial matich reports, which are subsequently refined by the reading community.

The system enhances the visibility of lower-tier matches, traditionally accessible only through data tables, greatly increasing the production and publication of textual summaries.



From match data (left) to a match report (right).

zerozero.pt

zerozero.pt is a prominent Portuguese sports portal, offering comprehensive access to match-related data and statistics, encompassing teams, players, goals, substitutions, fouls, and corners.

It also provides news coverage for premier competitions, curated by a team of journalists.

This coverage encompasses merely 1.4% of the matches cataloged in zerozero.pt 's database.

To take advantage of the vast volume of data, zerozero.pt pioneered the development of Prosebot, an automated system designed to craft text-based match reports predicated on structured match data.

This work presents the Prosebot platform, highlighting its role in engaging zerozero.pt's vast community in the creation of match summaries using automatically generated texts as a starting point.

Related Work

Natural Language Generation techniques have gained significant traction in journalism.

⇒ Academia: SumTime-Mousam [weather] (2005), Plachouras et al. [finance] (2016), PASS [football] (2017).

⇒ Commercial solutions: Arria Sports, WordSmith, Quill.

⇒ Application examples: LA Times [crime reports + earthquakes] (2015), BBC [elections] (2019)

One important distinction of the Prosebot platform is the involvement of the community of readers in the process.

Post-Editing in Text Generation

The Prosebot platform is original in the combination of NLG and the readers in crafting sports texts.

Related work can be found in post-editing and machine-in-the-loop story generation.

⇒ Stripada et al. (2005) report on the post-editing of machine generated weather reports.

⇒ Machine translation (Vaswani et al., Sennrich et al.) and automatic grammar checking (Dale et al.) are areas where text-to-text NLG is commonly used to support user work.

⇒ Machine-in-the-loop examples include text generation systems to support creative work, such as STORIUM (Akoury et al.), Writing Buddy (Samuel et al.) and Creative Help (Roemmele and Gordon).

Journalists' Perceptions towards Natural Language Generation

Survey to Journalists

To inform the design of the platform, we conducted a survey to journalists in zerozero.pt's newsroom.

This survey also aimed to gauge their perceptions regarding automated content generation tools.

The survey was conducted in 2021 and garnered responses from fifteen (15) journalists.

⇒ the survey revealed that 86.7% of the journalists believed that summaries should undergo approval before publication.

⇒ 73.3% of the journalists concurred that both the collaborators and Prosebot should be acknowledged as co-authors. Such an approach ensures transparency, allowing readers to be informed of all contributors to the text.

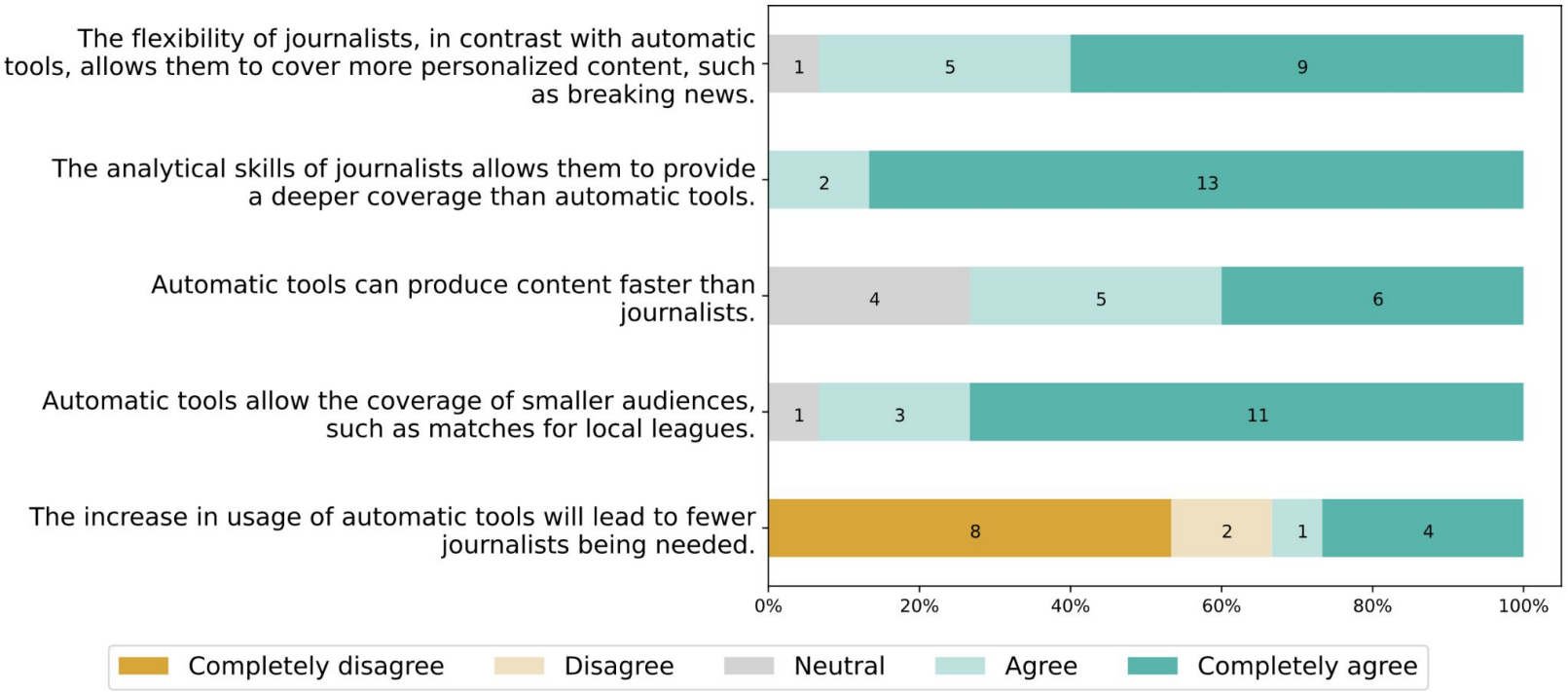


Figure 1: Overview of survey to zerozero.pt's journalists.

Prosebot Platform

Prosebot Platform

Data source ⇒ detailed mach information, including participating teams, competition, final score, goal contributors, antecedent match outcomes, and player infractions from the zerozero.pt database.

Template design ⇒ developed through a collaborative effort between zerozero.pt's engineering team and its journalists. Each template encompasses both content and an associated condition.

Document structure ⇒ generated match report is organized into seven sections: (1) title, (2) subtitle, (3) small text, (4) introduction, (5) events, (6) debriefing, and (7) trivia.

Prosebot Templates

Prosebot evaluates template conditions to decide if the content should be incorporated into the generated text.

This approach has been particularly important in diversifying the template pool, thereby reducing textual monotony.

Section	Category	Variants
Title		All-purpose, 2+ goal difference, 4+ goal difference, home team won, away team won, game ended in draw, match ended with penalties
Subtitle		All-purpose, favourite team lost, decisive goal in the last minutes, no goals scored, many goals scored
Small text		Final score, starter and benched relevant players
Introduction		Final score, previous results, best player
Events	Goal	All-purpose, first goal, only goal, own goal, hat-trick, poker, second goal, last goal, goal drew the match, goal increased/decreased the goal difference
	Substitution	Gamechanger player is subbed in, relevant player is subbed in/out
	Missed penalty	Goalkeeper saved, penalty taker missed
	Red card	Direct red card, accumulation of yellow cards
Debriefing		Post-match classification, next games, match stats
Trivia	Stats	Best/worst result of the season for the team, best/worst overall result of the season
	Streaks	Increased or broke a sequence of matches

Table 1: Available templates.

Section	Category	Variants
Title		All-purpose, 2+ goal difference, 4+ goal difference, home team won, away team won, game ended in draw, match ended with penalties
Subtitle		All-purpose, favourite team lost, decisive goal in the last minutes, no goals scored, many goals scored
Small text		Final score, starter and benched relevant players
Introduction		Final score, previous results, best player
Events	Goal	All-purpose, first goal, only goal, own goal, hat-trick, poker, second goal, last goal, goal drew the match, goal increased/decreased the goal difference
	Substitution	Gamechanger player is subbed in, relevant player is subbed in/out
	Missed penalty	Goalkeeper saved, penalty taker missed
	Red card	Direct red card, accumulation of yellow cards
Debriefing		Post-match classification, next games, match stats
Trivia	Stats	Best/worst result of the season for the team, best/worst overall result of the season
	Streaks	Increased or broke a sequence of matches

Table 1: Available templates.

Table 1: Prosebot available templates.

NLG Algorithm (materials)

The NLG algorithm starts by parsing the template, grammar, and entity manager files.

⇒ **Template** files contain the textual content that will appear in the final result, according to the validity of their conditions given the particular match's data.

⇒ **Grammar** files cater to language-specific utilities, aiding in the representation of numbers in both ordinal and cardinal forms; and also the inclusion of articles based on the gender and plurality of the subject.

⇒ **Entity** managers primarily serve to retrieve names of entities, such as players or teams, based on stored information and previously used names

NLG Algorithm (generation)

First, the system gathers match data from internal API requests.

For each match report section, text is generated recursively by selecting and applying templates.

⇒ Template selection is based on the defined per-template conditions.

⇒ Templates are applied by replacing match data or recursively expanding nested templates.

Example template – for the first game of the match in a game where there was more than one goal:

⇒ "text": "{template.time}, {scorer.name} opened the scoring for {team.name}, {template.goal_type}, {template.assisted}.",

⇒ "condition": "match_goal==1 && match_goals>1"

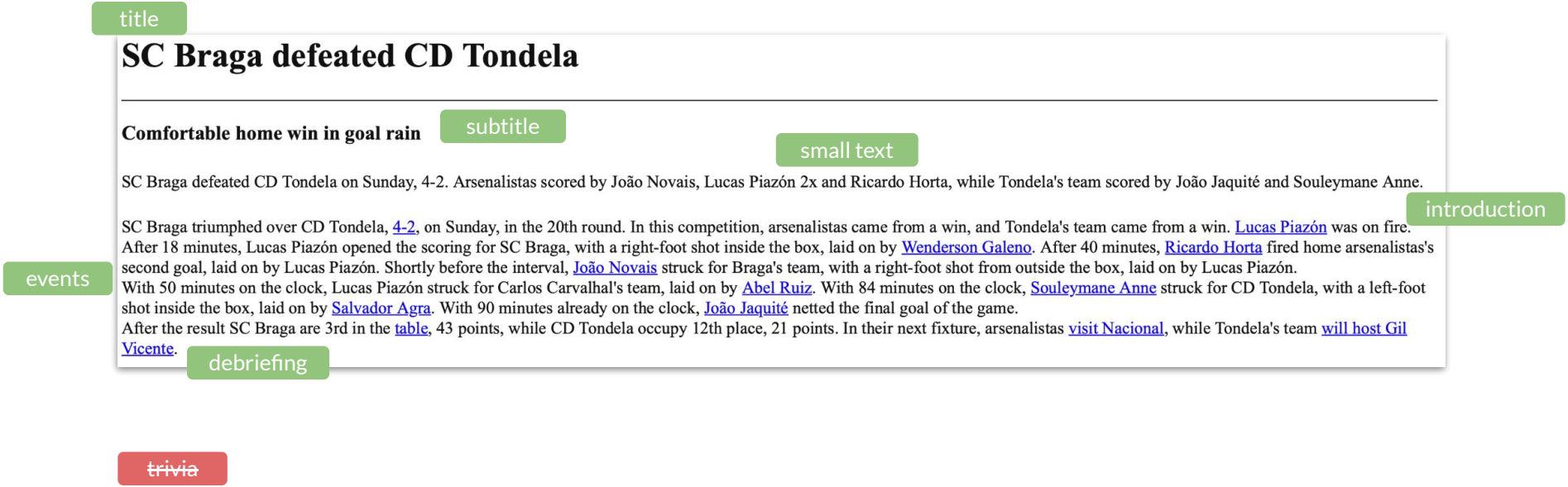
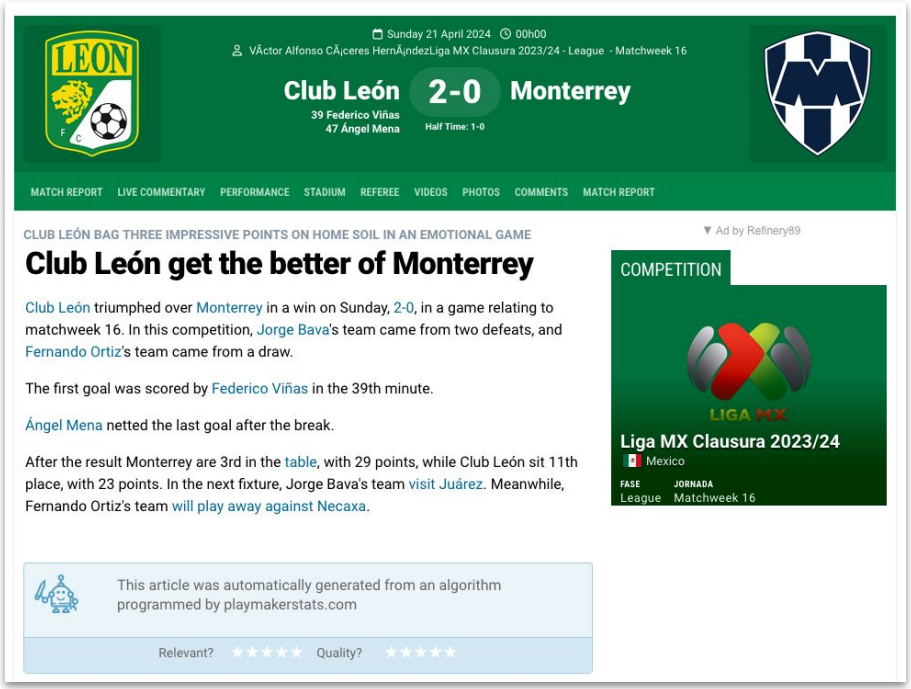


Figure 2: Sample of text generated by Prosebot in English.



Match report examples online.

Community-Based Platform (1)

zerozero.pt has always been a proponent of community engagement.

1) When a collaborator navigates to a match page, an option to craft a summary is presented.

2) Upon selection, they are directed to a dedicated interface.

3) This interface features a text editor pre-populated with content generated by Prosebot's NLG module.

For reference, the original generated text is also displayed. To guide collaborators, a brief overview of the initiative is provided, emphasizing certain guidelines. For instance, zerozero.pt advises against incorporating personal biases or subjective judgments into the summaries.

Community-Based Platform (2)

4) Upon submission of the summary, collaborators are presented with a brief survey, allowing them to share feedback on their user experience.

5) Once processed, the summary is displayed at the top of the respective match page.

In the spirit of transparency, an alert is displayed alongside the summary, crediting both Prosebot and the contributing collaborator. The platform also facilitates a collaborative review process: other community members can comment on and rate the summary based on its quality and relevance

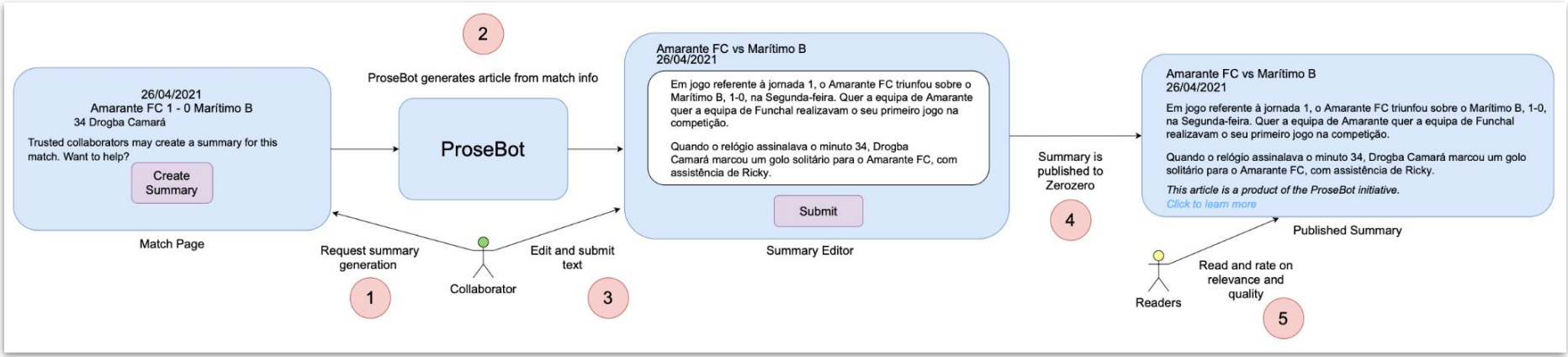


Figure 3: Overview of the Prosebot platform workflow.

Evaluation

Evaluation

To assess the efficacy and utility of the platform, we undertook a multifaceted evaluation approach.

First, we conducted an **automated comparison between the initial Prosebot drafts and the final texts** that were published on zerozero.pt.

Second, we **collected direct feedback from users through a post-submission survey**, aiming to capture their experiences and perceptions.

Third, we delved into an **analysis of the data**, examining the range of teams and competitions covered by the generated summaries and tracking the traffic these summaries attracted.

Changes to Initial Drafts

To assess the similarity between the initial drafts generated by Prosebot and the final published texts one month post-launch (174 texts), we used the Dice text similarity metric.

In a large majority of the published texts (70%), the Dice coefficient exceeds 90%, indicating a high degree of similarity between the drafts and the final texts.

Finally, we have applied a manual and automatic word frequency analysis, to understand what kind of text the collaborator usually adds or removes. Collaborators frequently added depth to the final paragraph that discusses team classifications post-match. They also highlight significant achievements, such as a team clinching the league title.

Post-Submission Survey

To gain a deeper understanding of user experiences and perceptions, we administered a survey comprising both quantitative and qualitative components.

The quantitative section consisted of seven (7) questions, each based on a 5-point Likert scale, where participants indicated their level of agreement with specific statements.

The qualitative section, on the other hand, consisted of two open- ended questions, allowing respondents to provide more detailed feedback.

We collected 46 responses, with an overall positive sentiment.

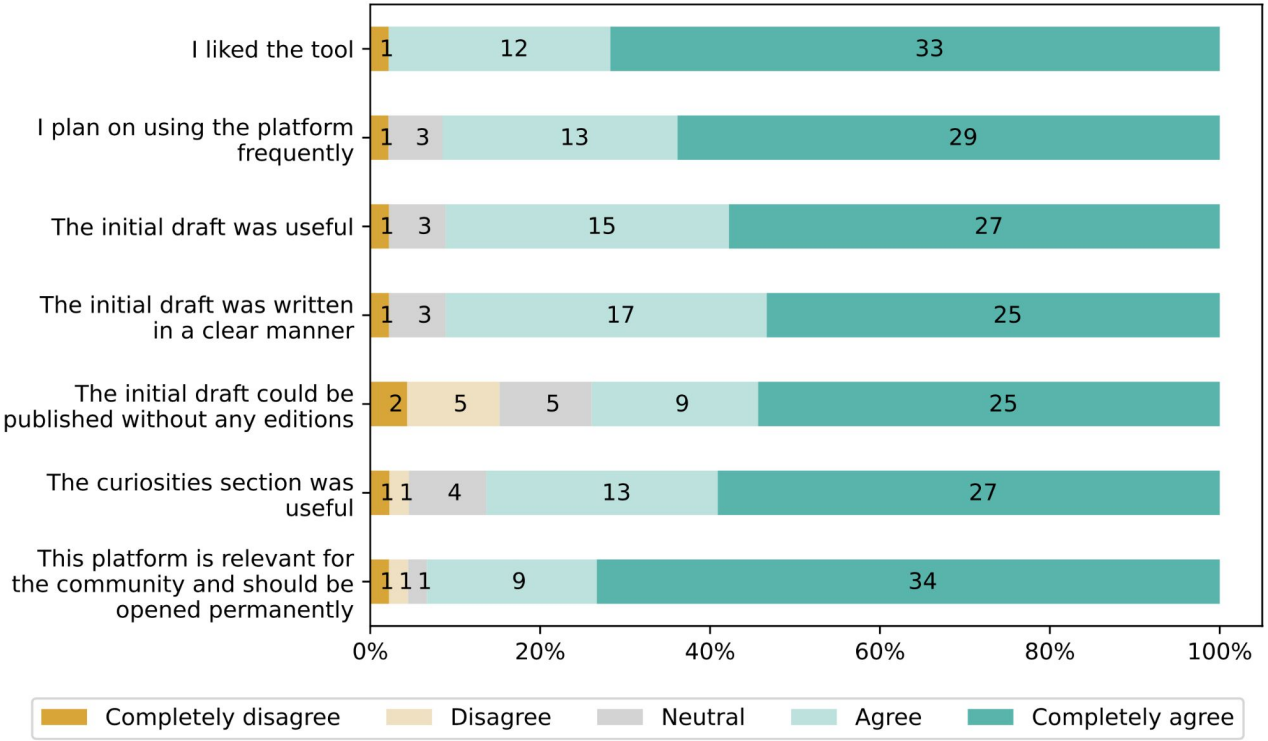


Figure 4. Results of the survey to Prosebot users.

Platform Usage and Engagement (2021)

The Prosebot platform was officially launched in early May 2021 – www.zerozero.pt/prosebot.php

⇒ A majority of authors produced between 1 to 4 reports.

⇒ The top five most active users were responsible for 40% of all reports generated.

⇒ Over half of the reports were published within two days of the match's conclusion.

⇒ A segment of users (18%) generated reports for matches from more than three months ago

Platform Usage and Engagement (2024)

In March 2024 we can observe a strong consolidation of the Prosebot platform within the community.

Over 8 thousand match summaries have been generated by more than 580 users, covering over 3 thousand distinct teams.

The top 5 users were responsible for 30% of all summaries generated and 14 users were responsible for 40%.

Indicator	May 2021	March 2024
Number of summaries	174	8,079
Number of authors	43	587
Number of teams covered	197	3,116
Number of competitions covered	56	295
Mean summaries per author	4.0	13.8
Mean summaries per team	1.8	5.1
Mean summaries per competition	3.1	27.4
Mean visits per summary	53.0	363.8

Table 2: Prosebot platform key statistics.

Discussion and Conclusions

Limitations

The Prosebot platform is used to produce match summaries derived from football match statistics.

The system is currently being used in a real-word setting with a user base of hundreds of thousands of users. Since its launch is has been used to generate thousands of texts.

One of the main limitations of a template-based system such as Prosebot is the diversity of the generated text. Although a large collection of templates have already been developed, content diversity is impacted and can be improved. To tackle this problem, automatic template extraction is being explored to augment the number of available templates.

Ethical Statement

Automatically generated media, also known as synthetic media, is increasingly common.

However, its societal impact is still under intense scrutiny and debate. In the design of the Prosebot platform, actions were taken to control the impact of such systems. Most importantly, all game summaries generated have a clear disclaimer stating that the text was generated by Prosebot.

Additionally, journalists were involved in the design of the system since early prototypes.

In the design of the Prosebot system we have found that initial fears of “machines replacing humans” have been replaced by a positive perspective of “machines helping writers”, as reflected on the survey the journalists reported in the paper.

Conclusions

In this work, we present the Prosebot platform, a pioneering effort in the Portuguese domain that integrates a community of readers into the post-editing process of automatically generated sports match summaries.

A survey conducted a month post-launch revealed a favorable reception among participants.

In the span of three years post-launch, Prosebot has been instrumental in generating millions of game summaries across various languages. This has notably amplified the coverage and visibility of matches at the grassroots level, which previously might have been overlooked.

Acknowledgements

We extend our sincere appreciation to ZOS for their ongoing and productive collaboration over the years, with special thanks to Marco Sousa and Pedro Dias.

This work is financed by national funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020 (DOI 10.54499/LA/P/0063/2020).

This presentation

A Community-Driven Data-to-Text Platform for Football Match Summaries

Pedro Fernandes[1], Luís António Santos[2], Sérgio Nunes[1]

[1] INESC TEC · FEUP, University of Porto
[2] ICS, University of Minho

INESCTEC

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Universidade do Minho
Instituto de Ciências Sociais

LREC-COLING 2024 | Turin, Italy