SocCor: A Multimodal-based Multilingual Soccer Corpus for Text Data Analytics

Paul Löhr and Jannik Strötgen
Karlsruhe University of Applied Sciences, Karlsruhe, Germany
{paul.loehr,jannik.stroetgen}@h-ka.de

Abstract

In this paper, we present SocCor, a novel dataset containing multimodal-based, multilingual media coverage of the UEFA EURO 2024. SocCor contains transcribed live commentary and video highlights as well as livetickers and reports in multiple European languages, with a total of more than 1.7 million tokens and 92,421 player mentions. By enriching SocCor with metadata about the games and curating it with normalized entity mentions based on a well-performing, tailored entity normalization approach, we provide the basis for sophisticated soccer-related text data analysis studies. Due to its multilingual nature and covering audio and text, real-time and post-game descriptions, SocCor opens ample opportunities for various types of analysis in multiple dimensions.

We demonstrate the value of SocCor by conducting and describing exploratory studies, which reveal sentiment variations, media biases, and opportunities to create player embeddings capturing soccer-specific relationships. Thus, this work provides a solid foundation for multilingual sports media research.

1 Introduction

Media coverage analyses can reveal interesting insights about events and how it is reported about them. For instance, there have been several media coverage studies on COVID-related topics (e.g., Hart et al., 2020; Mach et al., 2021). A sports domain of great interest to people is soccer. While there have been a few linguistic-centric studies in the past about this sports domain (e.g., Meier-Vieracker, 2017), there is only little work on soccer-related news media coverage analysis. One reason for this is probably the lack of well-curated datasets that allow for sophisticated studies.

In this paper, we present SocCor, a novel soccer news dataset containing multilingual media coverage of the UEFA EURO 2024. A key feature of SocCor is that it is multimodal-based: it consists of transcribed live commentary and video highlights as well as text data in the form of livetickers and reports. SocCor covers seven European languages, totaling to more than 1.7 million tokens with data from 20 sources, e.g., Kicker (Germany), BBC (England), L'Equipe (France), and Marca (Spain).

By enriching SocCor with metadata about the games, we increase the value of the data set for sophisticated multilingual sports media research. In addition, we increase SocCor's quality by tackling spelling mistakes from the original data and in particular mistakes from the transcription process, by developing a well-performing, tailored entity normalization approach to normalize entity mentions referring to players and coaches.

Due to its multilingual nature and covering audio and text, real-time and post-game descriptions, SocCor opens ample opportunities for various types of analysis across multiple dimensions. For instance, studies could be cross-language / cross-culture (e.g., are there differences regarding the sentiment of the reporting across languages and countries and does it depend on which teams are playing?) or cross-modality (e.g., is there a difference in live commenting in video vs. live commenting in "ticker"-style reports? Are ticker reports more descriptive to compensate for the missing image data?). By conducting and describing exploratory sample studies, we demonstrate the broad opportunities of SocCor-based data analysis, revealing, for instance, sentiment variations across media types and languages, country-specific media biases, and opportunities to create player embeddings capturing soccer-specific relationships.

In summary, this work provides a solid foundation for future research on diverse multilingual sports media.[1]

[1] SocCor and further resources related to this paper can be found at https://github.com/PaulLoehr/SocCor.

2 Related Work

Soccer datasets can be of different nature: text-based, statistical, and about tracking information. We develop a dataset for natural language processing (NLP) and text data analytics and thus focus on text-based data. This type of data conveys the emotional, linguistic, and cultural facets of soccer discourse. However, statistical data, when used as metadata (e.g., lineups, possession, results, and goals), are valuable for providing contextual match information that enriches text-based analyses.

This section reviews existing datasets, emphasizing the limitations of text-based resources and the absence of integrated statistical metadata, to highlight the gaps addressed by our work.

Text-Based Datasets Text-based datasets about soccer are rare. However, they are critical for NLP and text data analytics, enabling the analysis of language use, sentiment, and cultural dimensions. One dataset is the soccer linguistics corpus by Meier-Vieracker (2017), a multilingual dataset designed for studying soccer journalism. Covering 13 languages (e.g., English, Spanish, German, Portuguese) and with approximately 75 million words, it contains reports, livetickers, and live commentary transcripts. Reports offer detailed game analyses, covering team performances and tactical aspects. Livetickers provide real-time updates with spontaneous reactions to events like goals or fouls, while live commentary transcripts from radio broadcasts combine objective descriptions with subjective assessments, enhancing emotional engagement.

Despite its strengths, the soccer linguistics corpus has significant limitations. It does not focus on a specific tournament, instead aggregating texts across various matches and seasons. This lack of tournament-specific coverage prevents comprehensive analysis of all reports, livetickers, and commentary for individual games within a single tournament, limiting its utility for tasks like comparative sentiment analysis or event detection across multiple sources for the same match. Crucially, it does not include statistical metadata, such as lineups, possession statistics, or match results, which are essential for contextualizing textual analyses with general match information. With SocCor, we focus on one tournament (the UEFA Euro 2024).

Statistical and Tracking Data While text-based data are the focus of this work, statistical and tracking datasets provide complementary numerical insights that are particularly relevant as metadata. Statistical datasets, such as those on Kaggle by Sekhri (2024) and Dizdarevic (2024) for UEFA Euro 2024, offer match and player statistics, including goals, assists, and team performance metrics. FBref (2024)'s UEFA Euro 2024 dataset provides comprehensive data on results and player contributions. Tracking data from StatsBomb (2024), including event and tactical data for UEFA Euro 2024, enable detailed analyses of player movements and strategies. These datasets are valuable for objective analysis and can serve as metadata to contextualize text-based data, providing details like lineups, possession percentages, or substitutions that enhance NLP tasks, such as linking sentiment to specific game events.

However, existing statistical and tracking datasets are typically standalone and not integrated with text-based corpora. To bridge this gap, we release SocCor and provide a unified resource, which allows a variety of analytics tasks to be performed on a single resource covering linguistic aspects as well as metadata.

3 SocCor

The main contribution of this work is SocCor, a novel dataset designed for text data analytics with a focus on the UEFA Euro 2024 tournament. SocCor comprises multilingual text and transcribed audio data from diverse sources, including livetickers, match reports, video highlights, and complete game commentaries, complemented by statistical metadata (e.g., lineups, results) and automatically annotated player mentions.

The sources differ in various aspects: real-time (livetickers, complete game commentaries) vs. post-hoc (match reports, video highlights), as well as text-based (livetickers, match reports) vs. speech-based (complete game commentaries, video highlights) data. Furthermore, data are collected from sources in seven major European countries (Germany, France, Spain, England, Italy, the Netherlands, and Turkey) and thus multilingual (see Figure 1 for an overview of the distribution by language and media type). This heterogeneity of the data provides a comprehensive representation of soccer discourse, capturing linguistic, emotional, and cultural nuances, which allows for a huge diversity of analysis tasks.

In the following, we describe the collection of the data (Section 3.1), the preprocessing (Section 3.2), the automatic annotation processes (Section 3.3), and dataset statistics (Section 3.4).

3.1 Data Collection

SocCor aggregates multimodal data from media coverage of the UEFA Euro 2024 for all 51 matches. The dataset includes four primary data types and statistical metadata, collected to capture diverse textual and audio narratives of the tournament and its individual matches. To address the heterogeneity of sources, tailored Python scripts were developed for automated data acquisition, ensuring consistent and structured storage across text, audio, and metadata.

Livetickers Real-time text updates with minute-by-minute event descriptions and spontaneous reactions to goals, fouls, and other match events. Stored as `.json` files with event timestamps.

Reports Detailed journalistic analyses, offering summaries of team performances, key moments, and tactical insights, reflecting diverse linguistic and cultural perspectives. Stored as `.json` files.

Video Highlights Extracted audio tracks, capturing live commentary and crowd reactions. Initially stored as .mp3 files, with transcriptions to be stored as `.json` files, consistent with the liveticker format.

Game Commentaries Audio recordings of full matches, representing the most restricted data due to limited availability. Initially stored as .mp3 files, with transcriptions to be stored as `.json` files, mirroring the liveticker structure.

Metadata Match details (location, date, teams, lineups, results and goals) to contextualize textual and audio data. Stored as `.csv` files.

Data collection targeted comprehensive coverage from major soccer nations, leveraging their robust media ecosystems to ensure multiple perspectives on each match. Python scripts systematically scraped and processed data, transforming heterogeneous inputs into a unified format. All data are organized by unique match identifiers, enabling efficient retrieval for NLP and analytics tasks.

3.2 Preprocessing

Preprocessing SocCor's audio data is essential to convert spoken commentary into a text format suitable for analysis. Here, we describe the transcription of audio files from video highlights and complete game commentaries, ensuring consistency with the `.json` storage format for text data.

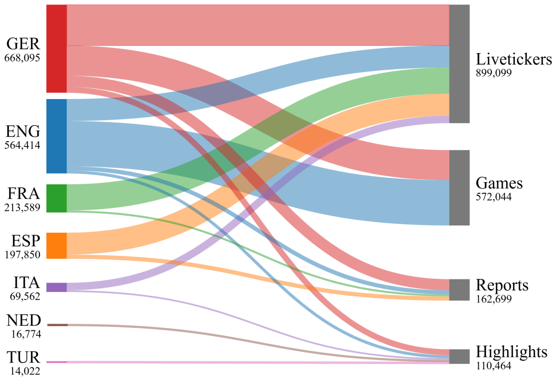

Figure 1: Sankey diagram of the SocCor dataset, showing the distribution of token counts across languages and media types.

Transcription of Audio Files Due to its robustness across multiple languages, we use the Whisper automatic speech recognition (ASR) model (Radford et al., 2023). However, technical limitations of Whisper require that the audio is segmented prior to transcription. To preserve contextual coherence, a pause-based segmentation strategy is employed, leveraging natural breaks in the commentary.

Audio files are initially stored as .mp3, with transcriptions saved as `.json` files, consistent with liveticker data, including timestamps and match IDs (e.g., "51_ENG_ESP") for event alignment.

However, the transcription of such sports commentary audio files faces several challenges:

- **Background Noise**: Stadium sounds and crowd noise, varying by broadcaster, reduce transcription accuracy, especially in high-volume segments.

- **Artifacts**: Unclear audio segments produce repeated (partially strange) text passages, necessitating post-processing to remove redundant (and incorrect) content – as detailed below.

- **Pronunciation Variations**: Inconsistent transcriptions of player names (e.g., "Füllkrug" in German vs. various misspellings in other languages) due to language-specific phonetics complicate entity normalization.

To address artifacts, a frequency-based detection system filters unwanted elements, including recognition failures (e.g., repetitive text due to noise), model hallucinations (e.g., "Das war ein guter Treffer" not present in audio), and broadcaster elements (e.g., jingles mistranscribed as "Subtitles by ZDF

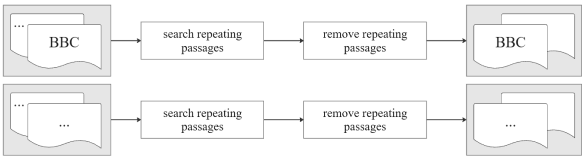
...
BBC
search repeating passages
remove repeating passages
BBC
...
...
search repeating passages
remove repeating passages
...

Figure 2: Artifact identification and removal workflow for systematic transcription refinement.

for funk, 2017”). Broadcaster-specific artifact inventories are compiled, capturing recurring phrases’ frequency distributions. Figure 2 shows the workflow in which a filtering algorithm refines the segmented transcriptions into clean `.json` outputs.

A critical example from *Sportschau*, a German broadcaster, involves the phrase “Sieg Heil! Sieg Heil! Sieg Heil!” appearing six times across two match transcriptions. This error arose due to the Whisper model misinterpreting loud, euphoric fan chants in the stadium as spoken words during commentary. Manual audio verification confirms that no Nazi salutations were sung in the stadium; the chants were enthusiastic but indistinct crowd noise. This demonstrates the need for careful verification to avoid misrepresentation in noisy environments.

Pronunciation variations and entity normalization challenges are addressed in subsequent processing steps, ensuring high-quality data for downstream NLP and analytics tasks.

3.3 Automatic Annotation

The main goal of the automatic annotation of the data in SocCor is to enrich the corpus with metadata about entity mentions. By normalizing player and coach names across multilingual text data from transcribed audio, livetickers, and match reports, the automatic annotation of SocCor enhances its utility for analytics tasks.

This subsection details the use of spaCy models for named entity recognition and our normalization approach against match-specific player lineups to resolve entity ambiguities, addressing challenges like pronunciation variations, spelling errors, and transcription errors.

Named Entity Recognition To extract named entities, we employ spaCy’s pre-trained large NER models (Honnibal and Montani, 2023) for each of SocCor’s language (English, German, Spanish, French, Italian, Dutch, and Turkish). These models, trained on general-domain corpora, effectively recognize PERSON entities in the sports media context of SocCor, offering consistency and computational efficiency.

	# entities	**prec**	**rec**	**f-score**
German	528	0.89	0.79	0.84
English	852	0.92	0.74	0.82
Italian	99	0.87	0.59	0.70

Table 1: NER performance on a subset of SocCor for three languages using spaCy’s large NER models. 2 English and 2 German full match transcripts, and 2 highlight summaries for all three languages, respectively.

For each input document $d \in D$ (e.g., a transcribed audio segment, liveticker entry, or match report), the pipeline extracts entities, retaining only those classified as PERSON, as only these are relevant for player and coach identification. This results in a set of entities $E_d = \{e_1, \ldots, e_n\}$.

To evaluate NER performance, a subset of SocCor was manually annotated using INCEpTION (Klie et al., 2018), creating a ground truth for PERSON entities. Despite differences in ASR transcription quality across broadcasters and languages, the evaluation results shown in Table 1 demonstrate robust performance across languages.

Pattern Matching with Match Lineups To normalize PERSON entities as specific players or coaches, extracted names are matched against match-specific lists from the dataset’s metadata. For each match, identified by unique IDs, the system loads the lineups of the two competing teams. This context-aware approach results in a high-quality corpus with player and coach names relevant to each match.

The system employs two complementary matching approaches to handle orthographic and phonetic variations in extracted entity mentions.

1. **Levenshtein-based Fuzzy Matching**: This method quantifies orthographic similarity using Levenshtein distance (Navarro, 2001), accounting for minor variations such as accents, spacing, or misspellings (e.g., *Fullkrug* vs. *Füllkrug*).

2. **Soundex-based Matching**: To address transcription errors, names are encoded using the Soundex algorithm (Zobel and Dart, 1996), generating 4-character codes that represent phonological patterns. These codes are compared using Levenshtein distance to identify matches despite incorrect transcriptions (e.g., *Tchouaméni* as *Schuamanni*).

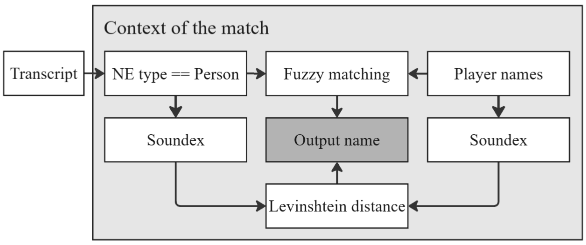

Figure 3: Our approach to perform player name normalization in multilingual transcriptions.

Figure 3 depicts our name correction pipeline to normalize entity mentions in multilingual transcriptions. The system evaluates three outcomes for each PERSON entity: retaining the original name, replacing it with the best Levenshtein fuzzy match, or replacing it with the best Soundex-based match. Levenshtein distance quantifies orthographic similarity (δ_f), while Soundex code comparisons (using Levenshtein distance) identify phonological correlations (δ_s), each producing a similarity coefficient as a confidence metric. If either method yields a high-confidence match (i.e., above a predefined threshold), the entity is corrected to the corresponding player name from the lineup, linking to the metadata. If both methods agree on the same match, the confidence is reinforced. Entities with insufficient confidence (i.e., low similarity scores) are kept unchanged, preserving potential non-player references (i.e., other persons). The output name is chosen based on similarity thresholds and the agreement between methods formalized as follows:

$$\text{name} = \begin{cases} n, & \text{if } \delta_f = 100\% \\ m_f, & \text{if } m_f = m_s \\ m_f, & \text{if } \delta_f \geq 80\% \wedge \delta_f \geq \delta_s \\ m_s, & \text{if } \delta_s \geq 85\% \wedge \delta_s > \delta_f \\ n, & \text{otherwise} \end{cases} \quad (1)$$

where m_f, m_s are best matches from fuzzy and Soundex methods, respectively, δ_f, δ_s are similarity scores (0–100%) for fuzzy and Soundex methods, respectively, and n is the original input name. Some sample corrections are shown in Table 2.

ASR output	Harbertz	Schumanni
Best Fuzzy	Havertz	Muani
Best Soundex	Havertz	Tchouaméni
Combined	Havertz	Tchouaméni
Ground Truth	Havertz	Tchouaméni

Table 2: Sample corrections of ASR-induced errors: orthographic and morphological variations.

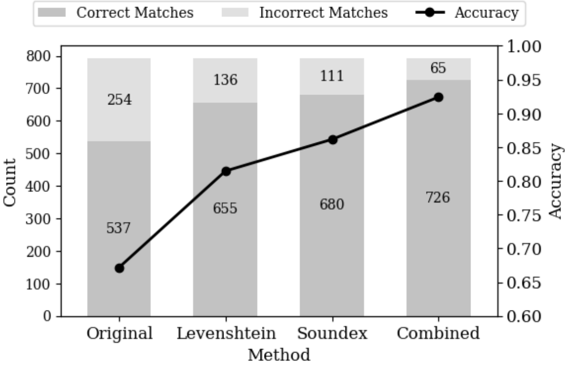

Figure 4: Comparative player name correction performance across methodologies.

To evaluate our entity normalization approach, 10 transcription documents in German, English, and Italian are manually verified. Our approach achieves a 92% normalization accuracy (726/791 entities resolved), a 24 percentage point improvement over the 68% baseline ASR accuracy (539/791), correcting 74.4% of errors (189/254 discrepancies), as shown in Figure 4.

Player Token Generation To normalize player and coach mentions across multilingual text data in SocCor, we implement a standardized tokenization approach that transforms names into consistent structured tokens. Full names, official roles (e.g., playing position or coach), and team identifiers are extracted from the metadata, filtered by match-specific identifiers derived from filename conventions. A rule-based text substitution pipeline applies Unicode normalization to convert text to ASCII, removing diacritical marks, followed by lexical pattern matching to detect complete names (e.g., *Harry Kane*) or surnames (e.g., *Kane*). Detected mentions are replaced with tokens in the format `<NAME_POSITION_TEAM>` (e.g., `<KANE_CENTRE-FORWARD_ENG>`), preserving positional context and team affiliation. This process addresses cross-lingual name variations, ensures consistent player references, and supports downstream NLP and analytics tasks.

3.4 SocCor Statistics

The SocCor dataset, sourced from UEFA Euro 2024 media coverage, consists of 1,180 files with a total storage size of 20.3 MB. It contains livetickers, game commentaries, written re-

Media	Initial Modality	Tokens
Livetickers	text	899,099
Games	audio	572,044
Reports	text	162,699
Highlights	audio	110,464
Sum		1,744.306

Table 3: Distribution of tokens across source types of different modalities in the SocCor dataset.

Media	Tokens	Player	Ratio [%]
Games	7,221.5	226	3.13
Highlights	332	23	6.93
Livetickers	1,661	112	6.74
Reports	586	26	4.43

Table 4: Median token and player mention counts, and player-to-token ratios across media types, showing denser player references in highlights and livetickers.

ports, and video highlights, all processed into text. Text-based sources (livetickers and reports) are scraped, with player names replaced by standardized tokens (e.g., *Kylian Mbappé* as `<MBAPPE_CENTRE-FORWARD_FRA>`). Audio sources (video highlights and game commentary) are transcribed using Whisper, with various player-related transcription errors corrected via the above-described named entity recognition and name normalization. The distribution of tokens across the four data sources in SocCor is summarized in Table 3.

On average, each match contains 34,202 tokens. Token counts range from approximately 19,548 (Slovenia vs. Serbia) to 58,853 (England vs. Switzerland, quarterfinal with extra time and penalties). In addition to the varying number of sources per match, this range indicates that token counts are also influenced by match duration and notable in-game events.

For all 51 matches of the UEFA Euro 2024, data is available across media types to ensure comprehensive coverage. Some game commentary files are missing either from *Sportschau* (German) or *BBC* (English), as these broadcasters did not provide data for all matches. *AS English* has incomplete livetickers due to scraping challenges with dynamic website structures. Written reports are also missing for some games due to minor scraping inconsistencies. However, a multi-source approach mitigates these gaps by leveraging complete data from alternative broadcasters' livetickers, game commentaries, and highlights, ensuring that every match is represented with adequate data for analysis.[2]

4 Sample Analytics Tasks on SocCor

To demonstrate the value of our novel SocCor dataset, we now present a series of text data analytics tasks conducted on SocCor. We apply NLP techniques to analyze player mentions, sentiment distributions, and player embeddings, addressing the research goal of understanding cross-modality and language-specific differences in soccer media narratives. This analysis reveals several sample findings: patterns in player prominence, sentiment variations across media types, and semantic relationships encoded in domain-specific embeddings. Thus, first insights into the representation of players and teams in the UEFA Euro 2024 media landscape can be gained.

4.1 Modality-based Player Mentions Analysis

Among the 92,421 player mentions in SocCor, *Kylian Mbappé* leads with 1,408 mentions, followed by *Harry Kane* (1,177), *Cristiano Ronaldo* (1,075), *Jude Bellingham* (1,020), and *Phil Foden* (1,003). This distribution reflects the media's emphasis on high-profile players during UEFA Euro 2024, with stars from prominent teams like France, England, and Spain dominating coverage due to their on-field impact and public recognition.

Variation Across Media Types Player mentions vary significantly across media types, influenced by the format and purpose of each modality. Table 4 presents the median word counts, player mention counts, and player-to-token ratios for each media type. Game commentaries exhibit the highest median player mentions (226), driven by comprehensive match coverage, but their large word count (7,221.5) results in the lowest player-to-token ratio (3.13%). Livetickers, with a median of 112 mentions and a ratio of 6.74%, and highlights, with 23 mentions and a ratio of 6.93%, show denser player references, reflecting their focus on key events. Reports, with 26 mentions and a ratio of 4.43%, balance player focus with tactical analysis, indicating modality-specific reporting styles.

Match-specific Mention Disparities Analysis of specific matches highlights disparities in player

[2] Further details about sources will be provided on the SocCor website.

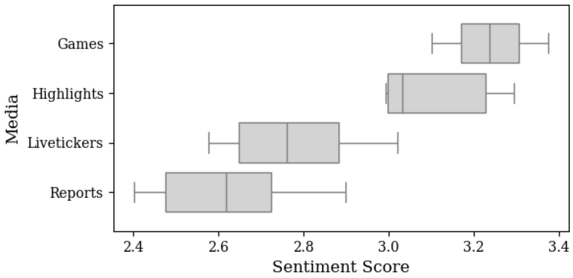

Figure 5: Boxplot of sentiment score distributions across media types in SocCor.

mentions, reflecting match outcomes and media focus. In the opening match between Germany and Scotland, which Germany won 5-1, the corpus records 1,333 mentions of German players compared to 566 for Scottish players across all broadcasters. This imbalance aligns with Germany's dominance, suggesting that media attention amplifies the winning team's players.

In the final match between Spain and England, where Spain secured a 2-1 victory, mentions are more balanced, with 1,145 for English players and 1,047 for Spanish players. However, broadcaster-specific patterns emerge: the BBC reports 299 English player mentions versus 208 Spanish in game coverage and 29 versus 19 in post-match reports, indicating a preference for the home team, while livetickers show near parity. These findings underscore how match context and broadcaster perspective shape player mention patterns across modalities.

4.2 Sentiment-based Analysis

We also split SocCor into individual sentences to perform a sentence-based sentiment analysis. Each sentence is processed using a pre-trained multilingual BERT-based (Devlin et al., 2019) sentiment model, which assigns a sentiment score on a scale from 1 (negative) to 5 (positive) (Hugging Face, 2020). When a sentence contains a player token, such as <KANE_CENTRE-FORWARD_ENG>, the sentiment score is assigned to multiple entities: the player, his team, the game, the broadcasting organization, and the media type (e.g., game commentaries, video highlights, livetickers and reports). This multi-level assignment enables detailed sentiment analysis across dataset dimensions.

Two sample sentences mentioning *Harry Kane* demonstrate the model's ability to capture tonal differences: "Brilliant work by Kane." (positive sentiment, 5) and "And Bjorn's not letting Kane go anywhere." (negative sentiment, 1). These corpus-derived examples highlight the model's sensitivity to sentiment variations in soccer media.

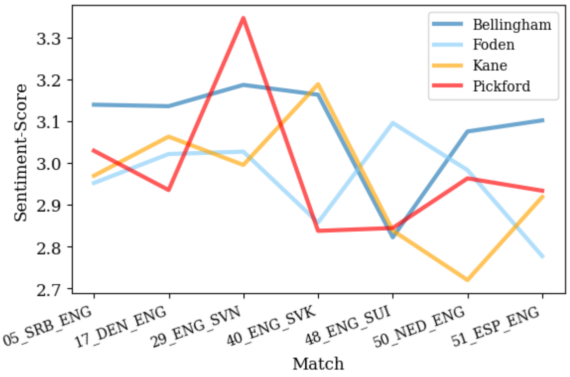

Figure 6: Mean sentiment scores of four players, reflecting media perceptions of performance per match.

Media Type Sentiment Distribution Mean sentiment scores across media types, visualized in Figure 5, reveal tonal differences. Game commentaries exhibit the highest mean score at 3.2, followed by highlights at 3.1, suggesting more positive coverage in these formats. Livetickers have a mean score of 2.7, while reports are lowest at 2.5, indicating more neutral or critical tones, likely due to their focus on continuous real-time updates (e.g., without slow motions of highlights in contrast to game commentaries) or rather objective post-game analysis. The boxplot illustrates the spread and variability of sentiment scores, highlighting modality-specific narrative styles in the UEFA Euro 2024 coverage.

Temporal Per-Game Sentiment Analysis Mean sentiment scores for selected English players are aggregated per game and analyzed across the tournament's matches, ordered chronologically, as visualized in Figure 6. This analysis reveals variations in media portrayal, reflecting perceived performance in individual matches.

For instance, goalkeeper *Jordan Pickford* achieves his highest sentiment score in the Slovenia match, a 0-0 draw, likely due to consistent positive coverage of his defensive actions that contributed to a clean sheet. In contrast, *Harry Kane* records his lowest sentiment score in the semi-final against the Netherlands, where, despite scoring a penalty, critical commentary on missed opportunities during regular play time lowers his aggregated score.

These per-game and temporal trends, supported by the visualization, highlight how match-specific events shape sentiment-based player ratings, of-

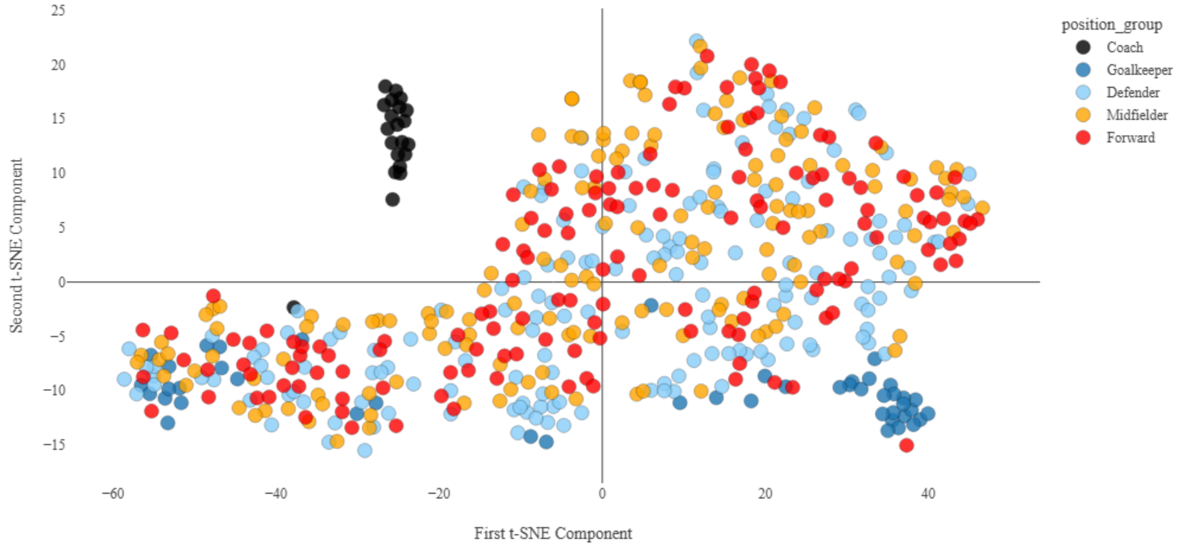

Figure 7: t-SNE visualization of German player embeddings in SocCor, colored by position group, showing clustering based on textual references.

fering a nuanced view of performance across the tournament.

4.3 SocCor-based Player Embeddings

To further demonstrate the value of SocCor, we generate static word embeddings for the German corpus of SocCor (650,000 tokens) using the Word2Vec CBOW model (Mikolov et al., 2013). Interesting future work opportunities could be, for instance, to create word embeddings for all languages and perform a detailed player analysis.

CBOW is chosen for its efficiency in capturing stable representations of domain-specific soccer terminology on a small corpus. Training uses a minimum token frequency of 5, filtering infrequent terms to ensure reliable vector representations while retaining relevant vocabulary. The resulting 300-dimensional embeddings encode semantic and syntactic relationships among players and soccer concepts, enabling downstream tasks like similarity analysis and visualizations.

t-SNE Visualization To explore the spatial structure of player embeddings, we project the 300-dimensional vectors into a two-dimensional space using t-SNE (van der Maaten and Hinton, 2008). Figure 7 illustrates this visualization, with players colored by position as one sample dimension.

Coaches form a distinct cluster (top left), reflecting their unique textual references, while goalkeepers cluster separately (lower-right), indicating distinct roles. Outfield players (defenders, midfielders, forwards) show overlapping distributions, suggesting textual similarity across these positions. Notably, the lower-left region includes players with limited playing time, hinting that the horizontal axis may capture match involvement. While t-SNE reduces complexity and may distort high-dimensional relationships, it reveals positional patterns and role distinctions.

Again, we expect that future work could perform much more detailed analyses on SocCor across various dimensions, e.g., to identify most similar players across countries.

Kroos Case Study A case study of *Toni Kroos*, frequently nicknamed in a negative manner as *Querpass Toni* for his horizontal passing, exemplifies the embeddings' insights. *Kroos*' vector (`<KROOS_MIDFIELDER_GER>`) has a cosine similarity of 0.86 to *querpass* (sideways pass), ranking him 234th among players for this term. Centerbacks *Faes* (0.96), *Vestergaard* (0.96), and *Christensen* (0.95) show much stronger associations with *querpass*, reflecting defensive passing patterns.

Kroos' embedding aligns more closely with technical passing terms: *steckpass* (one-touch pass, 0.92), *steilpass* (through ball, 0.92), *verlagern* (lateral distribution, 0.92), and *doppelpass* (give-and-go, 0.89). This suggests a versatile passing profile, contrasting with past media narratives emphasizing horizontal passes. The discrepancy highlights the embeddings' ability to uncover nuanced player roles, potentially limited by corpus coverage of stylistic nuances, and underscores their value for analyzing soccer media discourse.

Obviously, SocCor-based embeddings offer ample opportunities to perform person-centric analyses highlighting its value for future research.

5 Conclusion

In this paper, we presented SocCor, a novel dataset containing multimodal-based, multilingual media coverage of the UEFA EURO 2024. With its multilingual content covering audio and text as well as real-time and post-game descriptions, SocCor allows for various types of analysis across multiple dimensions, e.g., cross-language, cross-modality, and time-centric analyses.

By enriching SocCor with normalized player and coach mentions using a tailored, well-performing entity normalization approach, we provide the basis for future sophisticated text data analysis tasks. We exemplified this opportunity by performing and describing several sample analyses such as a player-centric sentiment analysis and SocCor-based player embeddings which can reveal interesting insights as we demonstrated about a sample player.

Limitations

Despite SocCor being a multilingual dataset, it covers only seven languages – and some of them do not cover all media types. Obviously, it would be nice if further languages and more languages across all media types were covered in the dataset. In addition, the data is not equally distributed across all covered languages so that language comparisons should probably be made with subsets of the dataset only. Despite covering multimodal-based sources, SocCor contains only a limited number of different news outlets and broadcasters due to difficulties in accessing additional sources.

Regarding the analysis performed on SocCor, we only present several sample analysis results and many further analyses could be performed. However, we want to emphasize that our sample analyses shall mainly demonstrate the value of SocCor besides providing interesting insights.

Finally, for generating word embeddings more sophisticated methods could be used, and embeddings could be generated on more than one language. However, again, we mainly aim at demonstrating opportunities to perform a variety of analysis tasks on SocCor, and leave more sophisticated approaches for future work.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.

Damir Dizdarevic. 2024. UEFA EURO 2024 - players. https://www.kaggle.com/datasets/damirdizdarevic/uefa-euro-2024-players.

FBref. 2024. UEFA Euro 2024 stats. https://fbref.com/en/comps/676/UEFA-Euro-Stats.

P. Sol Hart, Sedona Chinn, and Stuart Soroka. 2020. Politicization and polarization in covid-19 news coverage. *Science Communication*, 42(5):679–697.

Matthew Honnibal and Ines Montani. 2023. spaCy: Industrial-strength Natural Language Processing in Python. Version 3.7.2.

Hugging Face. 2020. nlptown/bert-base-multilingual-uncased-sentiment. Accessed: 2025-05-06.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics.

Katharine J. Mach, Raúl Salas Reyes, Brian Pentz, Jennifer Taylor, Clarissa A. Costa, Sandip G. Cruz, Kerronia E. Thomas, James C. Arnott, Rosalind Donald, Kripa Jagannathan, Christine J. Kirchhoff, Laura C. Rosella, and Nicole Klenk. 2021. News media coverage of covid-19 public health and policy information. *Humanities and Social Sciences Communications*, 8(1).

Simon Meier-Vieracker. 2017. Korpora zur Fußballlinguistik – eine mehrsprachige Forschungsressource zur Sprache der Fußballberichterstattung. In *Zeitschrift für germanistische Linguistik*, volume 45, pages 374–381.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM Computing Surveys (CSUR)*, 33(1):31–88.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, pages 28492 – 28518. JMLR.org.

Thamer Sekhri. 2024. Euro 2024 matches stats. https://www.kaggle.com/datasets/thamersekhri/euro-2024-matches.

StatsBomb. 2024. StatsBomb release free Euro 2024 data. https://statsbomb.com/news/statsbomb-release-free-euro-2024-data/.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

Justin Zobel and Philip Dart. 1996. Phonetic string matching: lessons from information retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, page 166–172, New York, NY, USA. Association for Computing Machinery.