# Soccer Commentary Mining
# Project Final Report for NLP Course, Winter 2023

**Szymon Maksymiuk**
Warsaw University of Technology
01131304@pw.edu.pl

**Adam Narożniak**
Warsaw University of Technology
01133060@pw.edu.pl

Władysław Olejnik
Warsaw University of Technology
01130735@pw.edu.pl

Patryk Świątek
Warsaw University of Technology
01151517@pw.edu.pl

Karolina Seweryn
NASK - National Research Institute
sewerynnkarolinaa@gmail.com

supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

## Abstract

Sports popularity is constantly growing throughout the world. Hence, it is no surprise that the most popular sport, which is soccer, with 3.5 billion estimated fans, brings the attention of scientists from various fields. In this work, we aim to analyze soccer matches from broadcasters' perspective to check how events on the field affect their emotions and, inherently, viewers' emotions as entertainment of the commentary is critical to good transmission. We will perform unsupervised sentiment analysis on transcriptions of the soccer match broadcasts. For that purpose, we will use established methods like lexicon-based VADER, which has already been proven successful in analyzing soccer data, and pre-trained state-of-the-art model *flair*. Obtained results will be later statistically compared with the on-field events like goals, cards, free kicks, etc., to asses whether occurrence or time between consecutive events significantly affects broadcaster sentiment. We will also use state-of-the-art models for deriving emotions directly from the recording and compare them with results acquired using transcriptions. The analysis will be based on the public SoccerNet databases consisting of hundreds of videos of soccer matches and collect transcriptions using the Whipser OpenAI tool. Overall, the work presents a novel approach to sentiment analysis in soccer by attempting analysis of data that has never been addressed before, which are broadcast's transcription, and comparing them with on-field events and audio characteristics.

## 1 Introduction

Sports is one of the most prominent people's entertainment worldwide, mainly due to the variety of channels through which we can experience multiple sports events. Among these, the most popular one is undoubtedly soccer, with an estimated 3.5 billion fans worldwide. That figure shows how important soccer is in today's global culture. Such a fact did not slip through the radars of scientists worldwide, who approach soccer from multiple directions, starting with physiotherapy, through pharmaceutics, and ending with statistics and artificial intelligence.

One of the most significant areas in the soccer industry is live broadcast, where experts comment on live events on the field. These experts' duties are not only to describe events happening during the game but also to bring valuable insights and, most important of all, to keep the audience entertained. The ability to engage people is naturally embedded deeply in the commentator's job description, and they are trained to voice their opinions in a matter accessible to the audience. Still, even these professionals can become overwhelmed by the events on the field. That's

precisely the place where the motivation for this project comes from. We aim to assess whether and by how much events on the field affect what emotions resonate from broadcasters' words. For that purpose, we will perform a sentiment analysis of the transcripted soccer match commentaries and statically compare the obtained results with the state of the matches at a given moment.

## 2 Related work

Nowadays, with the progressing globalization and constantly growing usage of various types of Internet media, we are nearly overwhelmed with data. People voice their opinions about almost everything through X (former Twitter) posts, service reviews, and journalistic articles. As underlined by the (Hamborg and Donnay, 2021) the fact that these forms encourage explicit opinion form of opinion, they were a perfect source of data for researchers exploring Sentiment Analysis in these areas for many years now (Pontiki et al., 2015), (Nakov et al., 2016), (Yin et al., 2020), (Sun et al., 2019), (Baccianella et al., 2010). Recently, (Ramnarine, 2023) used unsupervised sentiment analysis to derive whether social media posts about plastic surgeries are positive, neutral, or negative. Although the article has not yet been published, it certainly shows diverse appliances of sentiment analysis that span even medical care. As seen, the subject of Sentiment Analysis exploration is deeply embedded in the scientific community, so there is little surprise that there have been. Considering that, it is not a surprise that there were attempts to implement the same methods in the soccer world.

(Jai-Andaloussi et al., 2015) in their work used Twitter data to analyze sentiment associated with users interacting with each other about the games to predict who they are rooting for and possibly events on the field. Such analysis allowed authors to summarize soccer events efficiently. Similar work was done by (Patel and Passi, 2020), where a team of researchers analyzed historical data about the FIFA World Cup 2014 held in Brazil. The authors used sentiment analysis to detect emotions associated with particular messages covering the matches. It is worth pointing out that, among others, they used lexicon-based methods, which brought pretty successful results. While exploring the past research that has been done in sports sentiment analysis, we could not find anyone trying to analyze broadcasters' sentiments. Hence, we believe our approach to be novel.

Considering the fact that the dataset with the transcriptions is not available, it needs to be gathered from the audio data. The common approach is the open-source Whisper model (A. Radford and Sutskever, 2023). It is a deep learning-based automatic speech recognition (ASR) system developed by OpenAI, designed to convert spoken language into written text. It was trained on the 680,000 hours of multilingual and multitask supervised data collected from the web, which has increased its robustness compared to the previous similar models trained on significantly smaller labelled data or unsupervised learning models. Approximately one-third of the audio dataset is non-English and alternately tasked with transcription in the original language or translation into English. Therefore, the Whisper model can handle a wide range of tasks, including transcription for general speech, specific domain applications, and multilingual scenarios. The core architecture of the Whisper model is a simple end-to-end approach based on a transformer encoder-decoder framework. Input audio is split into 30-second chunks of data. Then, it is converted into a log-Mel spectrogram and processed by the encoder. The decoder is trained to capture the corresponding text, mixed with unique tokens, which direct the model to perform language identification, phrase-level timestamps, multilingual speech transcription, and translation to English.

Due to the nature of the data, which are transcripts of matches' broadcasts, we do not possess any labels associated with the sentiment. That means the task we aim to solve is unsupervised sentiment analysis. (Birjali et al., 2021) has prepared a profound and exciting survey on sentiment analysis and its applications in different tasks. Their comprehensive summary provided not only an exhaustive study of methods that had never before been gathered in one place but also provided some guidelines and heuristics on which method should be used, supplied with detailed comparison. The proposed general division of sentiment analysis task types was machine learning, lexicon-based, hybrid, and other techniques. In his work (Punetha and Jain, 2023) gave a similar division firmly based on the previous work. They are supervised methods, among which authors explicitly mention one of the oldest approaches, which are

machine learning classifiers like SVM (Joachims, 1998) but also some novel approaches that use deep learning (Basiri et al., 2021). The next category consists of semi-supervised methods that use co-training and graph-related methods for imputing labels and improving the model without access to the proper response. Such an approach was shown by (Lin et al., 2011), where they used Latent Dirichlet Allocation (LDA) (Blei et al., 2003) for simultaneous detection of both topics and sentiment from text. Last but not least, there are unsupervised methods where there is no access to the actual labels at all. The authors mention the computational limitations of clustering big datasets, also pointed out by (Vashishtha and Susan, 2019). Among many different methods of unsupervised sentiment analysis, the authors of the article mention lexicon-based methods, which have already been shown in the past to get along well with soccer nomenclature.

The sentiment lexicons methods (Kannan et al., 2016) enable the predictions typically based on the manually created lexicons for which every world is associated with the sentiment. The aspect of manual creation influences two main areas. Firstly, the set of words chosen for the analysis; secondly, the scores, which are the results of an expert decision. That makes the interpretability of the prediction very easy and intuitive. Typically, these methods are tight with a specific language, and the transformation of the dictionary would be needed to apply it to the language differently than designed after the translation of the entries.

We differentiate between two types of lexicons - semantic orientation (polarity-based) lexicons and sentiment intensity (valence-based) lexicons.

The dictionaries used for semantic orientation can have sets that specify a variety of categories. However, in the context of the sentiment analysis, only the positive, negative, and optionally natural categories are used. Each word from a selected sequence, e.g., a sentence or a paragraph, gets assigned a constant score, e.g., one and minus one for each occurrence that can be identified. Then, the score is typically normalized to be compared to other inputs of varying lengths.

The latter group, intensity lexicons, besides the classification to a category, also assigns a numeric significance of the word such that, e.g., the word excellent has a higher score than good.

Regardless of the scoring, both methods still do not encompass the information about modifiers, e.g., the negation ”not.” Therefore, they are often surrounded by some additional rule-based approach that can be modified, e.g., by multiplication, the score directly obtained from the dictionary. We call this method context aware because a single word does not determine the score, but the score is dependent on the larger context.

Vader (Hutto and Gilbert, 2014) is an example of a lexicon-based method encompassing the additional. This method was also created with not only concise messages and Twitter posts in mind but to work well also on longer text forms to ensure that the already available intensity lexicon scores were modified. Additional improvements involve context-based rules.

However, that is not the only way of approaching this problem. As reviews of the state-of-the-art methods done by (Mathew and Bindu, 2020) and (Birjali et al., 2021) show, one can use a supervised pre-trained model to address the issue of analyzing datasets without labels.

Flair is a popular and robust framework for natural language processing that enables the creation and deployment of cutting-edge models for various text analysis tasks. One of the main features of Flair is its ability to combine different types of word embeddings, such as GloVe (Pennington et al., 2014), FastText (Bojanowski et al., 2016), ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019), and Flair's contextual string embeddings (Akbik et al., 2019a).

The last ones are vital innovations that have improved the performance of many NLP tasks. These embeddings are vector representations that capture the meaning of words based on their surrounding context rather than relying on fixed or static terms. Flair uses these embeddings as the basis for all its models, allowing for more accurate and robust text analysis.

Flair has been used and evaluated in many research papers, demonstrating its effectiveness and versatility. Some examples are (Akbik et al., 2018) and (Akbik et al., 2019b), which show how Flair's embeddings enhance sequence labeling and named entity recognition.

In audio, sentiment analysis is typically called (speech) emotion recognition and might involve more classes than NLP methods. The depend on dataset annotation and might include anger, dis-

gust, fear, happiness, pleasant surprise, sadness, and neutral (Pichora-Fuller and Dupuis, 2020) or anger, happiness, sadness, neutral, excitement, frustration, fear, surprise, and other (Busso et al., 2008), or calm, happy, sad, angry, fearful, surprise, and disgust (Livingstone and Russo, 2018). The typical models used for this task are convolutional neural networks and recurrent neural networks due to the interpretation of the time series. A notable mention is the LSTM network (Purwins et al., 2019). The input to those models can differ from raw audio to feature creation like Mel-frequency cepstral coefficients (MFCCs), spectrograms, or embedding (Purwins et al., 2019). An interesting source of the embedding is output from self-supervised wav2vec2.0 (Baevski et al., 2020). The state-of-the-art solution that uses the described methodology is (Pepino et al., 2021). This solution still requires training on supervised datasets, yet it can transform the audio into more meaningful embeddings.

## 3 Datasets

The project is based on the publicly available dataset SoccerNet and its updated version, SoccerNet-v2. The SoccerNet (S. Giancola and Ghanem, 2018) emerged as a benchmark dataset for action spotting in soccer. The dataset is composed of 500 match broadcasts from the UEFA Champions League as well as the top five European leagues: English Premier League, Spanish La Liga, German Bundesliga, Italian Serie A, and French Ligue 1 (each .mpv file with a recording comprises one half of a particular match). It covers three seasons from 2014 to 2017 and a total duration of 764 hours. Additionally, the dataset contains the annotations of three main classes of events: goal, yellow/red card, and substitution. As the data represents the leagues of multiple countries, the match commentaries are in different languages. The dominant languages are English, Spanish, Russian, German, and French ( representing more than 97% of the observations). The dataset contains a few commentaries in Italian, Turkish, and Norwegian, as well as a recording of one-half of a match with Polish commentary.

The SoccerNet-v2 (A. Deliège and Droogenbroeck, 2021) is the expanded version of SoccerNet with significantly more manually specified action annotations, with over 110,000 of them in this version - an average of 221 per match, compared to 13 returned in the previous version of the dataset. Annotation types have been enriched by including events such as ball out of play, corner, direct/indirect free kick, or the penalty (total extension to 17 classes). With this update, a more in-depth analysis of the commentary is possible, depending on the actual events of the match.

Another version of the SoccerNet, SoccerNet-v3 (A. Cioppa and Droogenbroeck, 2022), has also been released. It was built upon the corpus of annotations provided in SoccerNet-v2, enriching them with associated frames from the replay clips. New action-specific data has also been coded, such as the soccer field line (straight or curved) within a given frame, bounding boxes for the players/the referees as well as the objects, including the ball, flag, or yellow/red card, the multi-view player correspondences and jersey numbers. Nevertheless, these additions were introduced with the view to providing a rich environment for the investigation of computer vision tasks related to football. They do not bring new insights into the project's problem of analyzing comments' sentiments. Therefore, previous versions of SoccerNet remain the primary source of the data for the project.

VisAudSoccer (X. Gao and Liu, 2020) is a similar dataset proposed as a benchmark for three different tasks that can be jointly used to produce highlights automatically, i.e., play-back detection, soccer event recognition, and commentator emotion classification. It contains broadcasts from 460 soccer games, 300 of which have been downloaded from the SoccerNet (S. Giancola and Ghanem, 2018) dataset. Audio data is available for 160 games with commentator voices categorized as "excited" and "not-excited." However, the dataset is not publicly available, so it will not be used in the project.

Analyses of voice commentary transcriptions or written live commentaries can be found in many articles. Huang et al. (Huang et al., 2020) conducted an analysis of football match commentary using transcriptions of matches from seven different leagues: Bundesliga, CSL, Europa, La Liga, Ligue 1, PL, Series A, UCL) from Sina Sports Live. The aim of the analysis was to create sports news reports that summarise the match based on its commentaries. Zhang et al. (Zhang et al., 2016) identified common features that have been widely used for general document summarisation and novel task-specific features aimed to gen-

erate sports news from live comments properly. This analysis was conducted based on downloaded commentary scripts from 150 football matches on Sina Sports Live.

## 4 Research methodology

In the upcoming stages of our project, our initial objective is to separate audio tracks from video recordings to isolate the football commentary for thorough scrutiny. We will then proceed to utilize the sophisticated capabilities of the Whisper model to transcribe the audio into text effectively. This critical stage will ensure that the fundamental characteristics of the original commentary are retained for comprehensive linguistic analysis.

Having obtained textual transcriptions, we intend to begin translating the commentary into English while maintaining the original languages to enable a thorough, multilingual investigation.

Afterward, we will proceed to data mining the audio content, which includes examining acoustic features like decibel levels and emotional inflections, referred to as audio sentiment. The aim is to capture the energy and emotions commentators convey during matches.

This study will perform an exploratory analysis focusing on the language complexity used in the commentary. We will examine whether the comment utilizes an uncomplicated, accessible vocabulary or if it features a more intricate diction that could appeal more to passionate football enthusiasts.

The study will include a thorough sentiment analysis using the VADER tool for English content and Flair for English and other language data. This analytical method aims to identify the emotional subtleties within the commentary during different stages of the matches.

Our ultimate goal is to establish a correlation between the emotional tone of the commentary and significant moments in football, including goal-scoring events and red card issuance. We aim to clarify the impact of these moments on the commentary style, thus enhancing our comprehension of the emotional storyline in football broadcasting.

## 5 Experiments and results

The analyses and experiments were carried out in a standard data processing scheme. Firstly, the exploratory data analysis was conducted, with particular emphasis on statistical analysis of comment transcriptions. Then, the sentiments were extracted from the transcriptions. As mentioned in the previous sections, two models were used for this purpose - Vader returning 4 measures describing sentiment intensity: positive, negative, neutral and compound, and Flair tagging the sentence as positive or negative with the certainty of prediction. In parallel, an analogous process was conducted to analyse emotions from audio data using the models dedicated to speech recognition described in the previous sections. The extracted sentiments from both data sources were subjected to further exploration and statistical analysis. Finally, various models were fitted separately for text and audio data to investigate the impact of different commentary emotions on the match event.

Transcriptions from 1046 broadcasts (each covering one-half of the match) were used for the entire analysis. The transcriptions returned by the Whisper model were divided into individual sentences, generating a source data size of 607581 rows. The labels indicating match events were available for 1000 broadcasts (the remaining videos were shared for the purposes of the challenges and competitions), resulting in a total of 110458 rows, each of which stored information on a particular action during the match.

### 5.1 Transcription EDA

We have performed an Exploratory Data Analysis of the transcripted matches data, which consisted of the static text analysis. The goal of this experiment is to assess the complexity and readability of the acquired text to have a better understanding of the linguistic composition and potentially look for some justification in chosen methods. To begin with, we grouped transcripted sentences by matches and merged them to form a unified text. Such an approach allowed us to differentiate potentially completely different matches commented on by different people and simultaneously allowed quantitive analysis of the differences between them. To run static text analysis, we have used *quanteda* (Benoit et al., 2018) R (R Core Team, 2023) package, which is a state-of-the-art tool for comprehensive quantitative text analysis in R language.

To begin with, we started with identifying stopwords, also known as negative dictionaries. They are words that are insignificant from the text representation perspective (Rajaraman and Ullman,

2011) and are usually removed before running any analysis. Such stop words are usually prepositions, articles, etc., or other words that connect particular parts of the text without any particular meaning associated with them. We prepared a copy of all the matches with English stop-words excluded and will compare both versions throughout this section whenever this makes sense. The reason is that we want to underline how great of an influence stop-words have on particular text statistics, which shows the specifics of the commentary textual data, in particular a high number of meaningless words.

Furthermore, we identified tokens and types for each particular match. Types are different unique words that occur at least once in the text. They can be understood as programming language classes - abstract reusable objects. Tokens on the other hand are all words divided into separate entities, they are instances of the aforementioned classes - types. The analysis of the number of tokens and types in the dataset can give a decent initial grasp of the dynamic in the dataset and its complexity. In fact, they are foundations for many different text-related statistics that describe the richness or complexity of the text.

Figure 1 shows a kernel density estimation for the density of the number of tokens for analyzed matches. The plot compares these densities between 2 groups, matches with and without stop-words included. This comparison shows an interesting fact about the specifics of collected transcriptions. These matches have high representations of insignificant words. We see two curves differ significantly and transition from a rather wide distribution without a clear center of mass for matches with stop-words included, to a distribution with a peak that is more packed when we exclude them. We believe that such a big representation of stop-words might be caused by the fact that there are always parts of the match when nothing interesting happens and commentators have to entertain the audience with some fact or improvised stories which are more likely to consist of such words. Certainly, such representation of stop-words will affect sentiment analysis as it might lead to a situation, especially for directory-based methods, there will be little-to-none information useful in assessing the sentiment. Figure 2 shows a relationship between the number of tokens and types for each match. We see, thanks to the soothing line and confidence intervals applied, that the initial linear relationship slows down and flattens reassembling the square root relationship. That notable slowdown in type growth with a bigger number of tokens shows that the vocabulary used when commenting matches is rather limited, even for matches with significantly more words spoken there were not many more new types. That shows limitations and already signals potential problems with differentiating between particular sentences which might affect the quality of the sentiment obtained.

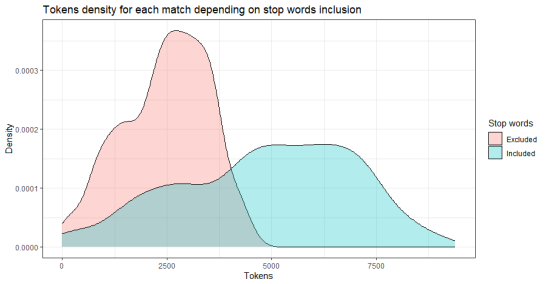


Figure 1: Kernel density estimation curves for matches treated as separate cohorts. Different colors stand for the same dataset with filtered-out stop words.

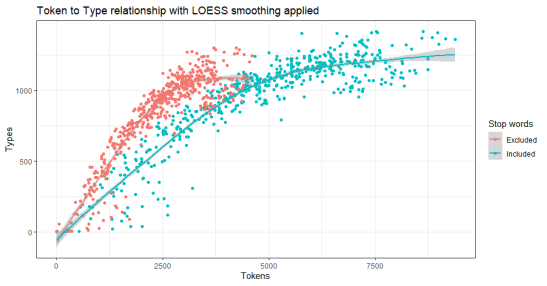


Figure 2: Tokens density depending on stop words inclusion. x-axis represents the number of tokens and y-axis is the number of types. Each dot represents an individual match understood as a separate text. Smoothing curves and confidence intervals are calculated with LOESS method (Cleveland, 1979).

An important feature of any text is its complexity. It affects almost all the areas in which such text could be used starting with reading it out loud for the audience and ending on predictive systems. That is why we found it important to assess the complexity of the transcripted text to get an even deeper understanding of the structure and specification of football match transcription that could

prove helpful to obtain and later understand acquired results. In this work we will use two most popular approaches for measuring text complexity which are Lexical Readability and Lexical Richness.

To calculate the Lexical Readability of acquired football matches transcription we will use the Flesch Reading Ease score. It is a formula suggested in the 1940s by the Associated Press consultant Rudolf Flesch who specialized in finding different metrics for assessing newspapers' readability. Now despite the formula being over 80 years old, it is still well-adopted and used by marketers, research communicators, policy writers, and many others who use it to assess the ease by which a piece of text will be understood and engaged. The higher the score Flesch score is, the more readable the text its. The formula itself is represented by the equation.

$$206.35 - 1.015\frac{n_{words}}{n_{sentences}} - 84.6\frac{n_{syllables}}{n_{words}} \quad (1)$$

The results received were experimentally compared to the United States grading system. Table 1 shows the results of these experiments and at the same time provides an overview of how particular values of the Flesch score correspond to real life.

Table 1: Table representing Flesch Reading Ease score grading and the relation between values and particular steps of the United States education system.

| Score | School Level | Reading Difficulty |
|---|---|---|
| **90–100** | 5th grade | Extremely easy to read |
| **80–90** | 6th grade | Consumers Conversational English |
| **70–80** | 7th grade | Quite easy to read |
| **60–70** | 8th-9th grade | Easily understood by 13-15-year-old students |
| **50–60** | 10th-12th grade | Somewhat difficult to read |
| **30–50** | College | Difficult to read |
| **0–30** | College graduate Professional | Extremely difficult to read |

Figure 3 shows the distribution of the Flesch Reading Ease score calculated for each match. The acquired distribution is unimodal with most of the mass centered in the neighborhood of 80, meaning most of the matches fall into the category between 6th and 7th grade and are easy to read day-to-day English. Most difficult-to-read matches had not less than 45 scores resulting in mid collage text, however, this group had very little probability mass associated with it.

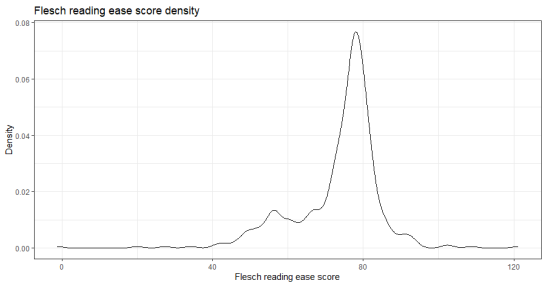


Figure 3: Kernel density estimation curves of the Flesch Reading Ease score calculated for each match. Different colors stand for the same dataset with filtered-out stop words. The context, which is the meaning of particular values and grading is presented in Table 1

There are multiple ways to derive the lexical richness of the text. One of the metrics once again features types versus tokens comparison. This time however we divide one value by the other obtaining a ratio which is a simple yet useful metric portraying the richness of the text. The idea behind that metric is rather straightforward, the more complex text is analyzed, the more varied vocabulary is used resulting in a higher number of types. That means the higher the TTR (token-to-type ratio), the richer the analyzed text is. It is important to note that one of the known drawbacks of this metric is its vulnerability to text length, but this does not apply in our case as particular matches were not long enough for that issue to occur. In Figure 4 we see density estimation for the token to type ratio calculated for each match. Like in the previous analysis, we have estimations for 2 cases, with stop-words excluded and not. Both curves show interesting results, the one presenting TTR with stop-words included once again confirms the high representation and significance of proper stop-words management in this dataset. On the other hand, the curve for TTR with stop-words excluded shows that for most matches, on average each unique word is used slightly less than 3 times, meaning transcription falls on the lower side of the richness. Additionally, if we use our personal experience regarding how the usual broadcast sounds, these rich fragments are almost certainly not uniformly distributed over the entire match which might prove to be problematic when deriving an assessment. Very similar results were achieved and shown in the Figure 5. It features yet another measure of text richness which is the

Hapax score. It is defined as the number of words that occur only once (so types with only one token instance) divided by the number of all tokens. The score follows a similar mechanism and the results match the aforementioned analysis of the TTR results.

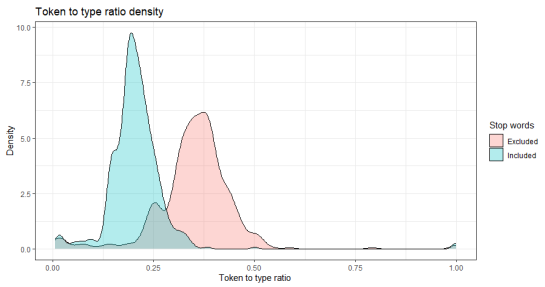


Figure 4: Kernel density estimation curves of the token to type ratio for each match. Different colors stand for the same dataset with filtered-out stop words.

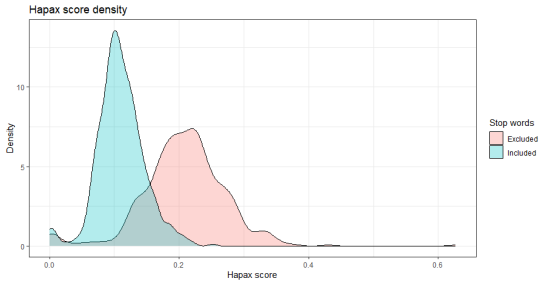


Figure 5: Kernel density estimation curves of the Hapax score for each match. Different colors stand for the same dataset with filtered-out stop words..

## 5.2 Textual sentiment results analysis

We extracted a sentiment from each of the transcriptions translated into English and achieved this in two ways using the Flair framework and Vader based on dictionaries. In this way, we obtained two sentiment classes for each sentence, a score for the sentiment determined by two methods, and added several flags to indicate whether an event of a given type occurred during the sentence.

We began our exploration of the results obtained by examining the sentiment distribution for both methods. This was presented in Figure 6. Note that Vader distinguishes between positive, negative, and neutral sentiments. Flair tries to classify sentences as positive or negative; hence, the corresponding class and its probability are returned.

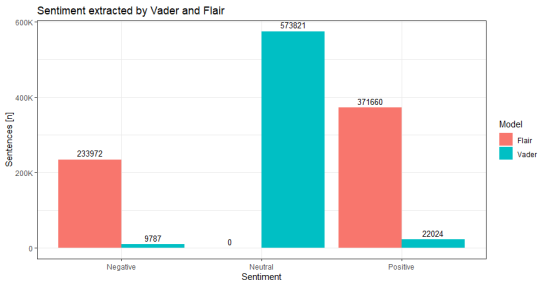


Figure 6: Sentiment labels extracted by Flair and Vader.

To be able to compare the two methods with similar metrics, among Flair's results, we separated out a subset of sentences for which the model was not very confident in the prediction and labeled it as the neutral class. Note, however, that the neutral class designated by Vader and the one transformed from the Flair results are not 1-1 the same. As can be seen from the unified results in Figure 7, the two methods behave differently. Flair was definitely confident in his predictions, classifying most of the sentences with high confidence. Vader, on the other hand, labeled only 5.5% of the sentences as other than neutral.

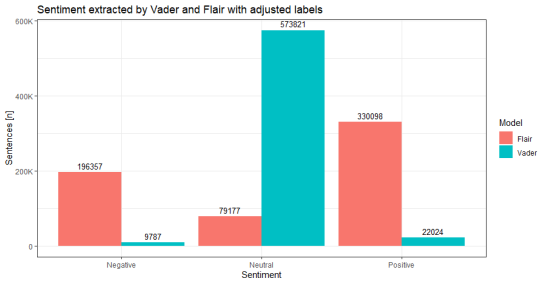


Figure 7: Adjusted sentiment labels extracted by Flair and Vader.

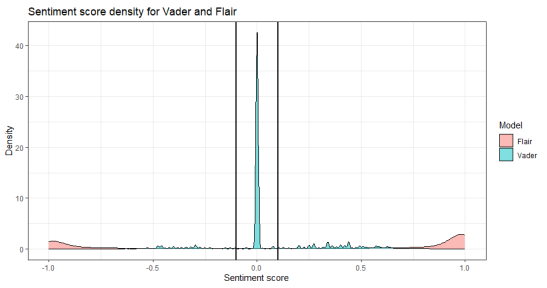


Figure 8: Sentiment score density for Flair and Vader

The designated classes and the scores assigned to them demonstrate the conclusions of the pre-

vious histograms. As a first step, we decided to examine the surrounding of zero for the score assigned by Vader (Figure 8). For the observations in this subset, we compared the percentage of sentences during which a significant event occurred in the game against the percentage in the population as a whole. The results (Figure 9) are not encouraging. It can be seen that practically all events in the game occur in the same percentage in the neutral score environment as outside it.

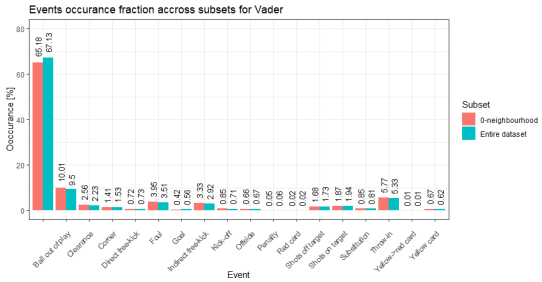


Figure 9: Events occurrence fraction for Vader 0-neighbourhood and whole population

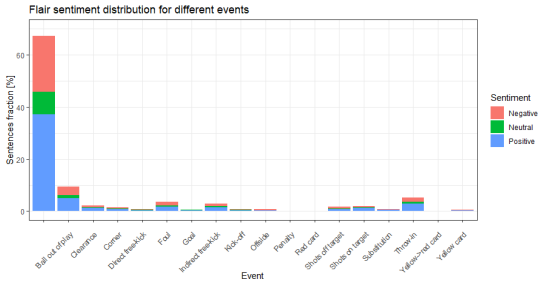


Figure 10: Flair sentiment distribution for different events

For the results obtained with both methods, we checked the distribution of sentence classifications among all events in the game. No significant pattern can be discerned for both Flair (Figure 10) and Vader (Figure 11). There are slight fluctuations within individual events, while all of the events show a significant similarity to the distribution of classes in sentences during which nothing happened (first column on the left).

Examining only single sentences, the relationships between what is on the pitch and what the commentator conveys may be decidedly fragmented. Their granularity can result in the blurring of existential influences amidst the noise. We used a moving average calculation approach for the results obtained with Flair. We calculated the average of the sentences surrounding our target for

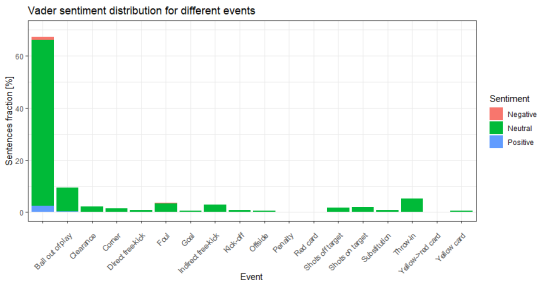


Figure 11: Vader sentiment distribution for different events

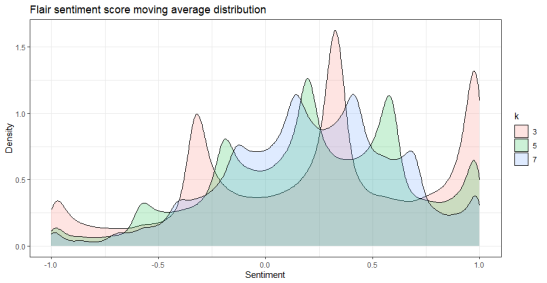


Figure 12: Flair sentiment score with moving average approach

a fixed window size of 3, 5, or 7. We hoped to observe the difference in sentiment around important events. Since the flair model was largely confident in its predictions, this resulted in summing numbers very close to 1 and -1 and then drawing an average from these. Hence, the peaks visible on the Figure 12. Peaks correspond to n sentences with positive sentiment for the rightmost peak and n sentences with negative sentiment for the leftmost peak.

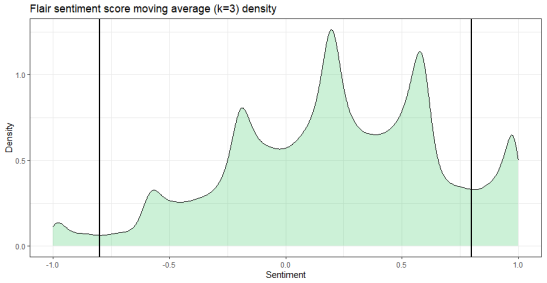


Figure 13: Flair sentiment score with moving average approach, with lines indicating top and bottom peak.

For a window width of 5, we investigated the distribution of individual events for the sentences with the highest and lowest average sentiment (Figure 13). In this case, the results turned out

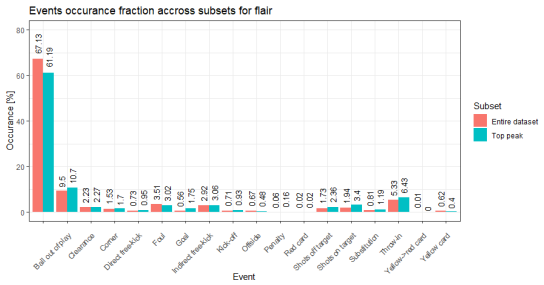


Figure 14: Events occurrence fraction for top peak.

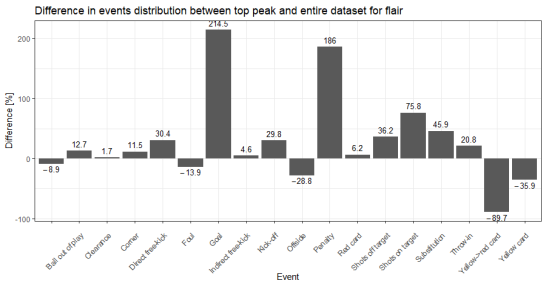


Figure 16: Difference in events distribution between top peak and the entire population.

to be slightly better. For both the top peak (Figure 14) and the bottom peak (Figure 15) one can see a difference in the distribution of events in relation to the population as a whole. These are even more apparent when considering the percentage change in the frequency of a given event for the peak relative to the population as a whole. For the sentences belonging to the top elevation (Figure 16) the percentage of goals increases by more than 200%, penalty kicks by more than 180%, and shots on target by 75%. The percentage of yellow cards, on the other hand, decreases by 36% and second yellow cards by almost 90%. The opposite relationships can be seen for the lower peak (Figure 17). For the sentences with the lowest average sentiment, the percentage of yellow cards increased by almost 200%, red cards by 180%, and shots off target and second yellow cards by about 120%.

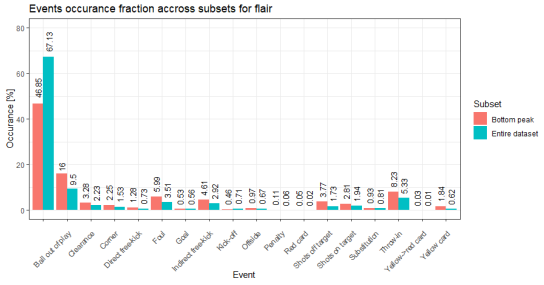


Figure 15: Events occurrence fraction for bottom peak.

However, it should be borne in mind that these are only percentage changes showing, on the one hand, that by actually examining the sentiment for a wider window than just one sentence, we can observe emerging correlations, but on the other hand, in relation to the population as a whole, the significant events in a match are lost amongst the noise. While a commentator utters thousands of sentences, goals, penalties, and shots will be at most a dozen or so.

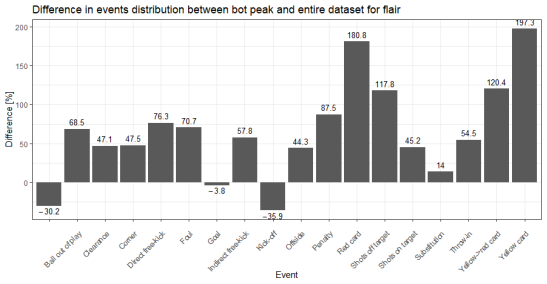


Figure 17: Difference in events distribution between bottom peak and the entire population.

| | Model | Accuracy | AUC | Recall | Prec. | F1 | Kappa | MCC | TT (Sec) |
|---|---|---|---|---|---|---|---|---|---|
| **catboost** | CatBoost Classifier | 0.6263 | 0.5829 | 0.6263 | 0.6144 | 0.5320 | 0.0668 | 0.1115 | 187.1740 |
| **xgboost** | Extreme Gradient Boosting | 0.6237 | 0.5751 | 0.6237 | 0.6046 | 0.5312 | 0.0625 | 0.1017 | 19.9180 |
| **lightgbm** | Light Gradient Boosting Machine | 0.6221 | 0.5689 | 0.6221 | 0.6501 | 0.4948 | 0.0312 | 0.0903 | 29.2900 |
| **gbc** | Gradient Boosting Classifier | 0.6160 | 0.5502 | 0.6160 | 0.6090 | 0.4771 | 0.0089 | 0.0395 | 532.7720 |
| **ada** | Ada Boost Classifier | 0.6149 | 0.5424 | 0.6149 | 0.5771 | 0.4835 | 0.0116 | 0.0352 | 112.0540 |
| **lda** | Linear Discriminant Analysis | 0.6149 | 0.5293 | 0.6149 | 0.5947 | 0.4712 | 0.0030 | 0.0207 | 17.1580 |
| **ridge** | Ridge Classifier | 0.6148 | 0.0000 | 0.6148 | 0.5943 | 0.4711 | 0.0029 | 0.0203 | 6.3260 |
| **dummy** | Dummy Classifier | 0.6145 | 0.5000 | 0.6145 | 0.3776 | 0.4677 | 0.0000 | 0.0000 | 6.8100 |
| **lr** | Logistic Regression | 0.6144 | 0.5265 | 0.6144 | 0.5061 | 0.4677 | 0.0000 | 0.0006 | 19.5060 |
| **qda** | Quadratic Discriminant Analysis | 0.5984 | 0.5313 | 0.5984 | 0.5472 | 0.5263 | 0.0253 | 0.0335 | 14.9280 |
| **nb** | Naive Bayes | 0.5894 | 0.5205 | 0.5894 | 0.5418 | 0.5325 | 0.0226 | 0.0273 | 7.3180 |
| **rf** | Random Forest Classifier | 0.5881 | 0.5726 | 0.5881 | 0.5763 | 0.5800 | 0.1038 | 0.1048 | 507.8360 |
| **et** | Extra Trees Classifier | 0.5825 | 0.5673 | 0.5825 | 0.5751 | 0.5780 | 0.1026 | 0.1030 | 233.4040 |
| **knn** | K Neighbors Classifier | 0.5792 | 0.5750 | 0.5792 | 0.5722 | 0.5744 | 0.0959 | 0.0965 | 620.6680 |
| **dt** | Decision Tree Classifier | 0.5658 | 0.5362 | 0.5658 | 0.5608 | 0.5630 | 0.0728 | 0.0729 | 61.4480 |
| **svm** | SVM - Linear Kernel | 0.4559 | 0.0000 | 0.4559 | 0.4549 | 0.3515 | 0.0012 | 0.0025 | 151.1920 |

Figure 18: Scores for different models.

The final step was to try to catch dependencies using models. We tried to predict the sentiment of a sentence depending on how much time has passed since the last event of a given type and how long it takes before the event happens again. On the one hand, we hoped that the model could predict how the commentators would behave in the setting of important events in the match. However, this task did not yield any significant results. Models trained using the PyCaret framework proved only marginally better than the random classifier

(Figure 18). On the other hand, using tree models, we wanted to see which features the model pays attention to in its prediction. By overfitting the model to the data, we tried to clarify what is considered when making a prediction. However, even this did not allow us to gather satisfactory results. Explaining the model was characteristic of explaining noise.

### 5.3 Audio emotions results analysis

The audio extracted from all matches was divided into 10-second non-overlapping windows. These extracts were used as input to the model predicting emotions (neutral, sad, happy, angry). Over 80% of the predictions constituted neutral emotion, about 10 % happy, 7 % sad, and 1 % angry. The results are presented in Figure 19.

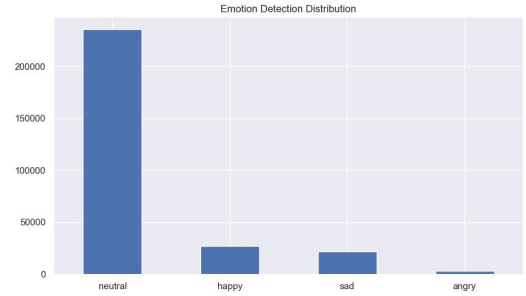


Figure 19: Audio Emotions Distribution.

We also explored the diversity of the emotions associated with different significant events identified in the dataset, like Ball out of play, Goal, and Cards. The results are presented in Figure 20.

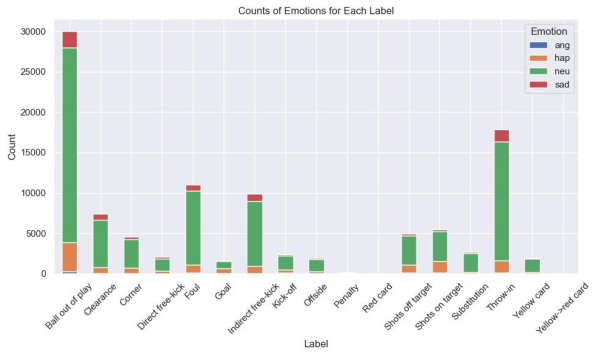


Figure 20: Audio Emotions Distribution.

Finally, we created a prediction model, where the goal was to predict if any of the significant events happened. The input data was the generated emotions from the window preceding the event. The results using different models and metrics are displayed in Table 2.

Table 2: Performance Metrics Comparison

| Classifier | Accuracy | AUC |
|---|---|---|
| LogReg | 0.706198 | 0.500683 |
| XGBoost | 0.705885 | 0.502156 |
| CatBoost | 0.706181 | 0.502027 |
| Dummy | 0.706198 | 0.500000 |
| AdaBoost | 0.706198 | 0.500010 |

The models were not able to learn from the provided data and the effects from the Dummy Classifier are on the same level as the remaining tested models.

## 6 Conclusions

Soccer event detection using emotions extracted based on audio presents a challenge with a large number of hyperparameters to choose from. The poor results might be caused by many factors that need further investigation. A few notable mentions are the window frame length over which the emotion is extracted, the number of historical frames used in the prediction model, and the more advanced model - recurrent neural networks.

When it comes to the textual data, as mentioned in the previous sections, the results are not encouraging. Nevertheless, they give indications of what approaches should be pursued in further analyses. From the analysis of the models as well as the statistical analysis described in the previous section, we can conclude that the tabulated data source from SoccerNet poorly estimates the sentiments extracted by Flair or Vader models. Even the distributions of sentiments in different labels have not shown any significant pattern. The most probable reason for the poor results is that the vast majority of the text data contains noise consisting of single words/parts of sentences typical for football comments, disrupting the models' sentiment detection process. It is also worth mentioning that some of the events stored in the labels data, such as 'Kick-off' or 'Indirect free-kick', were not visible in the video, which could neutralize comments even more.

The results confirm that textual data alone is not sufficient to detect more interesting relationships. Combined audio and text data will be necessary for this.

## 7 Future work

As mentioned in the conclusions, we need to combine audio and text data to achieve more interesting results. Thus, it will be our most important point in the future work of the project. We plan to analyze similarities and differences between extracted audio emotions and text sentiments, which may be crucial in the context of further statistical analyses or models to be defined.

Additionally, as we have focused on the comments only in English, we plan to compare the comments in other languages to explore whether there is a language or group of languages which differs from others in terms of sentiments or emotions.

What is also worth expanding on is the analysis of comments' emotions intensification after specific events during the match. Depending on the type of action, the reaction time of the commentators may vary and have different intensities.

We also plan to predict the next game event or group of events by applying some models based on recurrent neural networks. Considering the fact that some events may happen one after another or the comment's speech/text can be more intensified or emotional before the important action, we may be able to predict the event.

## References


S. Giancola B. Ghanem A. Cioppa, A. Deliège and M. Droogenbroeck. 2022. Scaling up soccernet with multi-view spatial localization and reidentifcation. *Scientific Data*.

S. Giancola M. J. Seikavandi J. V. Dueholm K. Nasrollahi B. Ghanem T. B. Moeslund A. Deliège, A. Cioppa and M. Van Droogenbroeck. 2021. Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

T. Xu G. Brockman C. McLeavey A. Radford, J. W. Kim and I. Sutskever. 2023. Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning, PMLR 202:28492-28518*.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019a. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Alan Akbik, Tanja Bergmann, and Roland Vollgraf. 2019b. Pooled contextualized embeddings for named entity recognition. 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, Volume 1 (Long and Short Papers)*, pages 724–728, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Mohammad Ehsan Basiri, Shahla Nemati, Moloud Abdar, Erik Cambria, and U. Rajendra Acharya. 2021. Abcdm: An attention-based bidirectional cnn-rnn deep model for sentiment analysis. *Future Generation Computer Systems*, 115:279–294.

Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. quanteda: An r package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774.

Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134.

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

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

William S Cleveland. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Felix Hamborg and Karsten Donnay. 2021. NewsMTSC: A dataset for (multi-)target-dependent sentiment classification in political news articles. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1663–1675, Online, April. Association for Computational Linguistics.

Kuan-Hao Huang, Chen Li, and Kai-Wei Chang. 2020. Generating sports news from live commentary: A Chinese dataset for sports game summarization. In Kam-Fai Wong, Kevin Knight, and Hua Wu, editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 609–615, Suzhou, China, December. Association for Computational Linguistics.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Said Jai-Andaloussi, Imane El Mourabit, Nabil Madrane, Samia Benabdellah Chaouni, and Abderrahim Sekkaki. 2015. Soccer events summarization by using sentiment analysis. In *2015 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 398–403.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer.

S. Kannan, S. Karuppusamy, A. Nedunchezhian, P. Venkateshan, P. Wang, N. Bojja, and A. Kejariwal. 2016. Chapter 3 - big data analytics for social media. In Rajkumar Buyya, Rodrigo N. Calheiros, and Amir Vahid Dastjerdi, editors, *Big Data*, pages 63–94. Morgan Kaufmann.

Chenghua Lin, Yulan He, Richard Everson, and Stefan Ruger. 2011. Weakly supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge and Data engineering*, 24(6):1134–1145.

Steven R Livingstone and Frank A Russo. 2018. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

Leeja Mathew and V R Bindu. 2020. A review of natural language processing techniques for sentiment analysis using pre-trained models. In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, pages 340–345.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. SemEval-2016 task 4: Sentiment analysis in Twitter. In Steven Bethard, Marine Carpuat, Daniel Cer, David Jurgens, Preslav Nakov, and Torsten Zesch, editors, *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18, San Diego, California, June. Association for Computational Linguistics.

Ravikumar Patel and Kalpdrum Passi. 2020. Sentiment analysis on twitter data of world cup soccer tournament using machine learning. *IoT*, 1(2):218–239.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Marilyn 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, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

M. Kathleen Pichora-Fuller and Kate Dupuis. 2020. Toronto emotional speech set (TESS).

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In Preslav Nakov, Torsten Zesch, Daniel Cer, and David Jurgens, editors, *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado, June. Association for Computational Linguistics.

Neha Punetha and Goonjan Jain. 2023. Game theory and mcdm-based unsupervised sentiment analysis of restaurant reviews. *Applied Intelligence*, 53(17):20152–20173, Sep.

Hendrik Purwins, Bo Li, Tuomas Virtanen, Jan Schlüter, Shuo-Yiin Chang, and Tara Sainath. 2019. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 13(2):206–219.

R Core Team, 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Anand Rajaraman and Jeffrey David Ullman, 2011. *Data Mining*, page 1–17. Cambridge University Press.

Alexandrea K. Ramnarine. 2023. Unsupervised sentiment analysis of plastic surgery social media posts.

T. Dghaily S. Giancola, M. Amine and B. Ghanem. 2018. Soccernet: A scalable dataset for action spotting in soccer videos. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. 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, Volume 1 (Long and Short Papers)*, pages 380–385, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Srishti Vashishtha and Seba Susan. 2019. Fuzzy rule based unsupervised sentiment analysis from social media posts. *Expert Systems with Applications*, 138:112834.

T. Yang G. Deng H. Peng Q. Zhang H. Li X. Gao, X. Liu and J. Liu. 2020. Automatic key moment extraction and highlights generation based on comprehensive soccer video understanding. *2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*.

Da Yin, Tao Meng, and Kai-Wei Chang. 2020. SentiBERT: A transferable transformer-based architecture for compositional sentiment semantics. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3695–3706, Online, July. Association for Computational Linguistics.

Jianmin Zhang, Jin-ge Yao, and Xiaojun Wan. 2016. Towards constructing sports news from live text commentary. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1361–1371, Berlin, Germany, August. Association for Computational Linguistics.