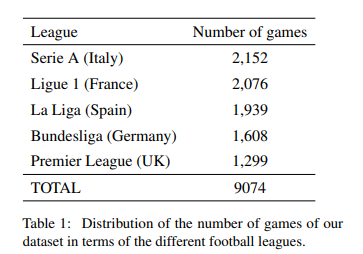
**Analyzing sports commentary in order to automatically recognize events and extract insights**

**Data sources**

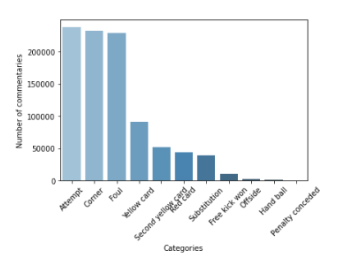
* **Audio dataset:** We obtained from the company Egoli Media a rich audio dataset. This dataset is extracted from **real live sports commentary of the 2021 Paralympic games** and of a **few English Premier League games**. To transcribe the audio into text, we used the Speech-to-Text API developed by Google (Chiu et al., 2018). This tool aims to convert speech into text by using Google’s AI technologies. However, when inspecting the transcription, the dataset seems to be very noisy. It would be very laborious to exploit it, For example: •“Like that shot that got put in on him by the three americans as a work in a dormant volcano cause he’s beginning to bubble” •“Babies and kids the and some of the united states of america so do about it goes as you could be easily” •“The first minute so gym roberts surveys the scene the usa team happy it’s a cluster the key and make things awkward for gb but robots finds is running part of our and fits” We notice that this transcription looks very inaccurate and unreliable. Building a high-performing event recognition and classification model with this data seems very difficult.
* **Textual dataset**: Therefore, we decide to mostly use other sources of data for building and training a model that would perform well. Firstly, we build a complete dataset of clean text commentaries by scraping live scores websites such as bbc.com, espn.com and onefootball.com. These live commentaries are only extracted from football events as they are the most enriched textual commentaries. These commentaries have the advantage to be as well accompanied with a timeline of events that would represent the labels for our training process.



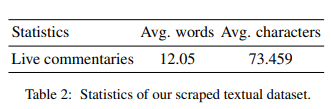
Moreover, we can also consider building the training dataset by scraping the automatic subtitles add-on of YouTube. Indeed, it is also possible to obtain the textual live commentaries from various sports events through the YouTube API. Nonetheless, this method remains laborious as it would imply to manually label a large number of sentences for the training of our model. Therefore, we only use this method to test and evaluate the accuracy of our model on a different dataset.

**Modeling techniques explored**

1. **Cleaning and preprocessing of the textual data:** The dataset scraped from the live score websites contains ≈ 941,000 sentences labeled with one of the 12 categories (Attempt, Corner, Foul, Yellow Card, . . . ). These textual commentaries are extracted from 9074 real games that occurred since 2011 in the five biggest European football leagues. An essential step in the preprocessing of this dataset was to check that the dataset was balanced enough to train a classification model on:

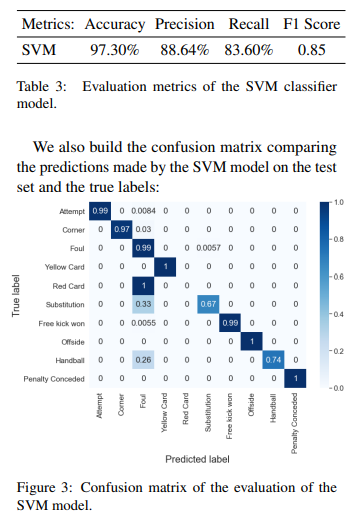
As we can notice on Figure 2, even if only a few commentaries belong to the categories “Offside”, “Hand ball” and “Penalty conceded”, there is not one category that holds more than 50% of the commentaries. However, we could still consider rebalancing the dataset using oversampling (Chawla et al., 2002). Furthermore, we apply some simple textual cleaning techniques:

• Removing punctuation, special characters, leading, trailing and extra white spaces/tabs • Stop-word removal • Stemming • Lemmatization



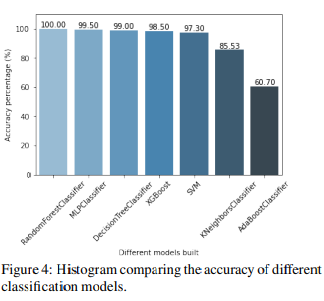
1. **Modeling using tf-Idf:** On one hand, for our word embedding, we build vectors from this cleaned textual data using the Term Frequency-Inverse Document Frequencies (tf-Idf). Using this method seems logical because in order to classify the textual live commentaries, it is essential to encode the frequencies of the most important term in each sentence with their relevancy (Liu et al., 2018). This is also the method that gives us the best results in building an accurate classification model. Equipped with these embedded sentences with tf-Idf, we are now able to train, evaluate and compare different classification algorithms.
2. **Modeling using BERT:** On the other hand, we also consider using BERT Transformers to model our text classification task. Indeed, BERT is designed to help computers understand the meaning of ambiguous language in text by using the bidirectional surrounding text to establish context (Devlin et al., 2019). Therefore, BERT is a State-of-the-Art language model that distinguishes itself from previous language models such as word2vec and GloVe, which are limited when interpreting context and polysemous words. This is very pertinent to our task. To classify the live commentaries efficiently, we would want our model to take into account the surrounding context of each word. This is why we will explore and evaluate this modeling technique more thoroughly.

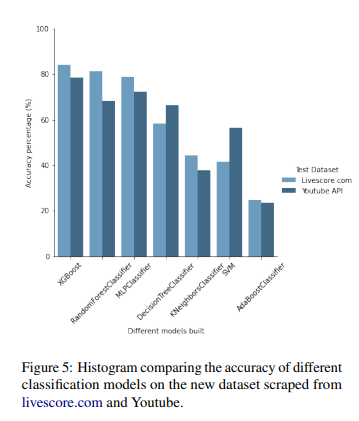
**How dataset is test and trained:**

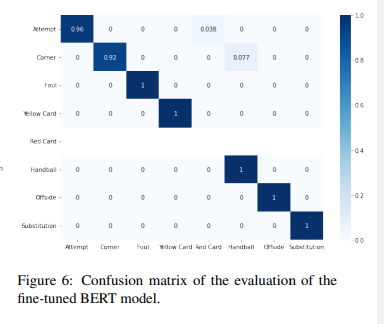
****The research paper (Minard et al., 2016) studies how an SVM model could be trained in order to detect and classify the main events in newspapers commentaries. As this study represents our baseline, we try to reproduce its methods and techniques using our large live commentaries dataset. Firstly, **we randomly split the tf-Idf embedded vectors into a train and a test subset using ratios of 80% / 20% respectively**. We decide to shuffle the data when splitting, as we want our model to be well-functioning on every type of game and for every commentary made at any time during the game. We then train our SVM classifier model and evaluate its metrics on the test set:

At first sight, we can observe on Figure 3 that most of the errors concern commentaries that are mislabeled as "Foul". For example 26% of the true labels "Handball" are being mislabeled as "Foul". However, these errors can be caused by the fact that most handballs in football are considered to be fouls as well. Hence, the clear distinction between the categories "Foul" and "Handball" could be quite ambiguous in football. On the other hand, we also notice that our SVM is still very powerful and can very accurately classify the main actions of various football games. **The baseline model built in the research paper achieved an F1 score of 0.71. In comparison, our SVM model attains an F1 score of 0.85** (Table 3). This important improvement could be explained by the fact that we have trained our model on a much larger dataset.

**Classification Model using tf\_idf:**

****In order to extend our analysis, we build various other classification models using the same embedded vectors with tf-Idf (Shaikh, 2017). We also use the same train/test splitting methods for training and evaluating the different classifiers. **To evaluate and compare these different classification models, we use the accuracy metric (accuracy = Number of correct predictions Total number of predictions ).** For our task, we are willing to get the rightest predictions for the commentary of each game. Therefore, we build and train the different classifications models. Firstly, **we compare the accuracy of each model on the same type of live sports commentaries**:

On the other hand, we want our classification model to be able to generalize well and to perform well on different types of live commentaries structures. For example, we would like our model to perform effectively on commentaries extracted from other sources like Youtube or livescore.com. Therefore**, we also compare the accuracy of each model on new these types of live sports commentaries which have different semantic structures and syntactic expressions**:

**BERT Classification Model:** In addition, to model our classification task, we study how implementing a BERT classification model could improve our results. Indeed, we expect that the use of BERT Transformers would confer us a better accuracy as this modeling technique is taking into account the surrounding context of each word. Therefore, we train on Google Collab3 our model with our dataset using the existing transformer pretrained on a large corpus of texts (’bert-base-uncased’). The dataset which we use to fine-tune the BERT model is similar to the one used previously. However, we do not proceed to the tf-Idf vectorizing step in this case. We evaluate our classification model on unseen live sports commentaries and obtain a very high accuracy of 99.8%. This new model is also evaluated on new types of sentences structures using the dataset scraped from livescore.com and Youtube. It attains a also very high accuracy of 92% in average.

**Summary of How data is collected and how models are trained:** The data collection process involves two main sources: an audio dataset obtained from Egoli Media, extracted from real live sports commentary of the 2021 Paralympic games and English Premier League games, transcribed using Google's Speech-to-Text API; and a textual dataset constructed by scraping clean text commentaries from live score websites like bbc.com, espn.com, and onefootball.com. The audio dataset, though noisy and challenging to exploit due to inaccuracies in transcription, provides real-world commentary content. The textual dataset, comprising approximately 941,000 sentences labeled with 12 categories from football events since 2011, undergoes preprocessing to ensure balance and includes steps like punctuation removal, stop-word removal, stemming, and lemmatization.

For the modeling techniques, two main approaches are explored. Firstly, the tf-Idf method is employed for word embedding, creating vectors from the cleaned textual data. This method, effective for encoding term frequencies and relevancy in sentences, allows the training, evaluation, and comparison of different classification algorithms. Secondly, BERT Transformers are considered for text classification. BERT, designed to understand ambiguous language by considering bidirectional surrounding text for context, distinguishes itself from previous models like word2vec and GloVe. The goal is to leverage BERT's ability to capture contextual information and effectively classify live commentaries. The modeling techniques are applied to the datasets to build classification models, and the performance is evaluated, providing insights into the suitability of each approach for event recognition and classification tasks.

**Summary of How dataset is test and trained:** In the study by Minard et al. (2016), the authors aimed to detect and classify main events in newspaper commentaries using an SVM model. Replicating their approach, the researchers utilized a large live commentaries dataset, initially splitting tf-Idf embedded vectors randomly into 80% for training and 20% for testing, with data shuffling to ensure model robustness. The SVM classifier was then trained, and evaluation metrics were applied to the test set. Notably, the SVM model exhibited a high F1 score of 0.85, surpassing the baseline model's 0.71, attributed to the larger training dataset. Expanding their analysis, the researchers built various classification models employing tf-Idf embedded vectors and assessed them using the accuracy metric. The models were tested on both similar and diverse live sports commentaries, showcasing their ability to generalize across various semantic and syntactic structures. Additionally, a BERT classification model was explored, leveraging the transformer's contextual understanding. Trained on a dataset similar to the previous models but without tf-Idf vectorization, the BERT model demonstrated exceptional accuracy, reaching 99.8% on unseen live sports commentaries and 92% on diverse sentence structures from livescore.com and YouTube datasets. Overall, the study highlights the efficacy of SVM models, the impact of dataset size on model performance, and the promising results achievable with advanced models like BERT.