TOWARDS COMMENTARY-DRIVEN SOCCER PLAYER ANALYTICS

A Thesis

by

RAHUL ASHOK BHAGAT

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	James Caverlee
Committee Members,	Alan Dabney
	Frank M. Shipman
Head of Department,	Dilma Da Silva

May 2018

Major Subject: Computer Science

ABSTRACT

Open information extraction (open IE) has been shown to be useful in a number of NLP Tasks, such as question answering, relation extraction, and information retrieval. Soccer is the most watched sport in the world. The dynamic nature of the game corresponds to the team strategy and individual contribution, which are the deciding factors for a team's success. Generally, companies collect sports event data manually and very rarely they allow free-access to these data by third parties. However, a large amount of data is available freely on various social media platforms where different types of users discuss these very events. To rely on expert data, we are currently using the live-match commentary as our rich and unexplored data-source.

Our aim out of this commentary analysis is to initially extract key events from each game and eventually key entities like players involved, player action and other player related attributes from these key events. We propose an end-to-end application to extract commentaries and extract player attributes from it. The study will primarily depend on an extensive crowd labelling of data involving precautionary periodical checks to prevent incorrectly tagged data. This research will contribute significantly towards analysis of commentary and acts as a cheap tool providing player performance analysis for smaller to intermediate budget soccer clubs.

ACKNOWLEDGMENTS

I would like to acknowledge and thank all the people who have helped me during my entire journey into this research. Firstly, I would like to thank my Committee Advisor, Dr. James Caverlee for providing me with such a challenge opportunity and constantly supporting and inspiring me into channelizing my passion into research. I have learned a lot under his guidance at both personal and professional levels.

I would also like to express my gratitude to my other committee members, Dr. Frank Shipman and Dr. Alan Dabney, for their positive feedback and encouragement throughout this process. I would also like to recognize and credit all the help I received from the members of InfoLab through lab meetings and interactions. I am specially grateful to Majid Alfifi, Parisa Kaghazgaran, Siddharth Verma, and Prafulla Choubey for their valuable inputs and suggestions.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a thesis committee consisting of Dr. James Caverlee [advisor] and Frank M. Shipman of the Department of Computer Science and Professor Alan Dabney of the Department of Statistics.

All work for the thesis (or) dissertation was completed by the student, under the advisement of Dr. James Caverlee of the Department of Computer Science.

Funding Sources

There are no outside funding contributions to acknowledge related to the research and compilation of this document.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Football (soccer) is like a religion to me. I worship the ball, and I treat it like a god. Too many players think of a football as something to kick. They should be taught to caress it and to treat it like a precious gem. – Pelé

Soccer is the most popular game in the world. It is a dynamic game, involving complex tactics, game strategies, and individual contributions. Making sense of this rich pool of activities via *analytics* is an important factor for (i) individual players – so that they can learn from previous mistakes, identify weaknesses in their opponents, and create new opportunities along with their teammates; as well as (ii) teams – so that they can scout new talent, identify successful lineups, and prepare for opponents. These analytics allow players and teams to better understand what has transpired in a recently completed performance, and how this performance fits into the pattern of cumulative athletic behavior over a year or season [14]. The traditional assessment of players and teams has been based on "subjective" observations by experienced scouts and coaches, but there is growing effort at creating new "objective" methods that are data-driven and hopefully not constrained by biases[1].

Although soccer is by far the most followed sports in the world, data analytics has yet to achieve the same level of sophistication as analytics in any other professional sports. A sport like baseball has clearly punctuated actions – a pitch, a hit, an out – that are straightforward to measure. A more dynamic game like basketball also has a multitude of actions – shots, blocks, assists – that give rich evidence to the flow of the game. In contrast, soccer is extremely fluid, with many games having only a single goal over 90 minutes of play. This lack of obvious measurable actions and fluidity of play makes it challenging to quantify the contributions of individual players beyond simple counts of goals, assists, and saves. However, there is little emphasis on other factors like blocks, passes and tackles that can be crucial in governing the flow of the game towards the final outcome.

[1]We use subjective and objective loosely here. Fairness and accountability in data-driven methods and learning-based approaches is a critical research challenge; algorithms can be just as biased as people.

In one direction, researchers and practitioners are exploiting new video-based methods to monitor and analyze player actions in soccer. For example, Xie et al. [15] exploited aspects of video analysis including color and motion intensity to classify video into play and break phases. A recent effort from Perin et al. [16] called *SoccerStories* supports the visual exploration of soccer "phases" (sequences of actions from one team until it loses the ball) to help experts gain better insights. While encouraging, such video analytics approaches typically require special camera setups or expensive processing.

In contrast, we aim in this thesis to exploit a rich, but relatively untapped resource about soccer – play-by-play commentaries. Commentaries are short though descriptive narratives of the play-by-play events of a game. The information contained in the commentary are a useful source for information extraction about team and player performance. For example, in a particular commentary at the 76th minute of the English Premier League match between Manchester United and Chelsea on the 5th November, 2017:

> "SAVE! *Hazard* continues to provide a spark in the attacking third for Chelsea and he runs into space, then firing at goal where *De Gea* can collect. The Blues are banging on the door again though."

The comments clearly defines that there was a beautiful run and a shot at goal by *Hazard* and an instantaneous save followed by *De Gea*. Additionally, it also gives an idea that *The Blues (Chelsea)* are continuously attacking at the moment.

Such commentaries often serve to compensate for imperfections of the visual modality of the medium in creating the live game more entertaining [17]. They also provide a good deal of analysis of the game as well as the player performance. Many soccer enthusiasts and fans keep up-to-date of their favorite teams' performance through game summaries or highlights. Usually, these summaries are an overview of key events of the game like goals, saves, cards, or any other crucial action influencing the final score for both teams [18]. Compared to video analysis that may provide detailed player positioning, tactical changes, team strategy, and many more, commentaries offer the potential of extracting key actions and events.

Figure 1.1: Old newspaper clippings for soccer matches. The image is a digital combination of 4 reprinted images from different sources cited in a clockwise order [1], [2], [3], and [4].

Commentaries also provide a potential window into the history of soccer games for which there are commentaries but no video evidence. There are numerous sources available for commentary. Old soccer transcripts can be used to extract information of players and teams of the past duration which can be then used for comparison of current player or team performance. Some of the match-turning commentary excerpts from as far in the past as 1966 have been given in the Figure 1.2 and 1.3. This level of information is intermediate to what we can gather from textual highlights or summaries and the video analysis. This rich and "clever" data remains still unexplored for soccer analysis and if analyzed deeply, it can serve as one of the best data source for sports analytics due to its vast variety and abundance.

However, taking advantage of these commentaries is not without its challenges. We faced the initial difficulty of the absence of a medium for the verification of our annotation of the data set.

		
Commentary	The West German defense was like a stone wall again England launched an offensive and again the German blockade could not be penetrated.	Now Germany on the left wing with Schäfer. Schäfer's pass to Morlock is blocked. Boszik, always Boszik, the right winger of Hungary, has the ball. But he's lost it to Schäfer. Schäfer crosses to the centre, header, it's cleared. Now Rahn could shoot from a distance, Rahn shoots ... Goal, goal, goal, goal!
Location	An intercept of England vs West Germany in the 1966 World Cup Final.	Radio commentator Herbert Zimmermann relates Helmut Rahn's winner for West Germany against Hungary in the final of the 1954 World Cup – the 'Miracle of Berne'

Figure 1.2: Commentaries from old matches Part 1. The images are digital snapshots of video highlights available from [5] and [6].

		
Commentary	Everything now is in the hands of Antonín Panenka. Antonín Panenka can decide it; if he hits the target, we are European champions. Antonín Panenka, 28-year-old midfielder of Bohemians Praha, Maier in the goal. Panenka! Goal! Goal! We are European champions for the year 1976! It really is true, the fight is over. Bravo, bravo, bravo!	"Goooooooooaaaal for Charisteas! We advance, they retreat! Check the history books again! Perhaps Vasco da Gama was Greek? Greece are on the verge of becoming European champions. I'm done commentating!"
Location	Czechoslovakia's 1976 UEFA European Championship final meeting with West Germany	Sport FM's Giorgos Chelakis, who had christened Greece's UEFA EURO 2004 team 'The Pirate Ship', celebrates Angelos Charisteas's winner against Portugal in the final

Figure 1.3: Commentaries from old matches Part 2. The images are digital snapshots of video highlights available from [7] and [8].

Since this field of text analytics in the sports domain is relatively untouched, we found it arduous to locate any related work to this research. At a time, we are dealing with a single commentary which can be composed of single or multiple statements and can be comprised of from zero to numerous entities (players, teams, coaches and referee) involved. Due to limitations in the performance of

Natural Language Processing based systems, we had troubles in differentiating these entities into categories along with identifying and associating the references of these entities in the text.

Hence, this thesis aims to exploit the potential of soccer commentaries as a rich resource for player analytics. By identifying key actions in these commentaries – including goals, assists, and saves, but also less obvious actions like creating chances or making a key mistake – we hope to provide a foundation for assessing player quality and complementing existing video analytics approaches. In summary, the main contributions of this thesis are:

- First, we propose an end-to-end framework as shown in Figure 1.4 for extracting key actions and players from soccer commentaries by utilizing web-scraping, natural language processing and machine learning approaches.

- Second, we develop a crowdsourcing data curation approach for labeling key actions and players in soccer commentaries. Using this approach, we curate a dataset containing 84910 commentaries from 936 matches across 9 international leagues. The crowdsourcing application is still active and we are able to collect hundreds of commentaries on a daily basis.

- Finally, we test a suite of classification methods for classifying self-curated labels. We find that we are getting an accuracy of around 98% for multi-class classification and an accuracy of 97% multi-label classification.

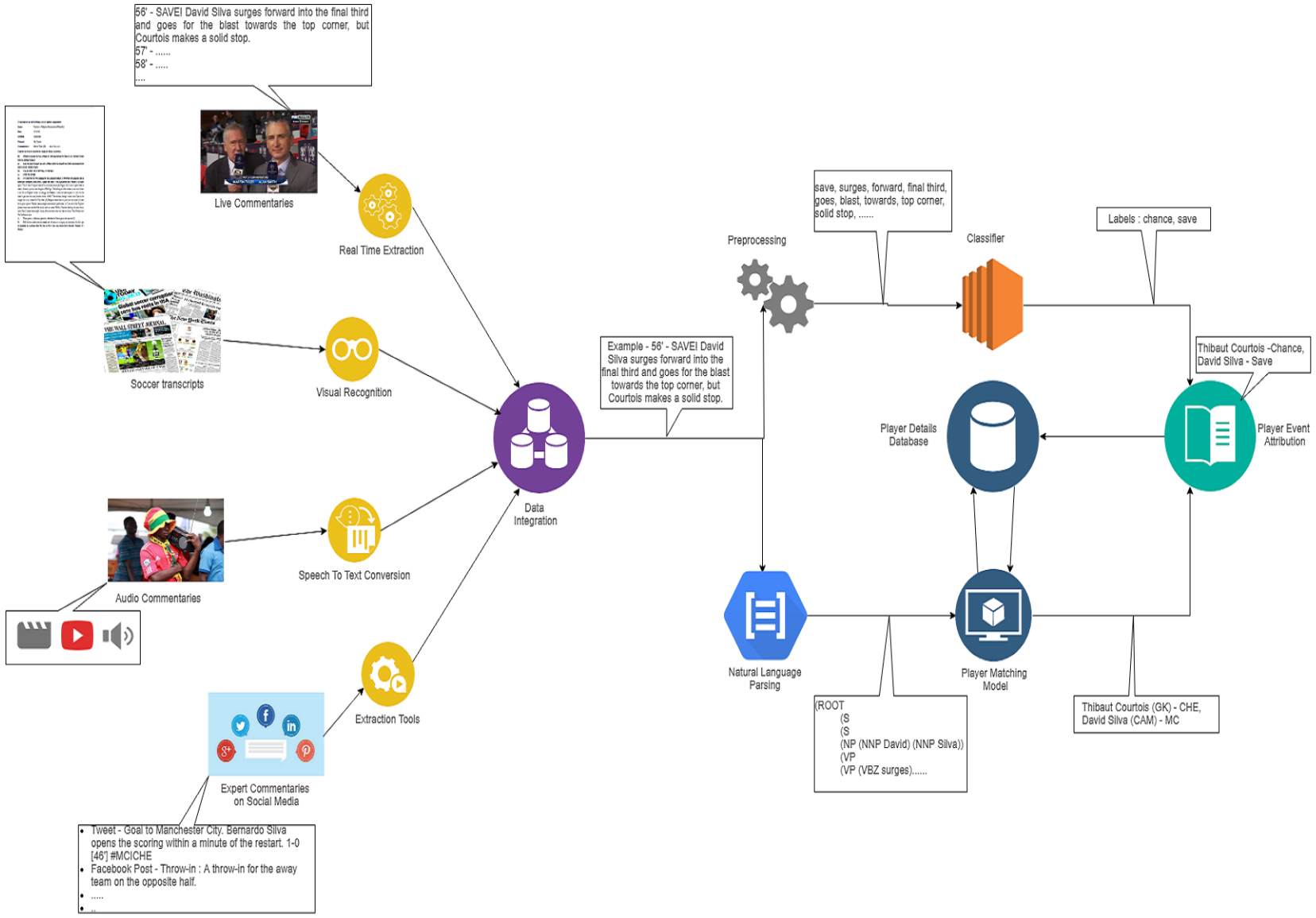


Figure 1.4: Workflow for an end-to-end system. The image is a combination of different reprinted images referenced from [9], [10], [11] and [12].

2. RELATED WORK

In this chapter, we highlight related work from video-based sports analytics, language model based approaches, and other audio and information extraction methodologies.

2.1 Video-based analytics

Over the last decade, analysis on soccer video has attracted much research and led to the deep insights into possible applications on a wide scale like analysis of tactics, auto-identification of highlight events, auto-summarization of play, verification of referee decision, player and team statistics evaluations, content based video compression, better video entertainment through graphical object overlapping, innovative advertisement insertion etc. For surveillance systems too, soccer context can be referenced as a specific application in some ways. Generally, TV broadcast cameras and specialized proprietary-fixed cameras (suitably placed) around the playing field are two possible sources of soccer image streams. Broadcast images, as worked in many papers, are described with the aim of recognizing significant events for different media streaming sources like television, mobile phone, and Internet services. Proprietary cameras are suitable for more specific tasks which are not deducible using broadcast cameras such as recording, analysis of team strategies, 2D/3D reconstruction and visualization of player actions [19]. Video is segmented based on *dominant color ratio* and *motion intensity* and classified into play and break phase using Hidden Markov Models in [15]. [16] designed *SoccerStories* - a system for the visual exploration of soccer 'phases' (a series of actions or events by one team until the ball possession is lost). *SoccerStories* explores the game through phases and help strategy experts to get a better insight by offering compact yet expressive standard visualizations.

[19] categorize three major application areas of soccer video analysis - video summarization, provision of augmented information like player identification, team recognition, and camera calibration, and high-level analysis which includes detecting of player skills, identifying team strategies, and extracting tactical formation. [20] identifies stable temporal structures (*T-patterns*) that provide

information about continuous and concurrent interaction among soccer contexts with respect to lateral position and zone.

Though video analysis is growing with a rapid pace, the technique involves specialized hardware requirements and more complex computations. For example, methodologies in surveillance tasks involves a lot of constraints : high fluctuations in the lighting conditions, rapid dynamic events, complex situations of occlusions, real time analysis, precise and accurate player position on the field, and so on, and thus cannot be directly applied in the context of a soccer match [19]. Other difficulties involved in player tracking, object detection and activity analysis are overlapping of players wearing the same uniform, unpredictable ball trajectories, adaptability to varying lighting conditions (natural and artificial) in the same match, ball invisibility due to bad lighting condition and wide-angled camera view, and other positions and ball-player interactions dependent complex events. It is for these reasons that analyzing soccer context is very challenging for the computer vision community.

2.2 Audio features-driven analytics

The majority of research in the field of soccer for event extraction is based on drawing out highlights from audio and video contents. In [21], the authors explored the ability to extract highlights automatically using audio-track features alone. Audio keyword provides more intuitive results for event detection in sports video, specifically soccer videos, compared with the method of event detection based on direct information extraction using low-level features like Zero Crossing Rate (ZCR), Spectral Power (SP), Linear Prediction Coefficient (LPC), LPC-derived Cepstral Coefficient (LPCC) and Mel-Frequency Cepstral Coefficient (MFCC) in [22].

2.3 Linguistic model

[23] exploited various approaches to recognize named entities and significant micro-events from large volumes of user-generated social data, specifically tweets, during a live sport event. They also described how combining linguistic features with background knowledge and using Twitter specific features can be used to achieve high precise detection results. [24] makes a com-

parison among the three national football league in the context of sports marketing, and illustrates how a range of factors influence the engagement of fans or followers on social media like Twitter. People are already discussing the game on various social media platforms like Twitter, Facebook, Instagram etc. Realizing this, [25] proposed an approach to use Twitter data for highlight detection in soccer and rugby matches by mining user tweets. They detected "interesting minutes" by looking at the sudden rise or peaks in the Twitter stream and their results were comparable to highlight detection from audio and video signals, however, it still suffered from a high number of false positives.

2.4 Other information extraction models

[26] presented a system with integration of Visual Analytics techniques into the analysis process for high-frequency position based soccer data at various levels of detail. Several work on game-related performance of the players and teams have also been presented. [27] defines Passing as a cardinal soccer skill and utilizes this fundamental observation to define and learn a spatial map of defensive weakness and strength of each team. The focus towards more sport-specific metrics like player movement and their similarity to other players and uniqueness in terms of their in-game movements have been analyzed in [28]. Research has also been performed on the prediction of the outcome of soccer matches to be used for betting on the winning team [18]. [29] presented mobile application usage for real-time opinion sharing and used the collected data to exemplify the aggregated sentiments corresponding to important moments, the outcome of which can be used to summarize the event.

Though, there have been a lot of research in the field of sports and soccer in particular as well. However, the direction of research that we are heading has still been relatively untouched. The work in [18] motivated us to move forward with our initial model of event classification and develop things on top of it. Dealing with commentaries is more about processing raw texts which is underlying definition of Natural Language Processing (NLP). As defined in [1], Natural language

[1] http://searchbusinessanalytics.techtarget.com/definition/natural-language-processing-NLP

processing (NLP) can be defined as the ability of a computer program to understand human language as it is spoken and comprehended. However, developing a program that understands natural language is a difficult problem [30]. NLP is still a hot topic and its future will be redefined as it goes through new phase of technological challenges and a governing force from the market for a move towards more user-friendly systems [30]. This dependence on the market and technological obstacle on the processing side has restricted our options.

3. GOALS, CHALLENGES, AND APPROACH

In this chapter, we address the overarching goals of this thesis, the key challenges we face, and the proposed structure of our approach.

3.1 Goals

In this work, we will develop an application which will enable crowd labelling to our datasets. We will restrict the task of tagging to individuals who satisfy certain criteria to the labelling standards set by our application. This is necessary to filter out junk and spam data. With the available crowd labelled dataset, we propose and evaluate methods to detect meaningful information from soccer commentaries. Our primary task is to identify key events in the given data set and later use that information to extract player attributes.

Our main focus for data collection is on live commentaries. We restricted our model to GOAL.com only as it is the most vibrant and widely popular soccer community in the world which provides regional as well as international soccer commentary, news, articles and entertainment through both internet and mobile platforms. Moreover, GOAL.com produces a monthly outright coverage of over 60 million football fans around the world. GOAL.com has a strong editorial team of 500+ experts who deliver football expertise and unique insight to thousands of pieces of every day gossips, drama, events and soccer stories in multiple languages suitable to the fans all over the globe.

3.2 Challenges

In this section, we define the challenges encountered while working on this research. We faced the initial difficulty of the absence of a medium for the verification of our annotation of the data set. Since this field of text analysis in the sports domain is relatively untouched, we found it arduous to locate any related work to this research. At a time, we are dealing with a single commentary which can be composed of single or multiple statements with zero to numerous entities (players, teams, coaches and referee) involved. Due to limitations in the performance of Natural Language

Processing based systems, we had troubles in differentiating these entities into categories along with identifying and associating the references of these entities in the text.

3.2.1 Lack of ground truth

Commentary is an excerpt of live play as given by an expert of the game. Since there is a human involvement in the process, commentaries are not always perfect. Sometimes, same commentary can be interpreted differently by two different persons. Many a time, same action or event in a live play can be described very differently by different commentators. Sporadically, commentaries in different languages are just a translation from one language to another, and this involves the translation errors involved in the process.

To understand this better, let us consider the commentary at the 18th minute of Brighton Hove Albion vs Chelsea match of Barclay's Premier League on 20th January 2018 -

> *SAVE! Chelsea work the ball down the right and space opens up for* ***Bakayoko*** *to strike from the edge of the box, but* ***Ryan*** *gets down to make the save.*

From the commentary, there is clear indication of a good save by Brighton Hove Albion goalkeeper *Ryan*, however, the ambiguity lies in deciding the action performed by *Bakayoko*. If one assumes *Bakayoko* as a striker, then the given scenario can be seen as a chance for the striker and his strike missed out on this opportunity. However, the scenario changes if we know consider his real position - the defensive midfielder, the commentary now has a different meaning with *Bakayoko* creating a chance from just outside of the box which has been well saved by *Ryan*.

Certain pattern recognition algorithms assume that the process of annotating data-set has been attained in a reasonably objective and reliable manner. However, this assumption is not valid as we can only have the subjective opinion(s) of experts. The application of supervised learning algorithm to a data set is restricted by two factors: (i) Examining the relative performance of expert opinion(s) and algorithms, and (ii) training a system in the absence of universal ground truth [31]. In our case, we can only afford the subjective assessment of data-set as explained in the example earlier and that bounds the performance of our system.

3.2.2 Player attribution

Apart from the basic attributes of physical fitness and ball-body coordination, the success in games involving ball-sports can be attributed to the ability of player to access and process information in the complex and quickly changing contexts. This ability to make decisions in such competitive environment that results in game changing scenarios is referred to as 'creativity' [32]. Attention to details and expertise in the game contribute significantly towards the development of creative thinking. The skills of these successful players can then be used as a benchmark for making comparison to a population norm. The characteristics of a good player is defined by his presence of mind, spatial attention, mentalizing capacity and working memory. These are important traits which segregates elite athletes from sub-elite or novice[33]. There has been an active focus on soccer development programmes by almost every football club to promote youth talent with an eye for "spotting a future star". These development programmes are designed by keeping in mind the huge financial implications to the youth soccer along with increased engagement of player towards sports participation through improved professional training and education [34]. However, scouting talent is not an easy task by itself. Hungarian Football Federation was going through drastic decline of the general quality and standard in Hungarian soccer in the last few decades. Several attempts (Bozsik Program: 2002–2004; Bozsik Program: 2006, Bozsik Program: 2011–present; Góliát McDonald's FC:1999–2004) to revive and renew the soccer standards were taken, however, all these programs were ineffective in providing remedy to their problem. The programs failed mainly due to insufficient utilization of resources and scarcity of methodically sound approaches to youth development [35]. The technology can play a crucial role here in identifying talent with an economically viable option for any soccer organization.

The match transcripts of games between different clubs can be readily available. Also, regional newspaper agencies write articles about local sports. So, there is ample amount of data available about player information in the form commentaries, blog articles or newspaper columns. This "rich" source of data set is full of information to identify player attributes as well as scout talent.

Figure 3.1 shows the WorldSoccer[1] [13] magazine articles outlining the player attribution. Figure 3.1a references the moments leading to the heroic and nefarious display of players and coaches. Figure 3.1b highlights the emerging young talent of diverse nationality. These articles exemplifies that enormous information attributing to talent identification existing in several forms across computer and new media devices.

(a) (b)

Figure 3.1: Clippings reprinted from WorldSoccer[1] Magazine's January 2018 edition [13]. (a)Heroes & Villians" of soccer. (b)"On the radar" - A spotlight on the players and coaches making waves.

The difficulty lies in attributing the action to the deserving player. The text can be very ambiguous at times, especially when the reader is novice to the subject. For example, take the following

[1]http://www.worldsoccer.com/publication/world-soccer/world-soccer-january-2018

commentary from an attacking move by Watford against Chelsea at the 73rd minute of the English Premier League encounter between the two teams on 5th February, 2018 -

> *CHANCE!* ***Deeney*** *intercepts when* ***Deulofeu*** *looks certain to score! A low cross from the left is heading straight for the Spaniard when his captain steps in, eventually attempting to tee up his team-mate who, by then, has no angle to aim at.*

The scenario describes that there was a low cross from the left wing heading straight for the Spaniard but the captain intercepted in between. There are two major occurrences of vagueness in this text commentary.

Domain knowledge and entity recognition. Firstly, any novice soccer enthusiast would be most certainly oblivious to the team composition and team dynamics. Without any further context, this would make it impossible for the follower to identify the team players and the corresponding team making the attack or defense. In this case, the reader would be incognizant of the fact that the two players mentioned in the commentary are from Watford Football Club and they are making an attacking move against Chelsea Football Club. This task of identifying the team and players is referred to as "Named Entity Recognition" in the field of Natural Language Processing (NLP). It aims at recognizing and classifying the named entities in text into categories such as persons, organizations, locations, species, temporal expressions, quantitative values and other proper names [36]. A perfect named entity recognition is a necessary preprocessing step to derive further knowledge from the textual data. Handcrafted rules and supervised learning based named entity recognizer are both expensive in terms of system engineering cost and a large collection of annotated data respectively [36]. A soccer specific system will require developing a system with domain specific set of rules and a huge collection of manually tagged data set associating named entities to its particular category.

Resolving player references. Secondly, any beginner or new follower won't be aware of the nationality or the role of either of the player and be unable to spot out the Spaniard from his captian. But with the context and knowledge about the soccer domain, one can spot out that the Spaniard

is referred to *Deulofeu* and the captain is *Deeney*. This is termed as "Coreference Resolution" in the domain of Natural Language Processing. Coreference Resolution is the task of identifying and grouping all the mentions referring to the same entity into an equivalence class. This grouping of textual mentions is prerequisite to information extraction from a textual data [37].

These problems are challenging to human comprehension and also call into question the machine interpretation. The machine interpretation is more difficult in the sense that it involves processing natural language which is a growing field in itself. Coreference resolution methodologies in the NLP literature has primarily evolved on the corporas of newspaper and biomedical sciences [38]. The sports domain is relatively unexplored. The results of coreference resolution challenge in [38] showed that the state-of-the-art medical coreference resolution systems have difficulties in decoding references which require domain knowledge. Additionally, the expense of instigating a soccer specific system is attributed previously in terms of engineering and data collection cost. So, the attribution of player specific details, lack of domain knowledge and expert annotated data-set remains a challenge that would benefit from future research.

3.2.3 Proposed approach

In this section, we discuss about our proposed approach to extract information from soccer commentaries and analyze player performance. Our problem statement is to utilize the data set available from live commentaries, voice to text conversion, newspaper columns, internet blogs or any other source of textual information. There is an abundance of textual data available today. Moreover, analysis on textual data will also allow us to reveal facts about some historical players, games or events for which there is more opportunity of existence of textual documents rather than audio or visual sources. The sports industry is moving towards a phase of augmenting emphasis on visualizing and investigating the facts and figures. The analysis of soccer requires a system capable of involving and evaluating all the aspects of the game both on and off the football pitch. It should not just include the spatial orientation of player and ball, but should also include aspects of the game like player-history, player-fitness, player-form, player-skills, team dynamics, team-balance and team-coordination. These factors, though can be derived individually, but as an aggregation are

difficult to feed into the system. An expert analysis is necessary to provide this mixture in a brisk manner considering the best possible outcome for an event. The online sports commentaries (OSC) also referred to as minute-by-minute or play-by-play reporting, is an agglomeration of a sequence of real-time events by professional sportswriters and experts [39]. So, commentary analysis, unlike video scrutiny, can serve as the first-rate review of real-time game action. The text analysis can be economically viable for lower division clubs which are bound to the financial constraints.

In order to deduce knowledge from the commentaries, we need a system that can have the soccer domain knowledge. To realize such a system, we need to train it using annotated data set. As a preliminary task, our focus was to prepare the system that can differentiate and categorize different events into equivalence classes. We defined a set of our own categories that can be utilized to compute team statistics and access player performance. As described in section 3.2, the commentaries can have multiple interpretations and individual annotations can be biased. To overcome this limitation, we decided to employ crowd-labeling on the commentary data set. We created our own platform to enable annotation through crowdsourcing our data to the public.

With some initial data, we started building our model that can be trained with the soccer domain knowledge. We kept our model simple and applied a few variations of machine learning algorithms on our data set to get a better performance. The model would basically learn to classify the commentaries into corresponding categories.

Once the system with some initial domain knowledge was built up, we started exploring ways to tackle with the challenge of player attribution. The primary classification will impart the principal highlight of the commentary action. Comments were issued on a minute-by-minute basis. Consequently, most of them had multiple events in the same textual review and as a result, we had multiple tags from the annotated data. Moreover, a single commentary also covered multiple participants. Our aim is to establish a one-to-one mapping between players and our predefined classes. To simplify our approach, we fragmented this problem into three parts:

1. Classification of commentaries,

2. Resolving player references, and

3. Associating commentary categories to player.

Classification model depends on the quality of annotated data. However, determining player referrals, and instituting a link between classes and players are difficult tasks to accomplish with the an acceptable rate of performance. We have investigated different approaches to obtain the desirable outcome which we will be discussing in the following sections.

4. DATA COLLECTION : CURATING SOCCER COMMENTARIES

In this section, we describe our primary source for data-gathering, how we obtained it, why did we decide to employ crowd sourcing and how did we process the data for crowd-sourcing. There are a lot of APIs and websites that provide the basic statistics of soccer games like league, venue, fixtures, team details, coach, game-verdict, and scorers. Our main focus for data collection was live commentaries. We have several websites that can dispense play-by-play updates like BBC Sports [1], GOAL.com [2], WhoScored [3], SportsMole [4], TalkSport [5], ESPN UK [6] and many more. We decided to initially restrict our data set to GOAL.com for the two reasons:

1. We wanted to initiate our research with a single and reliable source, build our model on top of it and later exploit it to with different data sources, and

2. GOAL.com is largest online soccer publication in the world [40] and one of the largest soccer communities in the world with around 500 contributors in 50 countries with 18 language versions. GOAL.com started in 2004 under an European digital media production firm, GMS and was modeled as a part of experimentation in the area of online publishing features [41]. GOAL.com has a strong editorial team of 500+ members to deliver soccer expertise and unique insight to thousands of pieces of content every day to almost 66 million fans across 38 location based-editions, social channels, mobile, and interactive TV apps [7].

4.1 Scraping commentaries

GOAL.com keeps the record of all the major league matches in the form of match preview context, line-ups, game statistics, live commentary and match report. Since GOAL.com doesn't

[1] https://www.bbc.com/sport
[2] http://www.goal.com/en-us
[3] https://www.whoscored.com/
[4] https://www.sportsmole.co.uk/
[5] https://talksport.com/
[6] http://www.espn.co.uk/
[7] www.performgroup.com/brands/goal/

have any underlying API, we extracted data through scraping the website. The need for web scraping has been diminished with the proliferation of Web Services, however, there are situations when web scraping is useful are as follows:

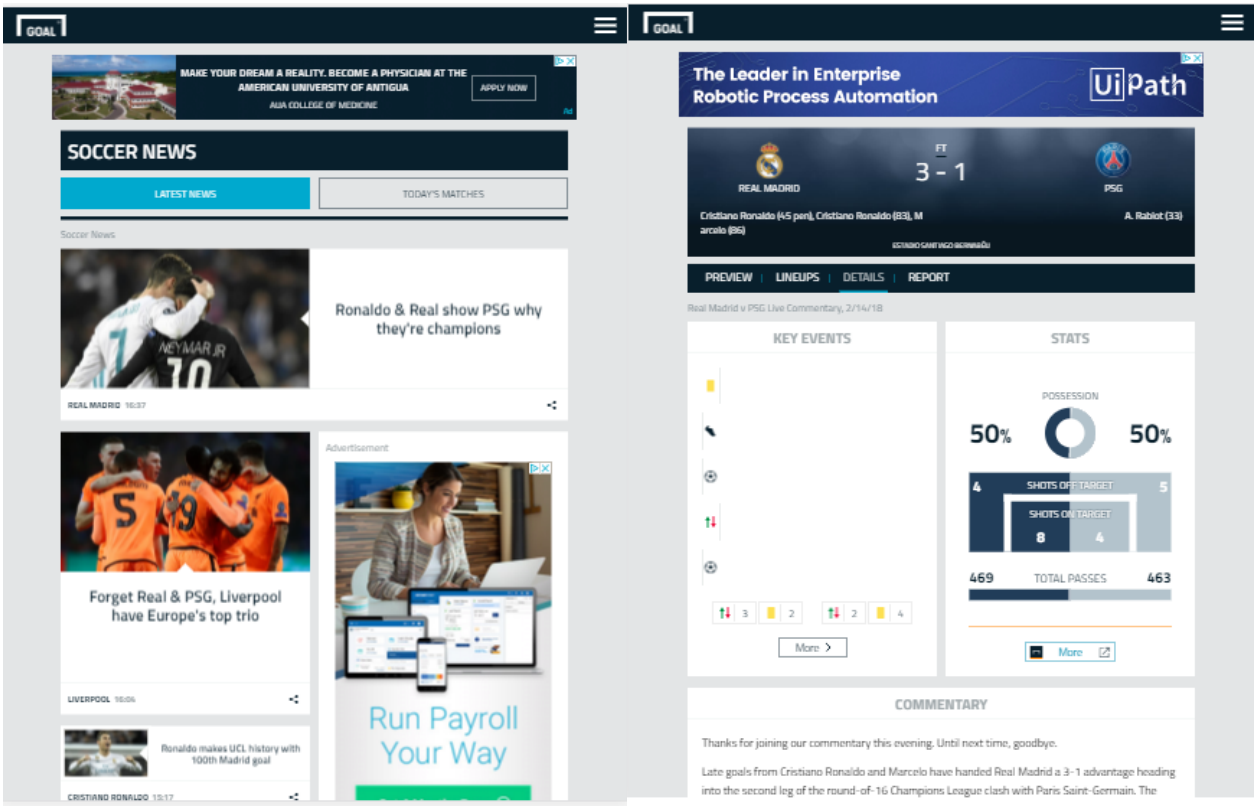
1. Independent web services with little scope for interoperability,
2. Restricted access to desired API services,
3. Operational cost of understanding API usage when such an investment is not justified, for example, during prototyping or source evaluation [42], and
4. Restriction on volume and rate of requests, unsuitable types and format of data available from the APIs [43].

Web Scraping is a method of gathering data from the Internet through any means other than a program interacting with an API or a human using a Web browser. This can be achieved generally by writing an automated program simulating human exploration of the Web that queries a web server, requests data and parses that data to extract required information [43, 44, 45]. There are different forms of scraping:

1. Screen Scraping - The output of a program is extracted as result for the end user instead for another program (usually for legacy applications with obsolete Input/Output Device or interface),
2. Web Scraping - Unstructured data from the web is extracted and processed into structured data to be stored in a database.

There are many ways to scrape the Web. This includes human-copy paste (feasible for small-scale projects), text grepping using regular expressions, HTTP programming, DOM parsing, HTML parsers, and making scraper sites (Websites created from scraping contents from other websites) [45]. [46] gave the perspective of HTML pages as containing two tokens - HTML Tag tokens and text tokens and represented HTML pages using a sequence of bits (0 - text, 1 - HTML tag). However, this approach was applicable to single body HTML documents only and would not be a viable

option for modern day multi-body HTML pages as it will take polynomial time for execution with a degree equal to number of bodies in the document. [47] used Document Object Model (DOM) tree for content extraction model by removing all the links from the page. But, this approach too is not usable for search engine websites like Google [8] and Bing [9] and multi-page websites. As shown in Figure 4.1, these two approaches cannot be used for scraping GOAL.com.



(a) (b)

Figure 4.1: An overview of GOAL.com website as adapted from [48]. (a) Homepage of GOAL.com shows navigation links to multiple web-pages. (b) *Real Madrid* vs *PSG* match page with multiple tab links to - *Preview*, *Lineups*, *Details* and *Reports*.

According to [49], there are three ways to parse the websites:

[8] www.google.com
[9] www.bing.com

Algorithm 1 Extracting Match Links using web-scraping

```
1:  procedure EXTRACTING MATCH LINKS
2:      matchDays ← A list of all the match days
3:      baseLink ← Link to GOAL.com matches section
4:      while matchDays is not empty do
5:          matchDay ← matchDays.pop()
6:          matchDayLink ← baseLink + matchDay
7:          page ← Fetch the link of the url using urllib2 libray
8:          soup ← Parse the pageusing HTML parser of Beautiful Soup
9:          competitions ← Find all competetion tags using class from soup
10:         for Each competiton in competitions do
11:             matches ← Find all match tags using class from competition
12:             for Each match in matches do
13:                 Extract matchlink from match and store it in a database match- table
14:             end for
15:         end for
16:     end while
17: end procedure
```

Algorithm 2 Extracting Commentary from match-links web-scraping

```
    procedure EXTRACTING COMMENTARIES
2:      matchLink ← Fetch Match links from database
        page ← Fetch the link of the match url using urllib2 libray
4:      soup ← Parse the pageusing HTML parser of Beautiful Soup
        commentaries ← Find all commentary tags using class from soup
6:      for Each commentaryTag in commentaries do
            Extract commentary, event, commentary-time, player-information from commentary and store it in a
8:      end for
    end procedure
```

1. Regular Expressions - It provides a fast option to scrape data, nevertheless, it is too fragile and will break easily as website gets updated.

2. Beautiful Soup - It is a popular Python module to pull out data from HTML and XML files [50]. It could correctly interpret the broken/invalid HTML tags and allows easy navigation to the elements. It is more verbose but easier to construct and understand. Regular Expressions are better in performance than Beautiful Soup, but are complex to implement.

3. Lxml - It is a python wrapper for C libraries *libxml2* and *libxslt* [51]. Like Beautiful Soup, it also parses invalid HTML and provides several options to navigate. On top of it, Lxml is much faster than Beautiful Soup but is difficult to install on some computers.

The factor of easy to install and use along with the past experience prompted us to use Beautiful Soup for web scraping. From GOAL.com, we scraped 122,748 commentaries from 35 game-days covering 1,320 matches of 9 major international soccer leagues - *'FA Cup'* [10], *'MLS'* [11], *'Liga MX'*[12], *'Bundesliga'*[13], *'Serie A'* [14], *'Premier League'* [15], *'Ligue 1'* [16], *'UEFA Champions League'*[17], *'La Liga Primera Division'* [18], and a separate international level matches category *'Friendlies'*.

Since most of the high profile matches occur over the weekends, we scraped the data of the matches over the weekend. GOAL.com provides with the feature of presenting all the fixtures over a particular date. So, we utilized this to get data for a particular historical match day.

[10]The Football Association Challenge Cup, is an annual knockout soccer championship in men's professional English football -http://www.thefa.com/about-football-association

[11] Major League Soccer is men's professional soccer league for teams in US and Canada - https://www.mlssoccer.com/

[12]Liga MX is Mexiican Professional Soccer League - http://www.ligabancomer.mx/

[13]Bundesliga is a German professional association soccer league - https://www.bundesliga.com/en/

[14]Serie A is professional football league competition for Italian http://www.legaseriea.it/en

[15]Premier League is English soccer league competition for top 20 clubs - https://www.premierleague.com/

[16]Ligue 1 is French professional league for men's association soccer clubs - http://www.ligue1.com/

[17]It is an annual prestigious club competition in Europe organized by Union of European Football Associations - http://www.uefa.com/uefachampionsleague/index.html

[18]La Liga Primera Division is soccer league for top Spanish clubs - http://www.laliga.es/en

4.2 Labeled actions

As we started scraping GOAL.com, we found that apart from commentary, we can also extract other preexisting valuable information like commentary-event, commentary-time, and player-information for selected events from GOAL.com.

4.2.1 Events and player information

GOAL.com provides information for 11 game-events in the form of pre-labelled tags for the commentaries. These events also encompasses information of the player(s) involved in it. At the initial stage, this data-set is a necessary part of developing our classifier. The 11 events are as follows :

Substitution. The event of substitution occurs when a player is replaced by a player from other team. The replacement can happen due to any of the reasons like injury or give rest to the substituted player or a strategy of game tactics. There are two players involved in the process : one leaving the field and another getting into the field.

> *Situation:* 79th minute substitution made by *Tottenham Hotspur* in their match against *Liverpool* on 4th February 2018.
>
> *Commentary:* “Wanyama comes on for Dembélé."
>
> *Players involved:* Victor Wanyama and Mousa Dembélé.

Yellow-card. A player is shown a yellow card by the referee to indicate that a player has been officially warned for a cautionable offense on or off the field. Following misdeeds results in a yellow card :

1. Unsporting or dirty behavior - Dissent by curse words or actions,
2. Consistently breaking rules,
3. Deliberate endeavors either to delay game play or to distract opponents, and

4. Any objectionable offense against the opponent player, a team-mate, a match-official or any other person or official [52].

For example:

Situation: 73rd minute action in EPL encounter between *Manchester City* and *Burnley* on 3rd February 2018.

Commentary: "Mee becomes the next Burnley player to go into the book, after he clattered into Silva on the half-way line. As I type, Fernadinho's clearance falls to Lennon 20 yards out, but his strike sails over."

Player involved: Ben Mee.

Goal. A goal is scored when a player either kicks or heads to score against opponent with the entire ball passing over the goal line between the goal posts and under the cross bars. This goal event, as defined in GOAL.com tags, does not include own-goals or penalty-goals as described in the next sub-sections.

Situation: 6th minute lead by *Arsenal* against *Everton* in their EPL game on 3rd February 2018.

Commentary: "GOOOOAAAAALLLLL!!! ARSENAL LEAD WITHIN SIX MINUTES! Well, Allardyce's game plan has already been proved wrong. Everton's cumbersome defence has not been able to deal with the hosts, and it was Williams and Mangala who were exposed then. Mkhitaryan found space down the right and delivered a low cross, with Ramsey on hand to tuck home."

Player involved: Aaron Ramsey.

Assist. An assist is accorded to a player when his/her pass results directly to a goal. The rules for assist are as follows:

1. Last pass to the goal-scorer,
2. Pass from second last holder of the ball provided it has direct influence on the outcome,
3. Player who is fouled and the goalscorer nets penalty or free-kick directly from the spot,
4. Rebound from shot on target by another player of the same team, and
5. No assist is awarded for goal scored by a solo run or dribble by the goal-scorer himself [53].

   Situation: The *Aaron Ramsey* goal described by 6th minute commentary in *Arsenal* against *Everton* EPL game on 3rd February 2018.

   Commentary: "An assist on his debut then for Mkhitaryan, who has just got back to make a brilliant challenge on Mangala and prevent the defender getting free in Arsenal's box."

   Player involved: Henrikh Mkhitaryan.

Penalty-goal. A foul/offense committed recklessly or carelessly or with excessive force inside the penalty area results in penalty-kick. The following actions within the penalty-box can lead to penalty :

1. Make or attempt to make contact with an opponent to gain ball possession,
2. Jumping or tripping on opponent,
3. Intentionally handling the ball, and
4. Any other offense leading to a yellow-card [54].

Penalty-goal occurs when a player scores penalty in the direct kick from the spot. Goal scored after the shot from player is saved by goalkeeper or rebounded from posts does not count as penalty-goal.

Situation: Extra-time penalty goal by *Tottenham Hotspur* against *Liverpool* at 90+5 minute on 4th February 2018.

Commentary: "GOALLLLLLLLL!!!! THIS TIME KANE CONVERTS! The striker comes forward with weight of the world on his shoulders. Kane fires low towards the bottom corner and it finds the net. He has his 100th Premier League goal."

Player involved: Harry Kane.

Own-goal. When a player accidentally kicks or heads the ball into his own-net while attempting to make a pass or clear the ball, the goal is considered as "own-goal". A shot deflected from the opponent player resulting in goal is not an own-goal given the shot is powerful enough to cause the goal [53].

Situation: 15th minute action between *Tottenham Hotspur* and *Southampton* in their EPL encounter on 21st January, 2018.

Commentary: "GOOOOAAAALLLL!! SOUTHAMPTON TAKE THE LEAD! Well, a huge slice of luck for the Saints, but they have got the goal that their positive start has deserved. Tadic and Bertrand combined brilliantly down the left - the latter thumping in a low cross that Sanchez could only prod into his own net."

Player involved: Davinson Sánchez.

Yellow-red. A player, who has been shown red-card, is sent-off the field and is not allowed to take part in the remaining game. There are two ways in which a player receives 'red card' :

1. **Yellow-Red** : When he/she is officially warned with second caution. This could result due to another offensive foul leading to one more yellow-card which is equivalent to showing a red-card, and
2. **Red-card** : When he/she is involved in a more serious offense than yellow-card eligible foul [54].

Situation: Double yellow card shown to *Chelsea* mid-fielder on 30th minute during their EPL clash with *Watford* on 5th February, 2018.

Commentary: "RED! A horrible showing meets a premature end for Bakayoko. He is off. A poor first touch allows Richarlison to nip in and a rash challenge brings a second yellow card in quick succession."

Player involved: Tiémoué Bakayoko.

Red-card. It is awarded for the gravest offenses. It is severest punishment in the soccer and can result in suspension of player for several games too. The player on red card has to leave the game immediately. As described above, a red card is shown for a serious offense as listed below :

1. Injury-prone foul play like two-footed slide tackle, savage tackle from behind that can be injurious to opponent, and long-foot tackles that could hit a players,
2. Violent unsporting conduct,
3. Deny goal or obvious goal scoring opportunity using unfair means like Handling the ball, and
4. Using threatening, insulting, foul language against opponent player, team player, referee, fans, or any other game official [52, 53].

Situation: Red card was shown to *Paul Pogba* for directly tagging his studs into *Belerin* at 76th minute during EPL match between *Arsenal* and *Manchester United* on 2nd December, 2017.

Commentary: "RED CARD! Pogba gets his marching orders! The former most expensive man in the world comes in with his studs showing on Bellerin and it's a deserved dismissal. Pogba ironically applauds and may well be punished further for that."

Player involved: Paul Pogba.

Penalty-save. An event is categorized as 'penalty-save' when penalty-kick is either punched or kicked or headed away or caught by the opponent goalkeeper. Rebound from goal-posts does not count as penalty save.

> *Situation:* 87th minute penalty awarded to *Tottenham Hotspur* is saved by *Liverpool* goalkeeper during their EPL clash on 4th February, 2018.
>
> *Commentary:* "MISS! KANE STRIKES THE BALL STRAIGHT AT KARIUS! The forward misses the opportunity to put the visitors ahead. He sends his effort straight down the middle and the Liverpool keeper makes the stop."
>
> *Player involved:* Loris Karius.

Missed-penalty. A penalty kick is missed by the player when he fails to score the goal from penalty-spot. The penalty-miss can occur either due to a brilliant save by opponent goalkeeper or mistake by the player taking penalty-kick for accidentally or deliberately shooting it outside. The missed-penalty event is attributed to the player shooting the penalty-kick.

> *Situation:* 75th minute penalty-kick missed by *Manchester City* striker during their EPL game with *Tottenham Hotspur* on 16th December, 2017.
>
> *Commentary:* "MISS! De Bruyne leaves the penalty to Jesus, who steps up and fires a thumping strike into the near post. The ball then flies out to Sterling, who tries to find the back of the net with his first-time follow-up but ends up sending it over the top of the crossbar instead."
>
> *Player involved:* Gabriel Jesus.

This tags as obtained from GOAL.com are annotated by experts. We initially used this pre-defined labels to gather some key insights into the commentary and utilize it to train our classifier.

4.3 Player details database

For player identification from the commentaries, our primary requirement was to establish a database with all the necessary information related to the players. We utilized SoFIFA.com [19] for collecting all the player details through web-scraping using Beautiful Soup. SoFIFA started as an online scouting tool for FIFA series career mode and later on evolved with more features growing community support. SoFIFA provided player information at individual level, club level, and international level. For this, we targeted all the major leagues involved in the commentary extraction and created a list of all the teams playing in those leagues.

Once, we identified the teams, we extracted player information and stored it in MongoDB[20]. MongoDB is document database which stores data in flexible, unrestricted and adaptable JSON-like documents and supports real time aggregation of data, ad hoc queries, and indexing. It is free and open source and supports text-search on the content of the fields. This functionality is very beneficial in establishing the player and team relationships. For each player, we stored his name, age, nationality, current playing position,all playing positions, jersey number, and his current club.

We collected data of around 9546 players of 141 different countries playing across 15 major leagues and international competitions. We encountered a few challenges in generating this player database :

1. The player informations are not easily available on the internet. Most of the website offer paid APIs to access the limited amount of data. Others had out-dated data from a few years back.

2. Player database was relatively latest as compared to commentaries. So, a few of the teams from commentary matches were either relegated to a lower division or were dissolved or disowned by their owners.

[19]https://sofifa.com/players/top
[20]https://www.mongodb.com/

3. Latest player data represents the current club of the player. However, our commentaries were relatively old and in the meanwhile, we encountered some players moved or transfered to another clubs.

4. GOAL.com uses different naming conventions for the team and player name. We had to figure out the way to relate both of them.

4.4 Additional labels

The pre-defined labels as we obtained from GOAL.com would only provide the basic statistics of a soccer game. Our research aims at analyzing those features which are important but does not contribute directly to any of this numbers. In soccer, the performance of a player is defined by figures and statistics. However, many times, the player involved in creating a goal or crucial defense of the time go unnoticed. For instances, in the English Premier League title winning run of *Chelsea* in 2016-17 and *Leicester City* in 2015-16, the role of *N'Golo Kanté* was very vital in both of them. *N'Golo Kanté* is a a french professional soccer player who currently plays for English Premier League Club *Chelsea* and French national team. He started as professional player with French League 2 club *Boulogne* and then moved League 1 club *Caen*. In 2015, he joined EPL club *Leicester City* and later joined *Chelsea* in 2016. He was awarded Professional Footballers' Association Players' Player of the Year award and Football Writers' Association Footballer of the Year for 2016-17. He is a crucial defensive midfielder and makes a lot of good runs, interceptions, and tackles. He has an amazing work-rate (The pace at which the player makes run chasing the ball while not in possession) and was praised by the managers of both the title-winning clubs for his importance in the team. He keeps a balance between attack and defense by launching or maintaining an attacking flow from the center and cover up for the defense as opponents push forward respectively. His innate ability to make sudden changes in his direction while dribbling through the opponent, to subtly change his speed when closing on adversary and ability to sense the game and act effectively upon it makes him one of the best soccer midfielder. In his 74 matches at *Chelsea*, he has scored just 3 goals and provided 1 assist. Though the numbers don't contribute

much towards his success as a player, the deeper game analysis would obviously give us a better statistics for a player of his caliber. Another example for such a player is Croatian Professional Soccer team's captain and *Real Madrid* central midfielder - *Luka Modrić*. He has just 9 and 12 goals, 23 and 17 assists in his 158 and 103 appearances for his club and country respectively. However, he is regarded by many as world's best midfielder. His previous manager *Carlo Ancelotti* described him as -

> *"**Luka Modrić** is definitely one of best midfield players in the world. He has great technical abilities, reads the game and has a strong personality that he has built over the years. Besides that, he is a very pleasant person".*

It is for players like *N'Golo Kanté* and *Luka Modrić*, we started looking for insights to generate a better statistics for their evaluation.

We added our 9 additional custom labels. A single commentary can include these multiple labels and therefore, we have more than one player in our player- information which will be different from those available at GOAL.com.

Chance. We defined **Chance** as an event which either results in creating an opportunity for a player to score goal or directly results in a goal. This include situations when player :

1. Hits a wonderful cross towards the penalty area intended for another player to score,
2. Has an opportunity to score from the free-kick or penalty-kick,
3. Dribbles through opponent's defense in order to score or create an opportunity to score, and
4. Plays a pass or through ball to another player who scores or has better chance to score.

The event-commentaries that are part of **Chance** label includes "goal", "assist", "own-goal", "penalty-goal", and any other comments with situations as described above.

> *Situation:* 69th minute commentary from *Tottenham Hotspur* vs *Manchester United* match in English Premier League (EPL) on 31st January, 2018.

Commentary: “CHANCE! Eriksen releases Son into the right inside channel and he has Kane screaming for the ball six yards out in the middle. However, he goes for goal himself, blasting his strike towards the near post and De Gea makes the stop.”

Player involved: We include both *Christian Eriksen* and *Son Heung-min* because the former provides a good pass to latter who also has opportunity to create another chance for *Harry Kane* or go for goal by himself.

Block. We include this label to add statistics for a good defensive performance. **Block** is defined as an intercepting or ball-clearing event which either momentarily or completely terminates that particular attacking move from the opponent. "Block" comprises of following scenarios :

1. Clearing a ball from incoming cross intended for the opponent player with an opportunity to score,
2. Obstructing a shot from opponent, and
3. Intercepting a through ball or pass leading to an opportunity for opponent to score.

Situation: 12th minute vital block by *Manchester United* defender during their EPL game against *Tottenham Hotspur* on 31st January, 2018.

Commentary: “BLOCK! The ball bounces kindly for Alli in the box and he goes for a strike, but Jones puts his body on the line and diverts it out for a corner.”

Player involved: Phil Jones.

Save. This class is intended to analyze the goal-keeper’s capabilities. Only goal-keeper will be part of the player information corresponding to this label. A commentary is classified as **save** when the goal-keeper prevents the opponent from either netting a goal or creating a chance. The events for "save" are when goal-keeper :

1. Punches or kicks or heads away the kick/pass/cross aimed towards creating a chance for or scoring goal, and

2. Handles the ball safely to prevent opponent from any opportunities.

Save includes successful protection by goalkeeper against the opponent invasion through penalty-kicks, free-kicks, long-range shots, headers, tappings, deflections, and accidental self or team-mate's mistake.

> *Situation:* 30th minute save from *Manchester United* goalkeeper during their EPL encounter with *Tottenham Hotspur* on 31st January, 2018.
>
> *Commentary:* "SAVE! Martial surges into the right inside channel and past the challenge of Sanchez. He fires towards the bottom corner, but Lloris is there to make a solid stop."
>
> *Player involved:* Hugo Lloris.

Foul. Foul is the criteria to assess the offensive or disruptive attributes of a player. It includes any activities resulting in yellow-card or red-card, handling the ball (players other than goal-keeper) and tackles or acts which are malicious or unsporting in nature. Depending upon the position and extent of foul, it can result in free-kick, penalty-kick or halting the play (in case of serious injury). The following incidences which can result in "foul":

1. Obstructing an opponent by holding the player or pulling the player's outfit when neither of them has ball-possession,
2. Kicking or pushing team-mate,
3. Deliberately delaying the start of play,
4. Arguing or disrespecting referee and other match officials, and
5. Involving in any activity that results in yellow-card or red-card.

> *Situation:* 85th minute foul in the EPL match between *Manchester City* and *West Bromwich Albion* on 31st January, 2018.

Commentary: "Ouch! Diaz breaks down the left wing on a marauding run before being clattered into by a high challenge from Phillips. The City players and fans all scream for a red card, but the referee only decides to book the winger. That's a big call, and it's one that Guardiola isn't at all happy about."

Player involved: Matt Phillips.

Block*(Corner). This is an additional label to identify corners. Corner occurs when ball crosses goal line, except when a goal is scored, with the last touch of defending team's player. The event leading to a corner is defined under **Block*(Corner)** label. However, corner-kick is considered as an opportunity to score goal and comes under **Chance** label. A player concedes corner when :

1. A shot from opponent is deflected from his body and goes across goal line,
2. He, as a goalkeeper, makes a save and/or accidentally allow ball to go through behind the goal line, and
3. The player either deliberately or unintentionally tackles/kicks/heads/chest-out ball outside the goal-line.

Corner is an interesting event in soccer as the event just before a corner can be a good defensive act and the one just after it can be a good attacking move.

Situation: 12th minute block by *Jones* leads to corner for *Tottenham Hotspur* during their EPL match with *Manchester United* on 31st January, 2018.

Commentary: "BLOCK! The ball bounces kindly for Alli in the box and he goes for a strike, but Jones puts his body on the line and diverts it out for a corner."

Player involved: Phil Jones.

Chance-missed. It is an important category to access the chance-converting ratio of players, especially strikers. Chance-conversion ratio is also applicable at club level as shown in Table 4.1.

Club	Shots At Goal	Goals	Ratio
Manchester City	382	79	20.7%
Manchester United	259	51	19.7%
Liverpool	342	61	17.8%
Chelsea	308	49	15.9%
Tottenham Hotspur	325	52	16.0%

Table 4.1: Conversion ratio of top 5 clubs in English Premier League as of 12 February, 2018 adapted from Transfermarkt statistics [55].

Player	Goals	Shots	Shots on Target	Shooting Accuracy	Shots per goal
Harry Kane	23	150	62	41%	6.52
Mohamed Salah	22	103	49	48%	4.68
Sergio Agüero	21	88	39	44%	4.19
Raheem Sterling	15	63	27	43%	4.2
Jamie Vardy	13	45	22	49%	3.46

Table 4.2: Shooting accuracy and shots per goal ratio of top 5 goal-scorers of English Premier League as of 12 February, 2018 adapted from PremierLeague statistics [56].

Due to lack of opportunity, certain players are not able to display their best talent om the pitch. This attribute is vital to provide statistics for such players. As we can observe from Table 4.2, though scoring the least number of goal among all, *Jamie Vardy* has the best shooting accuracy and conversion ratio too.

A player missing a chance is defined by any of the following scenarios :

1. Deserting an easy goal-scoring opportunity by shooting or heading ball away from the goal,
2. Failing to convert a free-kick or penalty-kick into a goal, and
3. Loosing out possession to an opponent during an attacking move.

Situation: 48th minute commentary from the EPL game between *Manchester City* and *West Bromwich Albion* on 31st January, 2018.

Commentary: "What a chance for Sterling to make it 2-0! Gundogan bursts through the middle of the pitch on a great run before picking out Sterling in a pocket of space in front of goal. He easily gets the better of Dawson before shooting, but he somehow fires his effort wide of the far post. He really should have buried that."

Player involved: Raheem Sterling.

Tackle. A tackle is when a player dispossesses an opponent through a challenge which does not result in a foul. This attribute defines the defensive capabilities of a player. There are three types of tackles:

1. **Block Tackle**: The player comes in path of the opponent and dispossess him by either blocking his shot or intervening in between to change the direction of ball and holding up the ball,
2. **Poke Tackle**: A single foot of the player is used to prod the ball away from the opponent, and
3. **Sliding Tackle** : It is the last resort slide by a defender to dispossess the opponent and deny any goal-scoring chance.

Generally, "Block Tackle" comes under "Block" label, but sometimes it can get ambiguous to differentiate among these 3 types of tackles from the commentary itself.

Situation: 34th minute tackle from the EPL game between *Chelsea* and *AFC Bournemouth* on 31st January, 2018.

Commentary: "TACKLE! Steve Cook makes an incredible last-ditch tackle on Hazard to deny him a shooting opportunity one-on-one with Begovic inside the box! Great work from the centre-back."

Player involved: Steve Cook.

Mistake. A mistake is a clear fault of a player in losing ball possession or any other personal errors/blunder resulting in creation of chance for opponents. There are no clearly defined scenarios for "mistake" label. It can vary from wide range of events like falling out on ball, communication-gap among team-mates, dribbling when passing is better option, unnecessary tackling and many more.

> *Situation:* 30th minute red-card commentary from the EPL encounter between *Chelsea* and *Watford* on 5th February, 2018.
>
> *Commentary:* "RED! A horrible showing meets a premature end for Bakayoko. He is off. A poor first touch allows Richarlison to nip in and a rash challenge brings a second yellow card in quick succession."
>
> *Player involved:* Tiémoué Bakayoko.

4.4.1 Information

In the game of 90 minutes, not every minute results in an action. There are times when a team just maintains ball possession prior to building their attack or a substitution occurs or an injured player is provided with treatment. For our classifier, keeping individual labels for these events is not necessary as we are currently assessing the attacking and defensive attributes of a player. However, these events are important and at later stage would be very significant to invest into deeper analysis using these details for player evaluation. In general, we classify those events as **information** which do not remotely relate to any of the labels mentioned earlier. Scenarios which are categorized as "information" are :

1. Event of substitution,
2. General announcement of start or end of playing half,
3. Description of ball possession by a team with no player specifications ,
4. Any kind of statistics related to player or team form or coach performance, and

5. Offside event.

   Situation: Commentators describing the atmosphere of stadium in 44th minute of EPL game between *Chelsea* and *AFC Bournemouth* on 5th February, 2018.

   Commentary: "You're getting sacked in the morning,' sing the Watford fans to Conte. Chelsea are struggling to regain a foothold in the game."

   Player involved: None.

These labels are necessary for our research. However, tagging commentary with these labels is quite a challenging task. We turned to crowd-sourcing to annotate a huge dataset as discussed in the next section.

4.4.2 Time

The underlying definition of commentary specifies the importance of play-by-play updates. To break down the game into these notifications, the commentaries are generally published on a minute-to-minute basis. Broadly, a single commentary summarizes the game-activity of entire one minute. Time is an important parameter to evaluate a lot of soccer statistics :

1. **Scoring Frequency**: Nowadays, evaluation of a striker is based on how frequently he scores in the match. This assessment requires tracking of minutes player by each player and time-stamp of each goal scored. As can be observed from Table 4.3, except *Álvaro Morata* and *Wayne Rooney*, the scoring frequency of top 10 goal scoring players in all competitons is greater than that in English Premier League (EPL). From this data, we can say that game-standards and quality of English Premier League clubs has improved and it has become more difficult to score goal in an EPL match than earlier.

2. **Goal Statistics**: We can also explore the highest probabilistic goal-scoring intervals for the opposing team. Table 4.4 shows goals scored and conceded by top 6 teams of the English Premier League at different intervals of the game in season 2017-18 as of 15 February, 2018.

Name	EPL minutes	EPL minutes per goal	All competitions minutes	All competitions minutes per goal
Harry Kane	2228	97	2735	78
Mohamed Salah	2036	93	2614	84
Sergio Aguero	1788	85	2204	73
Raheem Sterling	1922	128	2297	109
Jamie Vardy	2265	174	2268	162
Roberto Firmino	1954	163	2530	115
Romelu Lukaku	2251	188	2775	139
Eden Hazard	1617	147	2056	137
Álvaro Morata	1551	155	1904	159
Wayne Rooney	1638	164	1886	171

Table 4.3: Top 10 scorers of English Premier League as of 15 February, 2018 for the season 2017-18 adapted from WhoScored.com statistics [57].

We can notice that these clubs, especially *Manchester City*, are aggressively attacking in the last 15 minutes and this statistics can aid opponents/rivals to strategize their game-plan accordingly.

Club	0-15 min	16-30 min	31-45 min	46-60 min	61-75 min	76-90 min
Manchester City	6 - 3	12 - 1	10 - 5	14 - 2	14 - 4	23 - 5
Manchester United	5 - 3	5 - 3	11 - 4	9 - 3	6 - 2	15 - 4
Liverpool	5 - 4	14 - 2	8 - 7	11 - 7	12 - 1	11 - 10
Chelsea	6 - 3	10 - 1	6 - 5	6 - 2	9 - 6	12 - 6
Tottenham	6 - 6	9 - 2	6 - 3	12 - 0	8 - 4	11 - 9
Arsenal	8 - 5	7 - 5	7 - 2	10 - 7	10 - 9	9 - 8

Table 4.4: Goal scored and goal-conceded distribution of top 6 teams in English Premier League 2017-18 as of 15 February, 2018 adapted from SoccerSTATS statistics [58]. x - y in table corresponds to x = goals scored and, y = goals conceded.

3. **Clean Sheet Rate**: Analogous to goal count for a striker, an important characteristic for a goalkeeper is clean sheet. A goalkeeper or team's defense is attributed a "clean sheet" if they

prohibit their opponents from netting any goals during an entire match. However, maintaining a clean sheet is not the best criteria to assess goalkeeper or defense players. From table 4.5, we can see that even though *Thibaut Courtois* has more clean sheets than *Ederson*, but average number of minutes before conceding a goal (Minutes per goal conceded) is greater for latter.

Player	Matches	Clean-sheets	Minutes-played	Minutes per goal conceded
David de Gea	27	15	2430	128
Thibaut Courtois	26	13	2340	102
Ederson	27	12	2385	119
Hugo Lloris	25	11	2250	98
Nick Pope	24	10	2125	106

Table 4.5: Clean sheet statistics for top goalkeepers in English Premier League 2017-18 as of 15 February 2018 adapted from WhoScored.com statistics [57].

4.5 Annotating commentaries

Since many of our commentaries lack labels of specific actions and of which players are engaged with the action, we turn to a crowdsourcing approach to annotate commentaries. Crowdsourcing is the process of outsourcing a particular task to a large community of people with shared interest or open mass rather than local employees or closed organization. It has been successfully employed by companies like Threadless, iStockPhoto, InnoCentive, the GoldCorp Challenge and many more [59]. [60] defines crowdsourcing into 2 categories – *Explicit* where the crowd of users explicitly contribute towards the problem at hand, and *Implicit* where users collaborate to the crowd sourced task as a side effect of another task. According to study in [61], crowd sourcing from users can outperform the local employees in generating new ideas for a company. Based on the type of problem, [62] classifies 4 general approaches to crowdsourcing :

1. **Knowledge Discovery and Management** : It consists of a common platform for the online

community to share information related to the problem in a prescribed format which can be used as a general asset.

2. **Distributed Human Intelligence Tasking** : This approach is ideal when dataset is large and we require a cheap way of analyzing it. An organization or employer usually break down the large dataset into microtasks and post it on a common portal accessible by intended crowd.

3. **Broadcast Search** : This is generally applicable for seeking possible solutions for a scientific problems with a provable "right" answer. The problem statement is announced on an online community often with an incentive and/or prize money.

4. **Peer-Vetted Creative Production** : It is suitable for problems which does not have provable "right" answer but requires an aesthetic opinion or public support. Typically, in this category, organization or individual seeks an attractive, innovative idea (generally complemented with some reward) or public choice from the available options for a brief problem statement.

Of these 4, our problem statement requires *Distributed Human Intelligence Tasking* approach for crowdsourcing. To break down this into microtasks, we decided to allow the annotation of a single commentary at a time. We introduced these tasks through a simple web-application as discussed below.

We are motivated to use crowdsourcing because we require opinion of individuals for commentary analysis to remove the biasing effect. With crowdsourcing, we can attract users with common interest, skill-set or expertise to annotate our data set. Also, it allows them the opportunity to earn money, gain recognition from peers, and develop new skills.

4.5.1 Application design

We designed a simple web-application with *AngularJS*[21] front end and a *Django*[22] API providing with random commentaries as back end. In order to develop the back end API, we initially scrapped commentaries from GOAL.com and stored it in *MongoDB*[23](A NO SQL database with

[21]https://angular.io/
[22]https://www.djangoproject.com/
[23]https://www.mongodb.com/

stores the data in JSON like documents with schemas). Using this as our database, we developed our *Django*[22] application which would randomly fetch the commentaries from *MongoDB*[23] and store the user's response back to it. As of January 2018[24], the percentage of mobile phone users is

(a) (b)

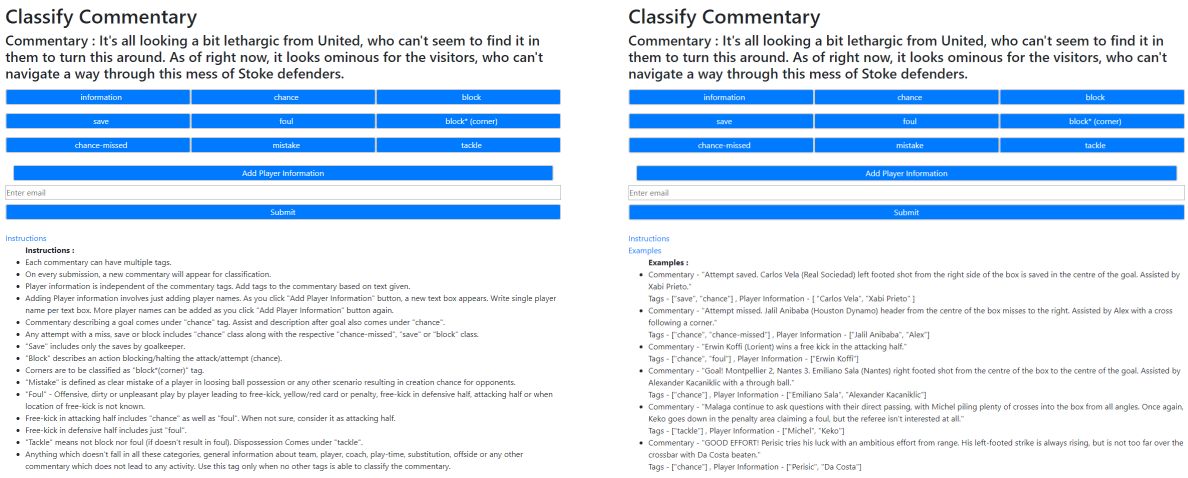
Figure 4.2: CrowdSourcing Application overview. (a)Desktop Display. (b)Mobile Display.

around 52% followed by desktop or laptop users at 44% and tablet users at 4%. According to [63], by 2013, around 90% of people were using multiple screens and mobile users were 5 times more likely to reject a website if it was not mobile-friendly. To allow more attention and flexibility of the users, we implemented web-responsive design. Figure 4.2 shows mobile and desktop presentation of our crowdsourcing application.

4.5.2 Labelling and player information

Since the application is open to crowdlabelling, we need to make sure that the user understands all the "labels" before he/she can proceed with annotation. For this, we added two additional divisions in our application - *Instructions*, and *Examples* as shown in figure 4.3a and 4.3b respectively. As explained earlier, we can have multiple labels in a single commentary. So, we provided buttons

[24]http://gs.statcounter.com/platform-market-share/desktop-mobile-tablet/worldwide/#monthly-201701-201801



(a) (b)

Figure 4.3: CrowdSourcing Application overview. (a)Desktop Display. (b)Mobile Display.

underlying check-box to support more than one selections per commentary. The check-box allows user to make a binary choice for each label i.e., either a label is suitable for the commentary or not. We have put a validation such that for each commentary, user needs to classify at least one label in order to move forward with the next commentary.

In addition to multi-labelling, a user can also add player information. We have added a "Add Player Information" button for filling out the name of player-involved in the commentary. With each click of this button, a new input box appears that asks for player name. A commentary can contain description pertaining to multiple players varying from zero to six (the most number of players in a commentary encountered till now). So, it is not mandatory for user to fill out player information and we have not added any check on it.

Example 1

> *Situation:* 52nd minute commentary of save by *Thibaut Courtois* on one-on-one opportunity to *Alexandre Lacazette* during English Premier League London derby clash between *Arsenal* and *Chelsea* on 3rd January, 2018.
>
> *Commentary:* “SAVE! Some excellent interplay and a stroke of fortune see Lacazette

with space to shoot on the left, but Courtois is out quickly to close down the space and block the Frenchman's fierce effort!"

Discussion: With proper comprehension of the commentary, we understand that *Alexandre Lacazette* had an opportunity to score goal, so we can say that he had a "Chance". However, *the Frenchman's fiere effort* (in this case *Alexandre Lacazette*'s effort) have been saved by *Courtois*, so there was a "Save" too. From this we can say that, two players were involved in this event - first creating or converting a "Chance", and second - Obstructing the goal-scoring chance for the first player by an excellent "Save".

Annotations: Chance, Save.

Players involved: Player 1 - *Lacazette*, Player 2 - *Courtois*.

Example 2

Situation: Commentary corresponding to the crucial miss at 70th minute by *Álvaro Morata* during EPL London derby clash between *Arsenal* and *Chelsea* on 3rd January, 2018.

Commentary: "MISS! Another glaring miss! Morata is played through again and Chambers tracks him as best he can, but he's the wrong side of the Spaniard. Morata tries a dinked effort at the near post, but it's into the side-netting! Chelsea have been so wasteful in front of goal."

Discussion: Morata has single handedly created a chance for himself after gliding past *Chambers*. He tries to take a dropping shot at the near post but it goes into side-netting and he ends up wasting the opportunity to score. So, only single player here is responsible for both producing the opportunity and missing out on it. Even though both the tags are attributed to *Morata*, we still include *Chambers* in our player-information because we thought this will give us a better insight to understand involvement of other

player in the game and also helps in training our model for player identification in the commentary.

Annotations: Chance, Chance-missed.

Players involved: Player 1 - *Morata*, Player 2 - *Chambers*.

Example 3

Situation: A needless action from frustration by *Mesut Özil* gets him booked. Following is commentary describing 67th minute event from EPL London derby match between *Arsenal* and *Chelsea* on 3rd January, 2018.

Commentary: "Ozil is booked for kicking the ball away after the penalty was awarded."

Discussion: The commentary is straight forward and it explains the reaction of *Ozil* was merely out of frustration. There was no need for such action which attributes it to a "mistake" and the following yellow card as per our labels would be a "foul".

Annotations: Mistake, Foul.

Players involved: Ozil.

Example 4

Situation: A general commentary of the game at 8th minute of the EPL game between *Tottenham Hotspur* and *Manchester United* on 31st January, 2018.

Commentary: "United have found acres of space in the Tottenham half, which will concern Pochettino in the early stages of the contest. Spurs are ahead, but are lacking control of the game."

Discussion: It just explains the ball possession with *Manchester United* in the opponent's half. This statement is very general and lacks any specific information about player or either team's attacking or defensive move. It is just an informative remark.

Annotations: Information.

Players involved: None.

Example 5

Situation: Description of 33rd minute corner by *David Silva* in the EPL game between *Manchester City* and *Newcastle United* on 20th January, 2018.

Commentary: "Silva, who has seen plenty of the ball so far, causes problems over on the left channel once again before sending a deflected cross behind for a corner. It's whipped in by the Spaniard and flies towards the near post, where it fails to find a team-mate and is easily cleared by Joselu."

Discussion: Silva has been a constant threat to the opponent and now, he has won a corner. The resulting corner kick is an opportunity for *Manchester City* to attack. It is taken up by *the Spaniard*(referring to *Silva* only) and is blocked by *Joselu* before it can lead to any further "chance" or the opponent. In brief, both the "chances" to cross and the consequent corner by *David Silva* has been blocked out by opponents - first one out of a corner and the second one through a clearance by *Joselu*.

Annotations: chance", "block* (corner), block.

Players involved: Player 1 - *Silva*, Player 2 - *Joselu*.

4.5.3 User validation

Crowdsourcing applications suffer from the drawback of inefficient workers who submit invalid or low quality work to get the incentive with reduced efforts [64]. The results of single dice

experiments in [65] showed that the amount of financial gain or incentive is not related to the degree of dishonesty in the users (workers). This means that we cannot entirely eliminate these users, however, we can put a check on their performance over the time and valuate them accordingly. For this, we have implemented the Majority Decision approach discussed in [64] with some modifications of our own to adjust our domain and reduce verification efforts.

In the Majority Decision based model, we randomly assign a few microtasks to workers and repeat those microtasks until we find a majority vote from the workers' responses. In our app, at each hit, we provide one random commentary from 300 commentaries (microtasks) to a worker. Once we get a majority vote for this microtask, we remove this commentary from the lot of 300 commentaries and add a new one from our database. In [64]'s model, only the workers who voted according to the majority vote are incentivized. However, we decided to pay every worker for their contribution as commentary can sometimes be very ambiguous as discussed in 3.2. On top of this, we also enforced a gold-standard check to verify the reliability of a user. We added some straight forward commentaries as part of our gold-standard check and we randomly added this in the commentary sequence for a worker. In every 10 microtasks, we had 3 gold-standard commentaries at random positions so that the worker's activity is regularly verified.In every 10 microtasks, we had 3 gold-standard commentaries at random positions so that the worker's activity is regularly verified. We kept track of correctness of user's annotation using this verification.

We collected the initial 11,813 commentaries annotated by 102 different users out of which only 31 went on to annotate the data corresponding to 30.3% of rater acceptance rate. The partial correctness score obtained from the goal-standard commentaries is averaged at 89.69%. Figure 4.4 shows that almost one third (32%) of users who do take up this task, annotate upto only 10 commentaries before they move away from this task. Only 13% of users maintained to continue with this task surpassing 1000 commentaries of data annotated. In these 11,813 commentaries, total 21,505 labels were annotated with *chance*, *information* and *foul* contributing 32%, 18% and 16% respectively. Figure 4.5 reflects the distribution of these 3 labels across various amount of

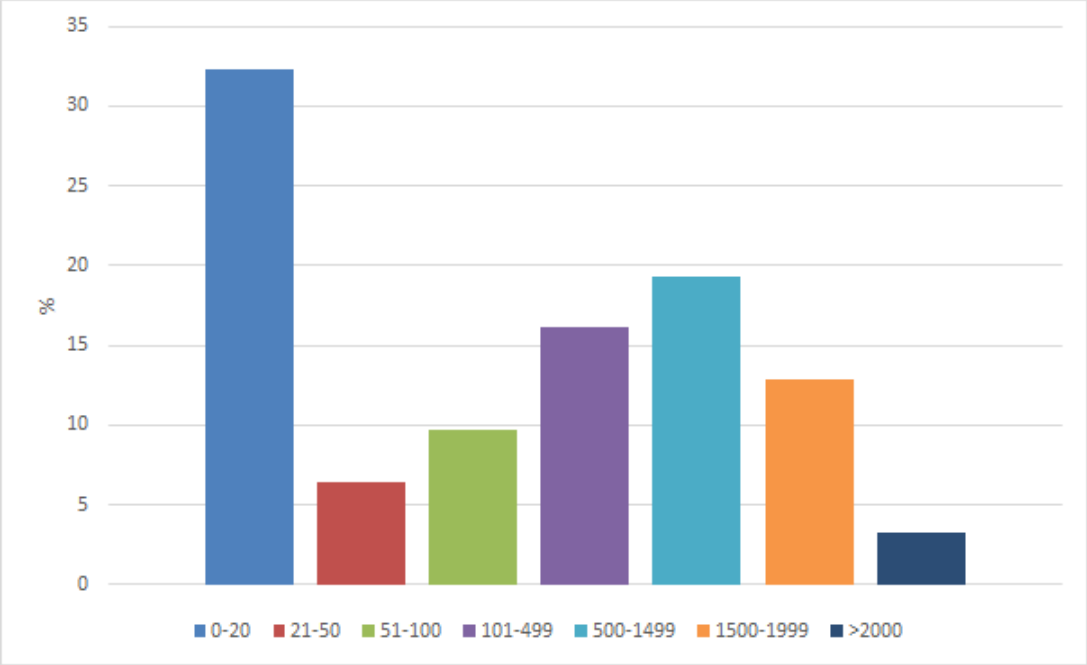


Figure 4.4: Percentage of commentaries annotated by users corresponding to the total commentaries annotated in the category of number of commentaries annotated.

commentaries annotated.Figure 4.6 and Figure 4.7 represents the similar distribution for moderately and least annotated labels. From figure 4.5, we can notice the gradual increase in the rater's preference for *chance* label and a gradual decrease for *Information* label as he/she annotates more commentaries. This can be attributed to rater's better understanding of commentary with more annotations. A sudden unusual bump for the rater's *Below 50* category can be due to the randomness of data being assigned. Figure 4.6 shows similar distribution for *Block* (corner)*, *Chance-missed*, and *Block* tags. Figure 4.7 does not show any such patterns. This might be due to the scarcity of *Tackle*, *Save*, and *Mistake* annotation type commentaries. We expect to more relationship between the labels and number of commentaries annotated by rater's as we gather more data.

4.5.4 Inter-judge agreement

In informational retrieval based research, a gold standard data set is very important to compare the performance and quality of the system. Generally, machine learning and IR based applications require a large, semantically annotated data and building them from the scratch is very time and

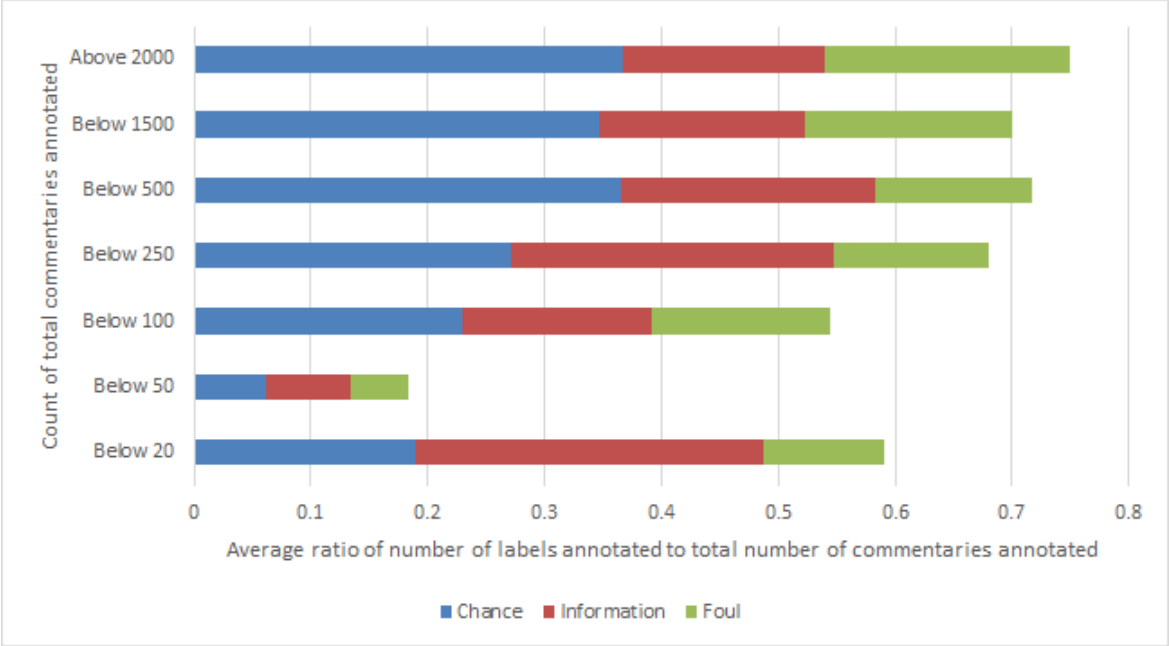


Figure 4.5: Distribution of top 3 annotated labels - *Chance*,*Information*, and *Foul* across the total count of commentaries annotated by different raters.

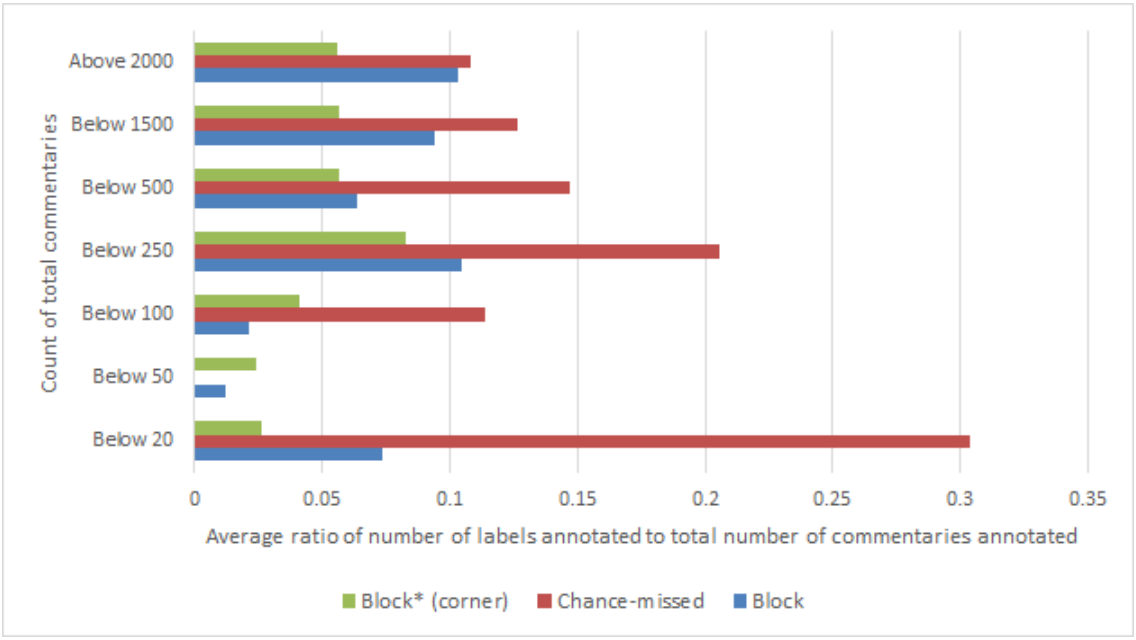


Figure 4.6: Distribution of next 3 annotated labels *Block* (Corner)*, *Chance-missed*, and *Block* across the total count of commentaries annotated by different raters.

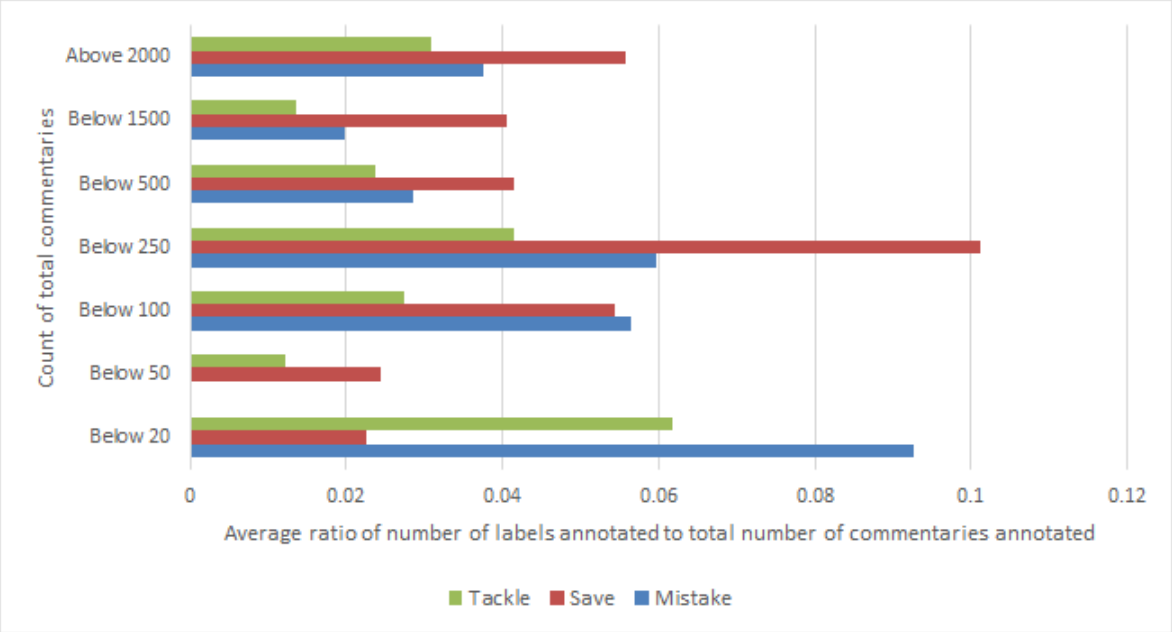


Figure 4.7: Distribution of least 3 annotated labels *Tackle*, *Save*, and *Mistake* across the total count of commentaries annotated by different raters.

cost consuming activity. Moreover, crowdsourcing this task through an application would require the same data-set to be annotated by multiple local experts. This makes it important to measure the degree of agreement among the raters known as *Inter-judge Agreement* or *Inter-rater Agreement*. It evaluates a score to measure the homogeneity and consensus among the ratings or annotations provided by the judges [66]. There are different methods to measure this score :

1. **Joint Probability of agreement** - It is an estimate of how many times the raters agree in a nominal rating system. It is based on the assumption that agreement cannot occur based on a chance.

2. **Kappa Statistics** - It takes into account the amount of agreement that could occur based on a chance.**Cohen's Kappa** κ coefficient is based on the chance-expected disagreement or alternatively, proportion of non-chance agreement. Cohen's Kappa is limited to agreement between two raters. However, **Fleiss's Kapp** is suitable for any number of raters annotating a fixed number of items.

3. **Correlation Coefficients** - It is used to evaluate the correlation between two ranked lists. Different coefficients like Kendall's τ, Spearman's ρ, or Pearson's r accesses the degree of correspondence between two rankings.

4. **Intraclass-corelation coefficient**- This is generally used to infer statistics when the data is organized into groups with a fixed degree of relatedness. It compares the variance in the ratings of the same category with the total variability across all the categories.

For our analysis, we have used **Kappa Statistics** as a measure of agreement between the raters. Kappa coefficient κ was introduced in 1960 as an assessment of agreement between two raters in annotating data into mutually exclusive categories. Kappa is defined as

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{4.1}$$

where p_o is the proportion of data agreed by raters, and p_e is the proportion of agreed data which is expected by chance [67]. There are two variation of kappa's - *Fixed Marginal Multirater Kappa* - when raters are needed to categorize a certain number of cases to an annotation, and *Free Marginal Multirater Kappa* - when raters are not assigned a fixed number of cases to be annotated into a category. *Fleiss' Kappa* and *Cohen's Kappa* are marginally dependent (fixed) coefficients where as *PABAK Kappa* and Brennan and Prediger's κ_m are marginally independent (free) coefficients. Kappa coefficient's value varies from 1 (complete agreement) among raters to -1 (complete disagreement) among raters. The value of 0 implies that current agreement is equivalent to agreement expected from random chances.

Our data set is more suitable for the free-marginal agreement based studies as we allow freedom to rater to assign any number of commentaries to each label irrespective of any fixed number. Table 4.6 reflects the inter-judge agreement kappa coefficient calculated using [68] for 31 users annotating 2000 majority voted commentaries. As mentioned in [69], for certain agreement studies, as in our case, application of fixed marginality to a free-marginal agreement data with the equal number of raters, categories and cases can cause the *Kappa coefficient* to vary significantly.

	Overall Agreement	Fixed Marginal Kappa	Free Marginal Kappa
chance	0.672114	0.128990	0.344229
chance-missed	0.839632	-0.046764	0.679265
foul	0.886619	0.657358	0.773238
block	0.902281	0.108939	0.804562
block* (corner)	0.964249	0.588015	0.928498
save	0.940075	-0.008688	0.880150
mistake	0.945863	0.020044	0.891726
tackle	0.963909	0.090825	0.927818
information	0.918965	0.793527	0.837930

Table 4.6: Individual Class based rater's agreement for the initial 2000 majority vote-based annotated commentaries from 31 different users calculated using [68].

5. CLASSIFIER AND PLAYER ATTRIBUTION

In order to work towards extracting player attribution, we wanted to initially start with identifying the commentary with labels. This would help in the player attribution process later as we will be aware of the essence of that description through those labels and the only task pending would be accrediting it to the players. To automate the process of commentary annotation, we need to build a classifier. In this section, we will discuss our model for classification, its performance and experiments with state of the art classifiers.

Our process of knowledge discover from unstructured textual dataset generally comes under the broad field of "Text Mining". [70] defines two primary components involved during the information extraction from text mining :

1. **Text Refining** - Converting raw text into an *Intermediate Form* - semi-structured documents such as graph, vectorized representation, or structured relation data. *Intermediate Form(IF)* be of two types -

   Document-based where deduction is based on patterns and relationships among documents as in Clustering, and Visualization.

   Concept-based where entities represent objects in specific domain and deduction is based on identifying patterns and relationships across entities as in Predictive modeling and associative discovery.

2. **Knowledge Distillation** - Inferring knowledge from the Intermediate Form like organizing documents for visualization, clustering or derive knowledge from projecting documents into concept-space.

In our approach, we are combining both types *Intermediate Form*. We first organize data into *Document-based IF* and then project them into *Concept-based IF* by annotating the data. Then, for *Knowledge Distillation*, we train our model to classify these entities into their respective categories.

5.1 Preprocessing data

Before we perform training of our classifier with actual data, we preprocessed the data to make it suitable to be fed into our classification model. There are three main steps involved in preprocessing :

1. **Entity Collection** : In order to support the language model better to our data set, the first step was to remove the entities like Person Name, Team name and Location from our data. For this, we employed Stanford Named Entity Recognizer (NER) to get the tagged data. Before training our model, we made a list of NER tagged data corresponding to each match and stored it in a Python dictionary. We employed Stanford CoreNLP NER[1] functionality through Python Natural Language Toolkit library[2]. It is necessary for us to remove all the entities to allow our classifier to train with only the words that can define the outcome of a commentary rather than the entity. So, we initially collected all the entities in our data and later after tokenization, we removed them from our data.

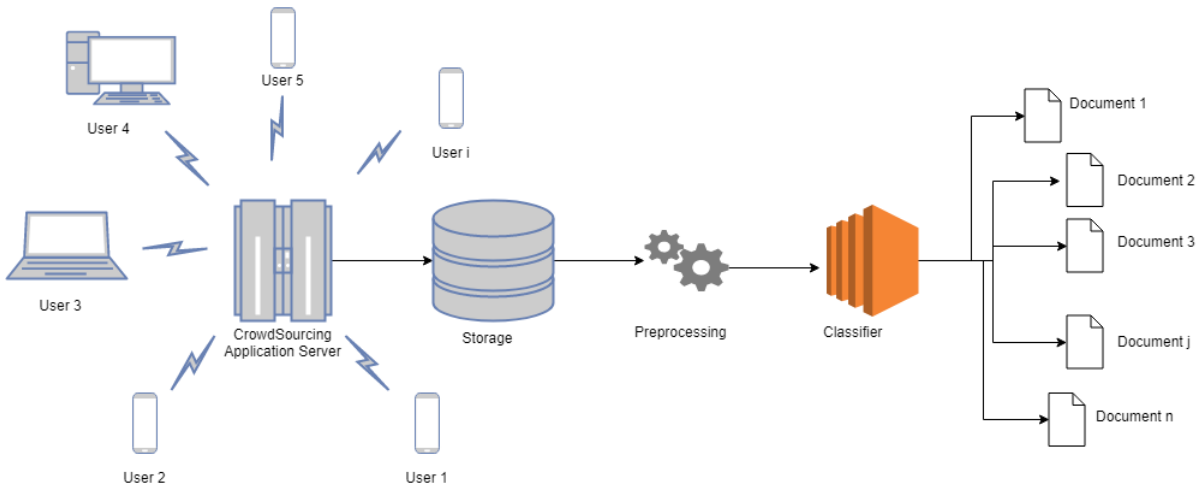


Figure 5.1: A block diagram representing flow of data.

[1]https://nlp.stanford.edu/software/CRF-NER.html
[2]http://www.nltk.org/

2. **Tokenization** : We first tokenized the words using space as separator. After tokenization, we all words were transformed to lowercase. We removed all the entities from these tokens using entity collection as described in the previous step. Tokenization refers to breaking up of the entire sentence into different components (words) and classifying them into different categories.Theoretically. splitting on white spacing should be considered as an approach. However, this method would end up in incorrectly classifying tokens corresponding to places, names or foreign language phrases. This would result in bad performance of the systems involving information retrieval from text. For example, search for Carolina University should not result in North Carolina University. This is the reason we implemented tokenization approach from Python Natural Language Toolkit[2]. To get the meaningful knowledge and a better realization of Natural Language Processing techniques, it is good practice to remove stop words. So, common words were removed after tokenization using the modified stopwords list from the NLTK [2].

3. **Data Set Creation** : This step involves formatting the data to make it suitable for training the classifier. The formatted data set would consist of information content - event type, commentary, classes and tokenized words. The data set was stored in DataFrame object of Pandas Python library 4 [3]. DataFrame object is a 2-dimensional data structure categorized by labels as in a spreadsheet, or a SQL table, or a dictionary of Series objects with potentially different types of columns.

5.2 Event classification

After filtering out the low information content and preprocessing the data, we want to know what kind of event it is. As mentioned in 4.4, we want to identify the event types for the commentaries. We have applied various algorithms for training these classifier and evaluated them based on their results.

We initially focused towards evaluating the performance of single label classification. Multi-class classification is different from multi-label classification. In multi-class classification, there

[3]http://pandas.pydata.org/

are multiple categories but each instance is labeled with only one. However in multi-label classification, each instance can be annotated with multiple categories. We have implement both multi-class and multi-label classification. For multi-class classification, we used Naive Bayes and Logistic Regression for evaluation and for multi-label we used 4 different approaches.

To start with, we implemented the Naive Bayes model for this classification. Naive Bayes is one of the earliest and widely popular approach for Information Retrieval models. Until a few years back, it accounted for most applications of supervised learning to information retrieval.[71]. This widely used Naive Bayes classification approach is based on simple yet novel *Baye's theorem* which describes the occurrence of an event given certain condition(s), based on prior knowledge of condition(s) that may or may not be related to the event. It is mathematically stated as :

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)} \tag{5.1}$$

In terms of classification problem, we can state it using two random variables C - representing k classes $(c_1, ...c_i, ...c_k)$ and X - vectorial representation of features $x = (x_1, ...x_j, ...x_d)$ with the length of each vector d, the number of documents. $P(C = c_k \mid X = x_d)$ is conditional probability describing that the document X belongs to class c_k given it has feature x_d. In terms of random variables, Baye's rule can be rewritten as :

$$P(c_k \mid x) = \frac{P(x \mid c_k)\, P(c_k)}{P(x)} \tag{5.2}$$

Term $P(x \mid c_k)$ is evaluated by calculating the probability of a feature across all the documents d , given the class c_k assuming feature i across d documents occurs statistically independent of each other -

$$P(x \mid c_k) = \prod_{i=1}^{d} P(x_i \mid c_k) \tag{5.3}$$

Probabilistic models represents distribution of possible outcomes and their likelihoods. This eval-

uation is an estimate of class c_k given a document x.

$$P(c_k \mid x) = \frac{\prod_{i=1}^{d} P(x_i \mid c_k)\, P(c_k)}{P(x)} \tag{5.4}$$

The classification model based on 5.4 is that the document with feature vector x belongs to class c_k such that it has the highest probability for all the classes. This type of model is described as *Naive Bayes Classifier*. It is termed as *naive* as it relies on 2 simple assumptions. Firstly, it assumes the conditional independence of feature documents given a class. Secondly, it makes classification decision based on the postulation that latent or hidden factors does not contribute to the prediction process [72]. Since denominator $P(x)$ is common across all classes, the minimum error classification model uses just the value of numerator for making decision [71], and this is termed as *maximum a posteriori (MAP)* hypothesis. MAP calculation involves multiplying a lot of small probabilities, so, generally logarithmic probabilities are used to avoid arithmetic underflow. There are different variations of Naive Bayes Classifiers :

1. **Gaussian Naive Bayes** - If we assume the normal distribution of numerical attributes of documents across each class, we can represent the distribution in terms of its mean and standard deviation as

   $$P(x \mid c_k) = \frac{1}{\sigma_{c_k}\sqrt{2\pi}} e^{-\left(x-\mu_{c_k}\right)^2 / 2\sigma_{c_k}^2} \tag{5.5}$$

   Gaussian Naive Bayes is useful when continuous values associated with each class are distributed according to a Gaussian distribution [72].

2. **Multi-variate Bernoulli Naive Bayes** - In this model, the document is represented as a binary vector over the space of dimensions (words). The vocabulary size V represents the number of dimensions equating each word of vocabulary with a dimension. Document i in dimension t is represented by a binary value, B_{it} corresponding to the presence or absence

of dimension(word w_t) in the document. Probability of a document x given class c_j [73]:

$$P(x \mid c_k) = \prod_{t=1}^{|V|} B_{xt} P(w_t \mid c_k) + (1 - B_{xt}(1 - P(w_t \mid c_k))) \tag{5.6}$$

To avoid the probabilities of 0 or 1 for conditional probability of word (dimension) in a document given the class, we use Laplacean prior - adding each word's count with 1. The probability of a word w_t in class c_j is given by

$$P(w_t \mid c_k) = \frac{1 + \sum_{i=1}^{|D|} B_{it} P(c_k|x)}{2 + \sum_{i=1}^{|D|} B_{it} P(c_k|x)} \tag{5.7}$$

This model is useful when the presence or absence of a feature is the deciding factor for classification.

3. **Multinomial Model** - In contrast to *Multi-variate Bernoulli Naive Bayes*, Multinomial model uses the word frequency in each document rather than just the occurrence or absence. The Multinomial Naive Bayes variation calculates the conditional probability of a particular word/term/token given a class as the relative frequency of term t in documents belonging to class c. The probability of a document x with d different tokens, being in class c_j is computed as

$$P(c_j|x) \propto P(c_j) \prod_{1 \leq k \leq d} P(t_k|c_j) \tag{5.8}$$

where $P(t_k|c_j)$ is the conditional probability of term t_k occurring in a document of class c_j. $P(t_k|c_j)$ is interpreted as a measure of how much weightage or evidence t_k contributes towards the correct classification of class c_j. $P(c_j)$ is the prior probability of a document occurring in class c_j. Using the laplacian prior as discussed earlier, the probability of a word w_t in class c_j with N_it unique tokens and a vocabulary size of $|V|$ is given by [73]

$$P(w_t \mid c_k) = \frac{1 + \sum_{i=1}^{|D|} N_{it} P(c_k|x)}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|D|} N_{is} P(c_k|x)} \tag{5.9}$$

For text based analytics in certain domains like news, capturing the word frequency is more informative as compared to *Multivariate Bernoulli Model*.

To understand the performance of Naive Bayes and the validation of its underlying assumptions, we implemented **Logistic Regression**. Logistic regression belongs to the family of exponential or log-linear classifiers. Similar to Naive Bayes, the log-linear classifier works by identifying a set of features, applying aapropriate weight to each of them, taking logarithmic operation and finally, combining them linearly. Technically, term "logistic regression" refers to a classifier that classifies an observation into one of two classes (binary classification), and multinomial logistic regression classifies into one among more than two classes [74]. The important difference between naive Bayes and logistic regression is that naive Bayes is a generative classifier while the latter one is a discriminative classifier. Generative classifier models like Naive Bayes and Hidden Markov Model, learns the joint probability $p(x, y)$ of the input x and class y and then classify the input using Baye's Theorem. On the other hand, discriminative models like A discriminative model like Support Vector Machine and Logistic Regression, directly learns the relationship between inputs and outputs. It takes this direct approach of discriminating among the different possible values of the class y by computing $P(y|x)$ rather than computing a likelihood first:

$$\hat{y} = \arg\max_{y} P(y|x) \tag{5.10}$$

The best way to achieve this is to model the conditional probability $P(y|x)$ as a function of x. The linear models are unbounded and suffers with diminishing returns when value of p is high enough. Logistic transformation of $logp$ is the solution to this approach :

$$log(\frac{p(x)}{(1-p(x)}) = \beta_0 + x\beta \tag{5.11}$$

Solving for $p(x))$ yields ,

$$p(x) = \frac{\exp(\beta_0 + x\beta)}{1 + \exp(\beta_0 + x\beta)} = \frac{1}{1 + \exp(-(\beta_0 + x\beta))} \tag{5.12}$$

We implemented this model using scikit-learn python package [4]. We used Stochastic average gradient (SAG)[5] method as the optimization method to iteratively find the minima of an objective function. Also, regularization strength set to 10 gave the best results. Only numerical feature were extracted using the CountVectorizer method from scikit-learn python package. Vectorization refers to the process of converting a collection of text documents into numerical feature vectors. CountVectorizer implements both tokenization and occurrence counting in a single class.

For multi-label classification, there are three approaches:

1. **Problem Transformation** : This approach deals with multi-label classification by splitting the problem into single-label classification problems. There are 3 approaches to this :

   Binary Relevance - It treats each label as a separate single-label classification problem. In our case, 9 classifiers corresponding to 9 classes would be trained and the prediction of each of them is aggregated in the end.

   Classifier Chains - It initially trains the first classifier with the training set. Then, the subsequent classifier(s) is trained on the aggregated chain of training data and the previous classifier(s). It is similar to *Binary Relevance* but the difference lies in maintaining the label correlation through chains.

   Label Powerset - It breaks down the multi-label classification to multi-class classification. A single multi-class classifier is trained on all the unique set of label combination found in the training data-set.

2. **Adapted Algorithm** : As the name suggests, the algorithm adapts to the multi-label classification problem rather than transforming into different problem set. MLKNN is one such approach. It is Multi-label version of K-Nearest Neighbors approach. For k nearest neighbors, we extract statistical information - the number of data element belonging to each class and then maximum a posteriori (MAP) principle is evaluated to predict the labels for the current instance[75].

[4]https://pypi.python.org/pypi/scikit-learn

[5]SAG is a stochastic approach towards the gradient descent optimization

3. **Ensemble Approach** : The data-set is trained on multiple models. A small subset of labels are considered at a time. In RAndom k-labELsets (RAKEL) algorithm, single-label classifier learns from the powerset of this subset.

5.3 Template matching

We introduce the concept of **auto-tagged commentaries** which we define as straight forward expert comments from which we extract data by applying pattern-matching. Due to majority vote based interjudge agreement policy, the amount of commentaries that are available for training on classifier were limited. Meanwhile, we discovered certain patterns in some of the commentaries. For example,

Foul by Milan Skriniar (Sampdoria).

Foul by Riccardo Meggiorini (Chievo).

These two commentaries are similar in context if we remove the player and team information. Any such commentaries starting with *"Foul by"* template can be labelled as *Foul*. Other templates are as follows :

Template : *""foul", "yellow card", "yellow-card", "free kick in the defensive half"*
Labels : *Foul*
Example :

Víctor Ruiz (Villarreal) wins a free kick in the defensive half.

Template : *"hand ball", "red-card", "red card", "yellow card for a bad foul"*
Labels : *Foul, Mistake*
Example :

Hand ball by Zlatko Junuzovic (SV Werder Bremen).

Template : *"offside", "substitution", "delay in match", "second half begins", "second half ends", "first half begins", "first half ends", "match ends", "another substitution", "delay over"*

Labels : *Information*

Example :

Substitution, Hamburger SV. Pierre-Michel Lasogga replaces Michael Gregoritsch.

Template : *"attempt blocked"*

Labels : *Chance, Block*

Example :

Attempt blocked. Marcel Risse (1. FC Köln) right footed shot from the right side of the box is blocked. Assisted by Dominique Heintz.

Template : *"attempt missed"*

Labels : *Chance, Chance-missed*

Example :

Attempt missed. Mathew Leckie (FC Ingolstadt 04) right footed shot from the centre of the box misses to the left. Assisted by Moritz Hartmann following a fast break.

Template : *"attempt saved"*

Labels : *Chance, Save*

Example :

Attempt saved. Philipp Max (FC Augsburg) left footed shot from outside the box is saved in the bottom right corner. Assisted by Dong-Won Ji.

Template : *"corner"*

Labels : *Block* (Corner)*

Example :

Corner, FC Ingolstadt 04. Conceded by Diego Benaglio.

Template : *"free kick in the attacking half", "free kick on the right wing", "free kick on the left wing"*

Labels : *Chance, Foul*

Example :

Dominik Kohr (FC Augsburg) wins a free kick in the attacking half.

Labels : *Chance, Foul*

All the auto-tagged commentaries are similar. They just differ in terms of player and team information. Moreover, while training the data, we do remove named-entites using NER annotator. So, such commentaries can be auto-tagged by considering the number of words per sentence, number of sentences in the commentary and any matching pattern to identify with label(s). This template-matching based tagging would result in better performance of our classifiers.

5.4 Preliminary results

In order to understand the feasibility of our approach, we started with building a classifier on the pre-labelled data available from GOAL.com. We worked on a developing a basic classification system initially to verify our approach. We collected data for 850 matches from November 2016 to April 2017. Of these 850 matches, we selected only 184 matches according to the information content in their commentaries. These 184 matches had 11384 events. We distinguished the commentaries based on the event-tags as :

1. **Action** : GOAL.com classifies all the non-key events as 'action'. This does not include any of the key events as described in 4.2.

2. **Not just Action** : Includes all events except the **Action** labelled commentaries. Game-events which results into any of the labels as defined by GOAL.com. These events are : yellow-card, substitution, assist, goal, penalty-goal, red-card, own-goal, missed-penalty, penalty-save and yellow-red.

Table 5.1 and 5.2 represents the evaluation matrix and classification metrics for these basic binary classifier.

	Action	Not Just Action
Action	8488	138
Not Just Action	410	2348

Table 5.1: Preliminary Evaluation Matrix for Binary Classifier using Multinomial NB in k-fold cross validation (k=10)

	Precision	Recall	F1-Score
Action	0.94	0.85	0.90
Not Just Action	0.95	0.98	0.97
Avg/Total	0.95	0.95	0.95

Table 5.2: Preliminary Classification Report for Binary Classifier using Multinomial NB in k-fold cross validation (k=10)

With the good performance of the binary classification system, we decided to include a second stage classifier. We started focusing towards classifying the 'Not just action' i.e.,relevant events further into 10 different classes as available pre-annotated from GOAL.com. Explained in 4.2, these labels are *action , yellow-card , substitution , assist , goal , penalty-goal , red-card , own-goal , missed-penalty , penalty-save* and *yellow-red*.

Table 5.3[6] and 5.4 shows performance of event classification system for GOAL.com generated labels on the same data set of 11384 commentaries. We shortlisted 184 matches out of 850 matches, for rich quality of commentary. Our model was based on only one feature - *bag of words*. This preliminary experiment was meant to verify our approach towards dealing with commentaries for soccer based information statistics retrieval.

[6]Sub - Substitution, YC - Yellow Card, PG - Penalty-Goal, OG - Own-Goal, YRed - Yellow Red Card, RC - Red Card, PS - Penalty-saved, MP - Missed-Penalty

	Sub	YC	Assist	Goal	PG	OG	YRed	RC	PS	MP
Sub	1045	0	0	0	0	0	0	0	0	0
YC	0	721	1	3	0	0	0	0	0	0
Goal	0	0	509	4	0	0	0	0	0	0
Assist	0	5	7	381	0	0	0	0	0	0
PG	0	0	9	0	18	0	0	0	0	0
OG	0	0	10	0	0	6	0	0	0	0
YRed	0	6	1	0	0	0	8	0	0	0
RC	0	1	2	0	0	0	2	3	0	0
PS	0	1	0	0	0	0	0	0	5	0
MP	0	0	1	0	3	0	0	0	0	6

Table 5.3: Preliminary Evaluation Matrix for Event Classifier using Linear Model - Logistic Regression in k-fold cross validation (k=10)

	Precision	Recall	F1-Score
Substitution	1.00	1.00	1.00
Yellow-card	0.98	0.99	0.99
Goal	0.94	0.99	0.97
Assist	0.98	0.97	0.98
Penalty-goal	0.86	0.67	0.75
Own-goal	1.00	0.38	0.55
Yellow-red	0.80	0.53	0.64
Red-card	1.00	0.38	0.55
Penalty-save	1.00	0.83	0.91
Missed-penalty	1.00	0.60	0.75

Table 5.4: Preliminary Classification Report for Event Classifier using Linear Model - Logistic Regression in k-fold cross validation (k=10)

During our preliminary experiments, we also ventured into exploring the combining the first and second stage of classification models and evaluate the combined performance. For the combined classification, we evaluated both the classifier trained with the same type of Naive Bayes approach as well as combination of models like Gaussian Gaussian Naive Bayes, Multinomial - Multinomial Naive Bayes, Gaussian - Multinomial Naive Bayes, Multinomial - Bernoulli Naive Bayes, Multinomial Naive Bayes - Gaussian Naive Bayes. In addition, we also trialled with a fu-

sion of MaxEnt model based logistic regression and Naive Bayes variations. The MaxEnt model based logistic regression variant worked best in both the classifications and also, it has better results as compared to combination of models too. The overall accuracy for the combined classification using this model was 97.4%. Table 5.5 and 5.6 shows the evaluation matrix and classification report for the combined classifier using Logistic Regression for the same data set. We notice that the performance of labels *Penalty-goal*, *Own-goal*, *Yellow-red*, *Red-card*, *Penalty-save* and *Missed-penalty* has been lower due to the scarcity in the availability of commentaries with those labels as compared to the other labels.

	Action	Sub	YC	Goal	Assist	PG	OG	YRed	RC	PS	MP
Action	8570	0	16	35	4	0	1	0	0	0	0
Sub	0	1045	0	0	0	0	0	0	0	0	0
YC	30	0	695	0	0	0	0	0	0	0	0
Goal	103	0	0	408	2	0	0	0	0	0	0
Assist	11	0	0	0	382	0	0	0	0	0	0
PG	4	0	0	11	0	12	0	0	0	0	0
OG	13	0	0	2	0	0	1	0	0	0	0
YRed	3	0	8	0	0	0	0	4	0	0	0
RC	4	0	0	0	0	0	0	0	4	0	0
PS	6	0	0	0	0	0	0	0	0	0	0
MP	7	0	0	0	0	1	0	0	0	0	2

Table 5.5: Preliminary Evaluation Matrix for Combined Classifier using Linear Model - Logistic Regression in k-fold cross validation (k=10)

With the good outcome of our preliminary results, we decided to move forward with crowd-sourcing the data for our manual tags and analyze our performance on it.

5.5 Player attribution

According to our research plan, we were simultaneously working towards gathering data and figuring out how to identify player in the commentary. Our approach was based on identifying

	Precision	Recall	F1-Score
Action	0.98	0.99	0.99
Substitution	1.00	1.00	1.00
Yellow-card	0.97	0.96	0.96
Goal	0.89	0.80	0.84
Assist	0.98	0.97	0.98
Penalty-goal	0.92	0.44	0.60
Own-goal	0.50	0.06	0.11
Yellow-red	1.00	0.27	0.42
Red-card	1.00	0.50	0.67
Penalty-save	0.00	0.00	0.00
Missed-penalty	1.00	0.20	0.33

Table 5.6: Preliminary Classification Report for Combined Event Classifier using Linear Model - Logistic Regression in k-fold cross validation (k=6)

"Person" named entities using Named Entity Recognition (NER) labels from Natural Language Processing. Named Entity Recognition Labels are the names of things, such as person and company names, or gene and protein names. We encorporated Stanford CoreNLP Library for this purpose [7]. The Named Entity Recognizer in the Stanford CoreNLP Library is based on the augmented *Conditional Random Fields (CRF)* framework (CRF utilizes probabilistic models to segment and label sequence data). It uses Gibbs Sampling, a simple Monte Carlo algorithm on non local dependencies to sequence models along with the existing CRF model[76]. Due to lack of availability of annotated training data, we cannot verify the performance of this library on soccer domain. However, according to [76], it has an F1-score of 92.3% on CMU seminar announcements and 86.86% on Computational Natural Language Learning (CoNLL) 2003 English named entity recognition dataset.

We initially used just Stanford Named Entity Recognition library to extract the player information from the commentary. However, we were not satisfied with the performance of these libraries and we decided to dig dipper into the Natural Language Processing algorithms. We used Stanford CoreNLP annotators to tokenize and split commentary sentences. Along with this, we also

[7] https://stanfordnlp.github.io/CoreNLP/

generated dependency parsing tree, NER labels, and coreference-resolution tags.

Danilo replaces Zinchenko for the home side.

For example, the parsing tree for the above commentary is:

```
(ROOT
  (S
    (NP (NNP Danilo))
    (VP (VBZ replaces)
      (NP
        (NP (NNP Zinchenko))
        (PP (IN for)
          (NP (DT the) (NN home) (NN side)))))
  (. .)))
```

The NER tagging is :

Danilo $PERSON$ replaces *Zinchenko* $PERSON$ for the home side.

The dependencies are :

$nsubj$(replaces-2, Danilo-1)

$root$(ROOT-0, replaces-2)

$dobj$(replaces-2, Zinchenko-3)

$case$(side-7, for-4)

det(side-7, the-5)

$compound$(side-7, home-6)

$nmod$(Zinchenko-3, side-7)

We identified all the tokens (words) with $PERSON$ in the NER tagging and looked for any $compound$ relations as sometimes first and last name both were not tagged as $PERSON$. The

dependencies in such case resolved by looking for $compound$ relationships which in most of the cases would relate the names correctly. In this way, we were able to extract player names from the commentary.

Post player identification, we need to match player name with that in our database. The text matching algorithm can be explained by various approaches like Cosine Similarity, Levenshtein Distance, Euclidean distance, block distance (Manhattan distance), Jaccard Similarity, term-frequency inverted document frequency (TF-IDF) score and many more [77]. We shortlist players based on the lineup or the squad of both the teams contesting. However, due to different naming convention of GOAL.com and SoFIFA.com, we need a text search functionality to link team names too. Here, we utilize *"Full-Text Search"* feature provided by MongoDB. *Full-Text Search* feature is based on stemming (reducing the word to its original root verb), removing stop words and indexing the text for a faster search response.

We have multiple labels from classification model and multiple player information from player matching using MongoDB. The next step is to identify event with the correct player. Our current model assigns events based on the player position. With this approach, we are able to correctly predict upto some extent the multi-player multi-label relation for *Chance*, *Chance-missed*, *Block*, *Tackle*, *Block* (corner)* and *Save* tags. This approach is based on the assumption that mostly strikers or mid-fielders are responsible for creating chances or missing upon any opportunities, defender or midfielders are accountable for tackles/blocks/corners, and goalkeeper for making saves. For tags *Foul* and *Mistake*, we are able to handle single-player single-label relation and some cases for multi-player multi-label relation provided that the labels contain only one of the either. However, this approach is not robust and will fail in the following scenarios :

1. When we want to relate two forward or midfield players with two entities which can be attributed to either of them using our approach. For example, if labels are *Chance and Chance-missed* and both the players mentioned in the commentary are strikers, then we cannot figure out which player to ascribe it to.

2. When a defender or attacker is responsible for an attacking move or block respectively, our

assumption fails under such scenario.

3. When commentary includes multi-label single-player relation, we cannot substantiate an association.

Our complete work flow for player attribution and classification is illustrated in Figure 1.4 with an example commentary. There can be multiple sources for text commentaries - live commentary from GOAL.com, old match transcript (image-to-text converison), video or audio matches (speech to text conversion) and expert reviews or updates on social media platforms. Once, a commentary is available our system could simultaneously classify it and extract player information as show in Figure 1.4. However, extracting player information requires a prior step to parse it using NLP models and it is very expensive in terms of resource utilization if we parse individual commentaries. For our current model, we process the entire data of a single match in a single run. Finally, linking process initiates when both the entities and events are available and the player database is updated with the corresponding statistics.

6. RESULTS

In this section, we will discuss our results for different types of classifiers with different combination of data-sets and our player attribution approach as described earlier. We will also analyze the performance of our methods and explore the reason behind it.

	Precision	Recall	F1-Score
Action	1.00	1.00	1.00
Not Just Action	0.98	0.98	0.98

Table 6.1: Classification Report for Binary Classifier using Multinomial NB in k-fold cross validation (k=10)

	Precision	Recall	F1-Score
Action	1.00	1.00	1.00
Substitution	0.99	1.00	0.99
Yellow-card	1.00	0.99	0.99
Goal	0.82	0.73	0.77
Assist	0.73	0.50	0.59
Penalty-goal	1.00	0.99	0.99
Own-goal	0.75	0.33	0.46
Yellow-red	0.00	0.00	0.00
Red-card	0.78	0.70	0.74
Penalty-save	0.00	0.00	0.00
Missed-penalty	1.00	0.50	0.67

Table 6.2: Classification Report for Combined Event Classifier using Linear Model - Logistic Regression in k-fold cross validation (k=6)

We gathered a data collection of **94221** commentaries for around 936 matches. This included **63610** auto-tagged commentaries and remaining crowdsourced majority voted commentaries. We

started with experimenting our model on GOAL.com pre-labelled data tags for this 94221 commentaries. Since our labelling does not influence the classification of these pre-labelled data tags, we experimented this model only with the entire collection of commentaries. Classification report for the combined event classifier model is shown in Table 6.2. We see a decrease in the performance mainly due to a mix quality of data. In contrast to preliminary results where we had explanatory commentaries, here, we did not restrict the commentaries to be of very rich quality and descriptive. Negligible F1-scores for labels *Yellow-red*, *Penalty-save* and *Missed -Penalty* can be attributed to the lack of sufficient commentaries pertaining to these tags.

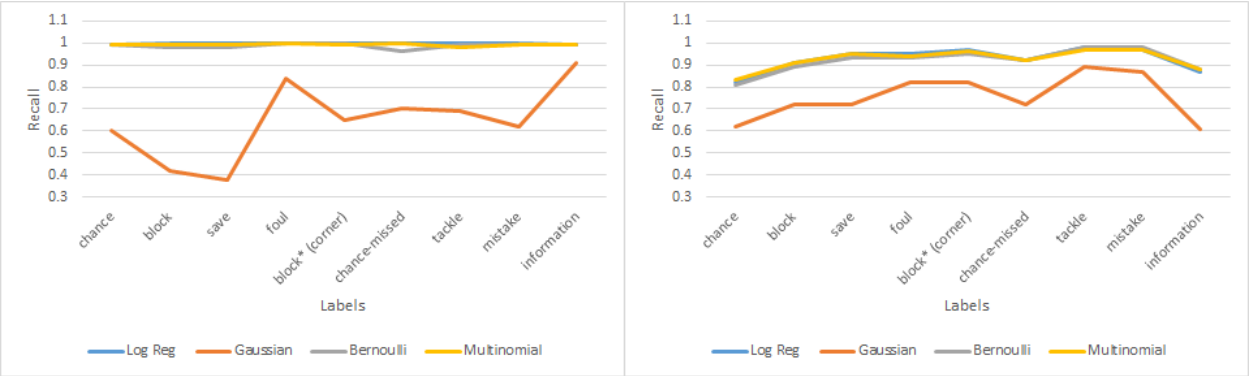
	Gaussian	Bernoulli	Multinomial	Logistic Regression
Chance	0.61	0.99	0.98	0.99
Not Chance	0.58	0.99	0.99	0.99
Block	0.16	0.85	0.91	0.97
Not Block	0.55	0.99	0.99	1.00
Save	0.15	0.86	0.96	0.98
Not Save	0.51	0.99	1.00	1.00
Foul	0.80	1.00	1.00	1.00
Not Foul	0.87	1.00	1.00	1.00
Block* (corner)	0.35	0.99	0.95	0.99
Not Block* (corner)	0.77	1.00	0.99	1.00
Chance-missed	0.38	0.81	0.98	0.98
Not Chance-missed	0.81	0.97	1.00	1.00
Tackle	0.0	0.13	0.09	0.46
Not Tackle	0.82	1.00	0.99	1.00
Mistake	0.17	0.89	0.88	0.97
Not Mistake	0.75	0.99	0.99	1.00
Information	0.80	0.97	0.97	0.98
Not Information	0.94	0.99	0.99	1.00

Table 6.3: F1-scores for single-label multi-class Additional Event Classifier using Naive Bayes' variations and Logistic Regression in k-fold cross validation (k=6) for combined data-set of auto-tagged and crowd-sourced commentaries

We verified our single-label multi-class additional-events based classification using the same models. We find that our performance was not so good when we used Gaussian Naive Bayes

	Gaussian	Bernoulli	Multinomial	Logistic Regression
Chance	0.55	0.83	0.84	0.84
Not Chance	0.68	0.79	0.82	0.80
Block	0.27	0.33	0.44	0.51
Not Block	0.83	0.94	0.95	0.95
Save	0.20	0.38	0.61	0.66
Not Save	0.83	0.96	0.97	0.97
Foul	0.64	0.83	0.85	0.88
Not Foul	0.88	0.96	0.96	0.97
Block* (corner)	0.26	0.14	0.48	0.71
Not Block* (corner)	0.90	0.97	0.98	0.98
Chance-missed	0.26	0.30	0.45	0.54
Not Chance-missed	0.83	0.96	0.96	0.96
Tackle	0.12	0.00	0.21	0.46
Not Tackle	0.94	0.99	0.99	0.99
Mistake	0.09	0.00	0.06	0.26
Not Mistake	0.93	0.99	0.99	0.98
Information	0.60	0.82	0.81	0.81
Not Information	0.62	0.91	0.91	0.90

Table 6.4: F1-scores for single-label multi-class Additional Event Classifier using Naive Bayes' variations and Logistic Regression in k-fold cross validation (k=6) for crowd-sourced commentaries only



(a) (b)

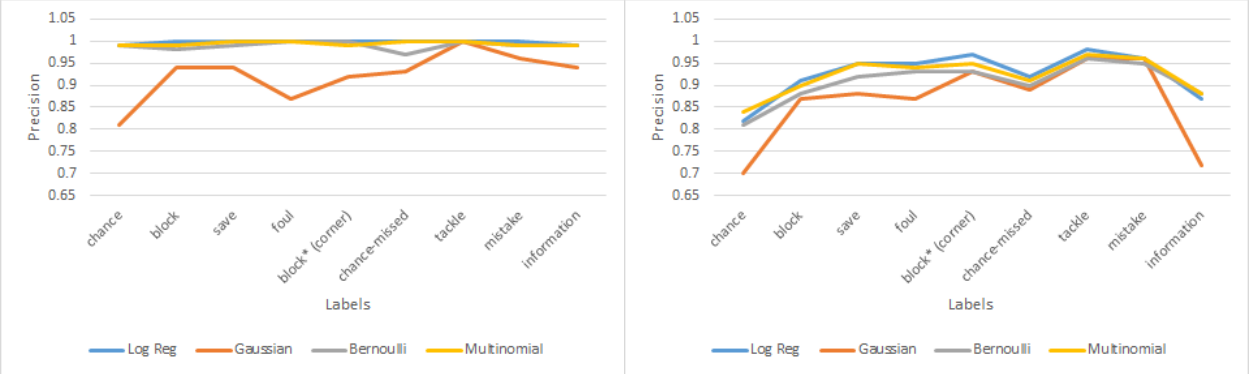
Figure 6.1: Average Recall for single-label multi-class classifier using Naive Bayes' variations and Logistic Regression models for (a) Combination of Auto-tagged and Crowd-sourced Commentaries, and (b) Crowd Sourced Commentaries Only.

model. This is due to the fact that Gaussian Naive Bayes is more suitable for normally distributed data-set. In our case, it was more random and the labels were more suitable for Bernoulli Naive Bayes and Multinomial Naive Bayes where more emphasis is given on word-presence and word-count. The bag of words feature focuses on the frequency and weight of each word according to its label which gives a good prediction output for these commentaries.The greater accuracy with less amount of data can be attributed to this feature. The event-tag for a commentary is decided by a few words either individually or in combination. As we read and comprehend a commentary, our annotation of events is based on the selected key words which provides us with a gist of the event. A commentary with words like *"foul"*, *"harsh challenge"*, and *"sliding tackle"* are more like to contribute to the *"foul"* label, however, words like *"shot"*, *"wonderful pass"*, *"opportunity"*, and *"goal"* shows more affinity to a *"chance"*.

F1 scores for single-label multi-class additional-events based classification for a data-set combination of both auto-tagged commentaries and crowd-sourced commentaries is shown in Table 6.3. Since we included two different sets of annotated data into our classifier, we also analyzed the performance of just the crowd-sourced annotated data-set. Table 6.4 represents F1 scores for crowd-sourced commentary data-set only. The corresponding precision and recall scores are also represented in Figure 6.1 and Figure 6.2, respectively.

Figure 6.3 illustrates the comparison of crowd-sourced annotated data-set with the combined data-set. The reduced performance of the crowd-labelled trained classifier is caused by the lack of crowd-sourced annotated data. We believe that with greater data-set the performance will improve. We obtained an average accuracy of **92.4%** on combined dataset with crowdsourced and auto-tagged commentaries and **62.4%** on only crowdsourced commentaries.

For multi-label multi-class classification, we trained on 4 different models as discussed earlier. For *Binary Relevance*, *Chain Classifier* and *Label Powerset* models, we implement three variations of Naive Bayes classifiers too. As we can observe from Table 6.5, we find that the accuracies for both crowdsourced data and combined data is low for Gaussian Naive Bayes models. However, it improves significantly for Bernoulli and Multinomial models. We also detected that Multi-Label



(a) (b)

Figure 6.2: Average Precision for single-label multi-class classifier using Naive Bayes' variations and Logistic Regression models for (a) Combination of Auto-tagged and Crowd-sourced Commentaries, and (b) Crowd Sourced Commentaries Only.

Figure 6.3: Comparison of average F1-Score for single-label multi-class classifier with Crowd-Sourced data-set and the entire 936 match data-set.

kNN (MLkNN) gave the best performance for multi-label classification. The adaptive approach resulted in better outcome for MLkNN but the computation cost was too high for it. Similarly, we made an attempt to test Logistic Regression with this model, however, it was very resource and time consuming and we decided to abandon it.

Models	Accuracy (Crowdsourced)	Accuracy (Combined)
BinaryRelevance (GaussianNB)	0.1827	0.1397
BinaryRelevance (BernoulliNB)	0.3733	0.9292
Binary Relevance (MultinomialNB)	0.4533	0.9648
Chain Classifier (GaussianNB)	0.1800	0.1420
Chain Classifier (BernoulliNB)	0.4056	0.9226
Chain Classifier (MultinomialNB)	0.4517	0.9676
Label Powerset (GaussianNB)	0.3340	0.3853
Label Powerset (BernoulliNB)	0.4528	0.9755
Label Powerset (MultinomialNB)	0.5687	0.9764
MLkNN	0.3722	0.9762

Table 6.5: Accuracy for Multi-label Event Classifier using different approaches for all 936 match commentaries.

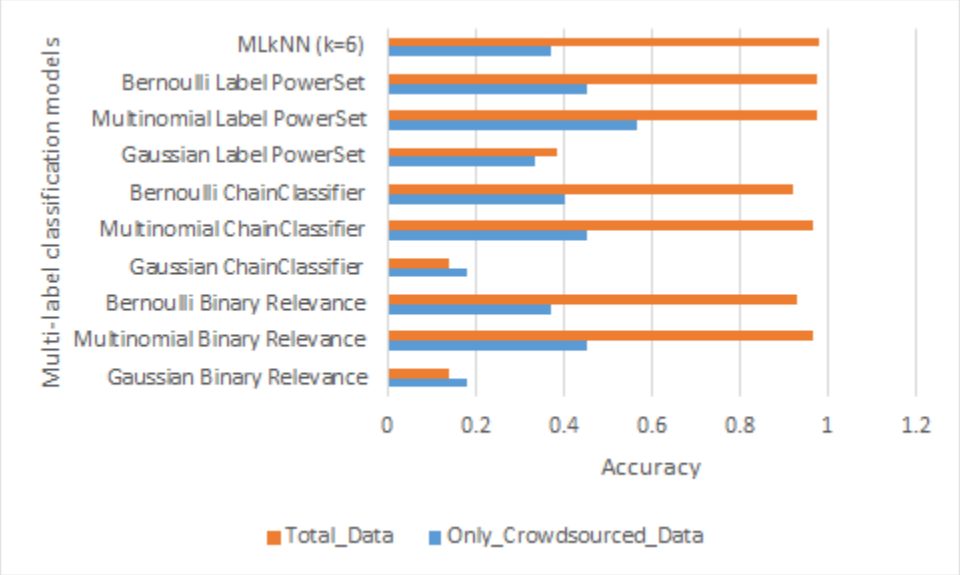


Figure 6.4: Accuracy comparison amongst multi-label classification models for Only Crowd-Sourced Data-set and Total Data-set.

We also analyzed the performance of multi-class multi-label classifiers for Crowd-source data-set only as presented in Figure 6.4. The reason for degraded accuracy of Crowd-sourced trained multi-label classifier is the irregularity in the annotated data. We do not have any means to verify the correctness of data and we rely on majority-vote for data validation. With the increase in

number of users, there will be more randomness in the data as compared to the auto-tagged data-set which is based on pattern matching.

For player attribution, we did not have enough player annotated data-set available. However, we verified our approach with a few auto-tagged data-set and it worked perfectly. We are aware of the limitations of this approach and with a proper dataset, we will be better equipped to support it with results.

7. FUTURE WORK

The analytics in the field of soccer especially based on text analysis is relatively untouched and our research is kick-start to various avenues for further research. Based on our approach and results, we believe that our system can be extended to other sports like Cricket. Moreover, we believe that the dataset generated from crowd-sourcing would serve as good data-source for researchers willing to explore this domain.

7.1 Other sports

We can explore into other domains of sports as well with this model. Currently, there has been enormous advancement going on in the field Natural Language Processing. Abstract Meaning Representation (AMR) is a semantic representation of the sentence encoded to its root. Once a model has been trained in a domain using AMR, it can be used in a new domain for predicting output based on the classes/labels of the new domain. Using this approach, we can robustly train our model to analyze soccer events and apply the model to different sports like Cricket, hockey and similar ball game sports.

7.2 Classification performance improvement

As the commentaries are very dynamic in nature, we will be more focussing towards improving the performance of crowdsourced commentary classification. Currently due to the lack of number of expert users, we lack the quality dataset needed to make our model more robust. With the aggregation of more labelled commentaries, we are planning to take a majority vote of 5 or more users instead of just 3 before assigning labels to a commentary. We will also exploit into word and time based features to improve the classification accuracy of the system.

7.3 Player attribution

The system proposed by us has limitations and is restricted towards a better evolution due to the lack of annotated data. The upcoming advancements in the linguistic models based analytics

has motivated us to dwell more into coming up with an approach to attribute player with events. We will also be focusing towards combining different sources of commentary like BBC sports, Twitter, and Radio commentaries into our system to overcome the uncertainty involved with a few commentaries. Also, processing individual commentaries for a single run is resource inefficient as loading NLP models consumes a good amount of computation power. We will also look into making it more real-time single commentary effective instead of evaluating for an entire game.

8. CONCLUSIONS

In our research, we studied different approaches to soccer analytics and proposed a unique end-to-end system to extract player and team information from soccer commentaries. We started off with a pre-labelled data set to verify our approach and with good results, we created our own labels to help in the process of gathering player statistics. Restricted by a good data-source, we developed from scratch our own data-source through crowdsourcing and built different classification models on top of it. We also identified ways to auto-label a portion of the commentary using template matching. This led to the segregation of our data-set into two sets of commentaries - Crowdsourced commentaries and auto-tagged commentaries. We understood the challenges in multi-label classification models as we were restricted by a large amount of crowd sourced dataset. However, we realized that multiple single label multi-class classifiers can be utilized to favor our approach. We achieved an overall accuracy of 92.4% using logistic regression model for single-label multi-class classification on combined dataset of auto-tagged commentaries and crowdsourced commentaries and an accuracy of 62.42% on crowd-dataset only. We obtained an accuracy of 97.6% on combined data-set and 56.87% on crowdsourced dataset using Multinomial Naive Bayes based Label PowerSet approach.

We tried to explore our way into extracting player information from the commentaries however, in our approach, we were restricted by the limitations of Natural Language Processing approaches to a relatively unique domain. During our research, we realized that soccer commentaries can be dubious to be able for a proper comprehension of the scenario. We limited our research to a single source of commentary.

REFERENCES

[1] Duncan Kuehn, GenealogyBank, "Omaha world herald (omaha, nebraska), 6 july 1954, page 13," 1954. [Online; accessed December 15, 2017].

[2] NewspaperArchive, "brandon-sun-jul-18-1966-p-7," 1966. [Online; accessed December 16, 2017].

[3] NewspaperArchive, "winnipeg-free-press-apr-18-1961-p-20," 1961. [Online; accessed December 20, 2017].

[4] TheCourierMail, "An article about george best during his 1983 stint in brisbane.," 1983. [Online; accessed December 20, 2017].

[5] S. History, "1966 world cup final (part 1) - england beat west germany (excellent colour footage)." https://www.youtube.com/watch?v=3T6IY2fz_Mc#t=3m30s. [Online; accessed December, 2017].

[6] Enki2011, "West germany - hungary. wc-1954. final (3-2)." https://www.youtube.com/watch?v=m9UjdKBzIdI&t=616s. [Online; accessed December, 2017].

[7] bracovcehorovce, "finale me 1976 - zapadne nemecko : Ceskoslovensko." https://www.youtube.com/watch?v=v_Hi0cIZFOI#t=13m19s. [Online; accessed December, 2017].

[8] NIXBLACK, "Greece 1-0 portugal 2004 euro final all goals & extended highlight hd 720p." https://www.youtube.com/watch?v=LWh2YUAjoNw&t=80s. [Online; accessed December, 2017].

[9] thingsthatannoyme.co.uk, "Football-commentators.." https://i.pinimg.com/736x/de/05/22/de0522ea8c1b1691adefc8e513b3c31e--humor-football.jpg, 2013. [Online; accessed January, 2018].

[10] MediaBuyersLLC, "Global soccer corruption case articles.." http://www.mediabuyers.com/wp-content/uploads/2015/12/edit_newsp.png. [Online; accessed January, 2018].

[11] IFEX.org, "A ghanaian fan listens to live commentary on the radio at the soccer village in accra, ghana, 26 june 2014," 2017. [Online; accessed January, 2018].

[12] CoSchedule.com, "Social media plan template," 2017. [Online; accessed January, 2018].

[13] T. I. U. L. S. . L. N. World Soccer, "World soccer january 2018," 2018.

[14] D. Partridge and I. M. Franks, "Computer-aided analysis of sport performance: an example from soccer," *Physical Educator*, vol. 50, no. 4, p. 208, 1993.

[15] L. Xie, P. Xu, S.-F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with domain knowledge and hidden markov models," *Pattern Recognition Letters*, vol. 25, no. 7, pp. 767–775, 2004.

[16] C. Perin, R. Vuillemot, and J.-D. Fekete, "Soccerstories: A kick-off for visual soccer analysis," *IEEE transactions on visualization and computer graphics*, vol. 19, no. 12, pp. 2506–2515, 2013.

[17] P. Comisky, J. Bryant, and D. Zillmann, "Commentary as a substitute for action," *Journal of Communication*, vol. 27, no. 3, pp. 150–153, 1977.

[18] G. Van Oorschot, M. Van Erp, and C. Dijkshoorn, "Automatic extraction of soccer game events from twitter," *Proceedings of the Workhop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2012)*, vol. 902, pp. 21–30, 2012.

[19] T. D'Orazio and M. Leo, "A review of vision-based systems for soccer video analysis," *Pattern recognition*, vol. 43, no. 8, pp. 2911–2926, 2010.

[20] O. F. Camerino, J. Chaverri, M. T. Anguera, and G. K. Jonsson, "Dynamics of the game in soccer: Detection of t-patterns," *European Journal of Sport Science*, vol. 12, no. 3, pp. 216–224, 2012.

[21] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for tv baseball programs," in *Proceedings of the eighth ACM international conference on Multimedia*, pp. 105–115, ACM, 2000.

[22] M. Xu, N. C. Maddage, C. Xu, M. Kankanhalli, and Q. Tian, "Creating audio keywords for event detection in soccer video," in *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, vol. 2, pp. II–281, IEEE, 2003.

[23] S. Choudhury and J. G. Breslin, "Extracting semantic entities and events from sports tweets," 2011.

[24] A. Bruns, K. Weller, and S. Harrington, "Twitter and sports: Football fandom in emerging and established markets," in *Twitter and society*, vol. 89, pp. 263–280, Peter Lang, 2014.

[25] J. Lanagan and A. F. Smeaton, "Using twitter to detect and tag important events in live sports," *Artificial Intelligence*, vol. 29, no. 2, pp. 542–545, 2011.

[26] D. Sacha, M. Stein, T. Schreck, D. A. Keim, O. Deussen, *et al.*, "Feature-driven visual analytics of soccer data," in *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, pp. 13–22, IEEE, 2014.

[27] I. Bojinov and L. Bornn, "The pressing game: optimal defensive disruption in soccer," *Procedings of MIT Sloan Sports Analytics*, 2016.

[28] L. Gyarmati and M. Hefeeda, "Analyzing in-game movements of soccer players at scale," *arXiv preprint arXiv:1603.05583*, 2016.

[29] A. Sahami Shirazi, M. Rohs, R. Schleicher, S. Kratz, A. Müller, and A. Schmidt, "Real-time nonverbal opinion sharing through mobile phones during sports events," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 307–310, ACM, 2011.

[30] M. C. Surabhi, "Natural language processing future," in *Optical Imaging Sensor and Security (ICOSS), 2013 International Conference on*, pp. 1–3, IEEE, 2013.

[31] P. Smyth, U. M. Fayyad, M. C. Burl, P. Perona, and P. Baldi, "Inferring ground truth from subjective labelling of venus images," in *Advances in neural information processing systems*, pp. 1085–1092, 1995.

[32] D. Memmert, "Creativity, expertise, and attention: Exploring their development and their relationships," *Journal of sports sciences*, vol. 29, no. 1, pp. 93–102, 2011.

[33] T. Vestberg, R. Gustafson, L. Maurex, M. Ingvar, and P. Petrovic, "Executive functions predict the success of top-soccer players," *PloS one*, vol. 7, no. 4, p. e34731, 2012.

[34] G. Stratton, *Youth soccer: From science to performance*. Psychology Press, 2004.

[35] I. Csáki, G. Géczi, L. Kassay, D. Déri, L. Révész, Z. Dávid, and J. Bognár, "The new system of the talent development program in hungarian soccer," *Biomedical Human Kinetics*, vol. 6, no. 1, 2014.

[36] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[37] E. Bengtson and D. Roth, "Understanding the value of features for coreference resolution," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 294–303, Association for Computational Linguistics, 2008.

[38] O. Uzuner, A. Bodnari, S. Shen, T. Forbush, J. Pestian, and B. R. South, "Evaluating the state of the art in coreference resolution for electronic medical records," *Journal of the American Medical Informatics Association*, vol. 19, no. 5, pp. 786–791, 2012.

[39] M. Lewandowski, "The language of online sports commentary in a comparative perspective," *lingua posnaniensis*, vol. 54, no. 1, pp. 65–76, 2012.

[40] A. T. Sites, "MS Windows NT kernel description." `https://www.alexa.com/topsites/category/Top/Sports/Soccer`, 2017. [Online; accessed August, 2017].

[41] P. J. Boczkowski and J. A. Ferris, "Multiple media, convergent processes, and divergent products: Organizational innovation in digital media production at a european firm," *The Annals of the American Academy of Political and Social Science*, vol. 597, no. 1, pp. 32–47, 2005.

[42] D. Glez-Peña, A. Lourenço, H. López-Fernández, M. Reboiro-Jato, and F. Fdez-Riverola, "Web scraping technologies in an api world," *Briefings in bioinformatics*, vol. 15, no. 5, pp. 788–797, 2013.

[43] R. Mitchell, *Web scraping with Python: collecting data from the modern web*. " O'Reilly Media, Inc.", 2015.

[44] E. Vargiu and M. Urru, "Exploiting web scraping in a collaborative filtering-based approach to web advertising," *Artificial Intelligence Research*, vol. 2, no. 1, p. 44, 2012.

[45] A. Mehlführer, *Web scraping: A tool evaluation*. na, 2009.

[46] A. Finn, N. Kushmerick, and B. Smyth, "Fact or fiction: Content classification for digital libraries," 2001.

[47] S. Gupta, G. Kaiser, D. Neistadt, and P. Grimm, "Dom-based content extraction of html documents," in *Proceedings of the 12th international conference on World Wide Web*, pp. 207–214, ACM, 2003.

[48] P. M. S. Limited, "Soccer news, live scores, results and transfers goal." `http://www.goal.com/en-us`, 2018. [Online; accessed February, 2018].

[49] R. Lawson, *Web scraping with Python*. Packt Publishing Ltd, 2015.

[50] L. Richardson, "Beautiful soup," *Crummy: The Site*, 2013.

[51] S. Behnel, M. Faassen, and I. Bicking, "lxml: Xml and html with python," 2005.

[52] SoccerRules.org, "Soccer rules soccer fouls."

[53] N. C. A. ASSOCIATION, "MS Windows NT kernel description," 2016. [Online; accessed August, 2017].

[54] T. I. F. A. Board, "MS Windows NT kernel description." http://static-3eb8.kxcdn.com/documents/278/082040_220517_LotG_17_18_FINAL_EN_150dpi.pdf, 2017. [Online; accessed May, 2017].

[55] T. G. . C. KG, "Premier league - conversion rate 17/18 transfermarkt." https://www.transfermarkt.com/premier-league/chancenverwertung/wettbewerb/GB1, 2018.

[56] P. L. (the "Competition"), "Premier league players – overview and stats." https://www.premierleague.com/players, 2018.

[57] W. R. F. Statistics, "Football statistics | soccer statistics." https://www.whoscored.com/Statistics, 2018. [Online; accessed February, 2018].

[58] SoccerSTATS.com, "England - premier league - goal times." http://www.soccerstats.com/timing.asp?league=england, 2018. [Online; accessed February, 2018].

[59] D. C. Brabham, "Crowdsourcing as a model for problem solving: An introduction and cases," *Convergence*, vol. 14, no. 1, pp. 75–90, 2008.

[60] A. Doan, R. Ramakrishnan, and A. Y. Halevy, "Crowdsourcing systems on the world-wide web," *Communications of the ACM*, vol. 54, no. 4, pp. 86–96, 2011.

[61] M. K. Poetz and M. Schreier, "The value of crowdsourcing: can users really compete with professionals in generating new product ideas?," *Journal of product innovation management*, vol. 29, no. 2, pp. 245–256, 2012.

[62] D. C. Brabham, *Crowdsourcing*. Wiley Online Library, 2013.

[63] S. Mohorovičić, "Implementing responsive web design for enhanced web presence," in *Information & Communication Technology Electronics & Microelectronics (MIPRO), 2013 36th International Convention on*, pp. 1206–1210, IEEE, 2013.

[64] M. Hirth, T. Hoßfeld, and P. Tran-Gia, "Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms," *Mathematical and Computer Modelling*, vol. 57, no. 11-12, pp. 2918–2932, 2013.

[65] S. Suri, D. G. Goldstein, and W. A. Mason, "Honesty in an online labor market.," *Human Computation*, vol. 11, no. 11, 2011.

[66] S. Nowak and S. Rüger, "How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation," in *Proceedings of the international conference on Multimedia information retrieval*, pp. 557–566, ACM, 2010.

[67] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[68] J. Randolph, "Online kappa calculator," 2018. [Online; accessed January, 2018].

[69] J. J. Randolph, "Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa.," *Online submission*, 2005.

[70] A.-H. Tan *et al.*, "Text mining: The state of the art and the challenges," in *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, vol. 8, pp. 65–70, sn, 1999.

[71] D. D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," in *European conference on machine learning*, pp. 4–15, Springer, 1998.

[72] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 338–345, Morgan Kaufmann Publishers Inc., 1995.

[73] A. McCallum, K. Nigam, *et al.*, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752, pp. 41–48, Citeseer, 1998.

[74] D. Jurafsky, *Speech & language processing*. Pearson Education India, 2000.

[75] M.-L. Zhang and Z.-H. Zhou, "Ml-knn: A lazy learning approach to multi-label learning," *Pattern recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[76] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd annual meeting on association for computational linguistics*, pp. 363–370, Association for Computational Linguistics, 2005.

[77] W. H. Gomaa and A. A. Fahmy, "A survey of text similarity approaches," *International Journal of Computer Applications*, vol. 68, no. 13, 2013.