computers

MDPI

Article

Machine Learning for Predicting Key Factors to Identify Misinformation in Football Transfer News

Ife Runsewe [1,*], Majid Latifi [2], Mominul Ahsan [2] and Julfikar Haider [3,*]

[1] Foremore B.V., Arthur van Schendellaan 4, 6711DC Ede, The Netherlands
[2] Department of Computer Science, University of York, York YO10 5GH, UK; majid.latifi@york.ac.uk (M.L.); mominul.ahsan2@gmail.com (M.A.)
[3] Department of Engineering, Manchester Metropolitan University, John Dalton Building, Chester Street, Manchester M1 5GD, UK
* Correspondence: iferunsewe@gmail.com (I.R.); j.haider@mmu.ac.uk (J.H.)

Abstract: The spread of misinformation in football transfer news has become a growing concern. To address this challenge, this study introduces a novel approach by employing ensemble learning techniques to identify key factors for predicting such misinformation. The performance of three ensemble learning models, namely Random Forest, AdaBoost, and XGBoost, was analyzed on a dataset of transfer rumours. Natural language processing (NLP) techniques were employed to extract structured data from the text, and the veracity of each rumor was verified using factual transfer data. The study also investigated the relationships between specific features and rumor veracity. Key predictive features such as a player's market value, age, and timing of the transfer window were identified. The Random Forest model outperformed the other two models, achieving a cross-validated accuracy of 95.54%. The top features identified by the model were a player's market value, time to the start/end of the transfer window, and age. The study revealed weak negative relationships between a player's age, time to the start/end of the transfer window, and rumor veracity, suggesting that for older players and times further from the transfer window, rumors are slightly less likely to be true. In contrast, a player's market value did not have a statistically significant relationship with rumor veracity. This study contributes to the existing knowledge of misinformation detection and ensemble learning techniques. Despite some limitations, this study has significant implications for media agencies, football clubs, and fans. By discerning the credibility of transfer news, stakeholders can make informed decisions, reduce the spread of misinformation, and foster a more transparent transfer market.

Keywords: football transfer news; machine learning; prediction; random forest; AdaBoost; XGBoost; natural language processing

Citation: Runsewe, I.; Latifi, M.; Ahsan, M.; Haider, J. Machine Learning for Predicting Key Factors to Identify Misinformation in Football Transfer News. *Computers* **2024**, *13*, 127. https://doi.org/10.3390/computers13060127

Academic Editor: Paolo Bellavista

Received: 19 March 2024
Revised: 11 May 2024
Accepted: 21 May 2024
Published: 23 May 2024

1. Introduction

The landscape of interpretation across various sectors—ranging from e-commerce and law to medicine and media—has been revolutionized by the use of data analytics and artificial intelligence, both of which offer unforeseen access to extensive datasets and innovative techniques for pattern analysis [1]. The problem of misinformation, which is fake or misleading information that spreads unintentionally [2], has gained significant attention in recent years due to its potential to impact public opinion, political decisions, and public health [3]. It is essential to clarify that in the context of this study, "misinformation" refers to incorrect or misleading information about football transfers, whether shared unintentionally or as speculative content. This includes a wide range of information, from rumors and hypothetical statements about potential transfers to news about deals that were under negotiation but did not conclude for various reasons. In the traditional sense, not all incorrect claims may be deliberate misinformation or "fake news", as some may be based on genuine speculation or incomplete information available at the time.

In a report by the University of Baltimore and CHEQ, a cybersecurity firm, it was estimated that the impacts of misinformation cost USD 78 billion to the global economy. Thus, there is a large economic incentive to study this issue [4]. An industry that contributes to that cost is the USD 47 billion revenue-generating football industry and, more specifically, the greatest competition in terms of revenue, the English Premier League (EPL) [5]. In the EPL, transfer news, or news about the sale of one player to another club, offers the best chance for misinformation to occur, as there are many stakeholders involved with different agendas [6]. This news is generated daily [7], and substantial amounts of money are involved in transfers; for example, people spent GBP 2.8 billion on transfers over the 2022/23 season [8]. The news can create unrealistic expectations for fans and clubs, affecting the value of players and clubs and decreasing trust in media organizations [9]. In addition, although the EPL is a private entity, it can be argued that it is in the public interest to reduce misinformation in the EPL, as it contributes a significant amount to the UK economy [10] through the number of viewers it attracts; for example, Reuters found that it reaches 800 million homes in 188 countries [11].

Machine learning (ML) can be used to combat the effects of misinformation on the EPL. ML is a subset of artificial intelligence (AI) that enables computers to adjust their algorithms and predict outcomes without the need for explicitly programmed instructions [12]. A study by Deloitte in 2020 [13] revealed that 67% of companies are currently using ML to increase productivity, and Accenture reported that productivity is expected to increase by 40% due to AI by 2035 [14]. The value of ML is being recognized, and it is a great tool for solving this problem due to its ability to analyze vast amounts of data, detect patterns, and make predictions with a high degree of accuracy. More specifically, ML can be used to assess the likelihood of a news report being identified as misinformation and to highlight the factors contributing to it.

By applying ML techniques to EPL transfer news, it is possible to identify and filter out false or misleading information, ensuring that only accurate news is disseminated to the public. This can help to reduce the economic impact of misinformation on the football industry and restore trust in media organizations [15]. Furthermore, the application of ML in the EPL can serve as a valuable case study for other industries struggling with the problem of misinformation. In this paper, the potential of ML in combating misinformation in the EPL is explored, and a framework for its implementation is proposed.

The primary aim of this study is to identify key factors for predicting misinformation in football transfer news using ensemble learning techniques. To achieve this aim, the study has identified the following research objectives:

- The literature on misinformation detection and ML algorithms in the context of football transfer news is reviewed to identify gaps and limitations.
- A large dataset of football transfer news was collected and preprocessed using natural language processing techniques, and the veracity of each report was verified.
- Suitable ensemble learning methods for detecting misinformation in football transfer news are investigated and selected, considering performance and interpretability.
- The performances of the selected ML algorithms are trained, evaluated, and compared, identifying strengths, weaknesses, and opportunities for improvement.
- This study provides recommendations for future research and the implementation of ML in football and other industries.

This study makes several novel contributions to the field of misinformation detection in football transfer news as follows:

- Factor Analysis for False News:
  - ✓ Identification and understanding of the key factors prevalent in false transfer news.
  - ✓ The potential to guide journalists in publishing more reliable and authentic transfer news.
- Decision-making Tool for Clubs:

- ✓ Informed decisions regarding player acquisitions and sales can be made.
- ✓ Managing the expectations of fans and media more effectively.

- Benchmarking ML algorithms:
  - ✓ Comprehensive analysis of different machine learning algorithms for misinformation detection.
  - ✓ This study provides a foundation for future research aiming to detect misinformation in broader contexts beyond football transfer news.

The remainder of this paper is organized as follows. After this introduction, Section 2 offers an extensive examination of the current body of literature concerning the detection of misinformation and the utilization of machine learning algorithms in the field of sports. Section 3 outlines the research methodology, encompassing data gathering, data preprocessing, and the various machine learning algorithms employed in the study. Section 4 showcases the outcomes generated by these machine learning algorithms and engages in a discourse about these discoveries. Ultimately, in Section 5, the paper is concluded, summarizing the principal findings and presenting suggestions for future research.

2. Recent Advancements

Efforts to address misinformation using machine learning have surged due to the rise of social media [16]. This review focuses on key methodologies, particularly ensemble learning, relevant to detecting misinformation in football transfer news.

2.1. ML Algorithms for Detecting Misinformation

Research by Alghamdi et al. [17] and Chen et al. [18] highlights the increasing efficacy of deep learning techniques in misinformation detection, as they compare classical ML algorithms such as logistic regression, decision trees, and Naïve Bayes with more advanced techniques such as convolutional neural networks (CNNs), bidirectional long short-term memory (Bi-LSTM), and bidirectional gated recurrent units (Bi-GRU). Additionally, this study evaluates two deep learning transformer-based models, BERTbase and RoBERTabase. The results reveal that deep learning techniques, specifically BERTbase, outperform other models across all four datasets, suggesting that employing advanced deep learning models could yield better performance in detecting misinformation [17]. Chen et al. [18] reported an accuracy of 99% in detecting fake statements in long-sentence English texts using deep learning methods, emphasizing their applicability across languages. These findings underscore the potential of employing advanced machine learning models in misinformation detection, especially given the rapid propagation of fake news on platforms such as Twitter, as shown by Liu et al. [19].

2.2. Ensemble Learning for Predicting Misinformation

Ensemble learning, a machine learning method that combines several algorithms to produce more accurate predictions than any single model, has shown increasing prominence in misinformation detection. Hansrajh et al. [20] created a blended model using various machine learning algorithms and found it to be more effective than classical approaches in predicting the veracity of news reports. Similarly, Singh et al. [21] demonstrated that boosting ensemble classifiers such as AdaBoost and XGBoost significantly enhanced performance metrics such as accuracy and F1-score. Ensemble learning was used to greatly affect other industries when Sahil et al. [22] used it to develop a credit card fraud detection model. Models such as Logistic Regression, AdaBoost, Random Forest, bagging, etc., were used to create a model that had 99% accuracy, outperforming all of its base models. These studies underline the contribution of ensemble techniques in misinformation detection and their potential to improve upon classical machine learning methods. However, it is important to note that the effectiveness of these ensemble methods may vary based on dataset selection and algorithmic parameters, indicating a need for further research to validate their general applicability.

2.3. Important Factors in Misinformation Detection

Feature selection has emerged as a crucial element in the optimization of machine learning algorithms for the identification of misinformation. Zhao et al. [23] demonstrated an 85% accuracy rate in health misinformation detection by focusing on behavioral features, which describe an individual's actions and interactions with data, e.g., a feature in the report is the number of messages a user sends on a social network. Buzea et al. [24] emphasized the enhancement of algorithmic performance through the inclusion of various features such as sentiment and irony. Vosoughi et al. [25] explored how emotional factors and the novelty of information play a role in the dissemination of false news. Collectively, these studies indicate that the incorporation of a range of features and an understanding of human behavior can improve the accuracy of misinformation detection. However, it is worth mentioning that these findings may be subject to contextual variables, suggesting the need for further exploratory research.

2.4. Natural Language Processing (NLP) for Structuring Data

Advancements in NLP, notably the use of large language models such as GPT-3, have revolutionized the structuring of unstructured data for misinformation analysis. Dunn et al. demonstrated that complex scientific information could be extracted from GPT-3, albeit cautioning about the risks of data 'hallucination' [26]. Agrawal et al. [27] showcased GPT-3's effectiveness in clinical information extraction, outperforming existing baselines even without domain-specific training. These studies indicate that NLP technologies offer innovative ways to enhance data structuring in misinformation detection, although their reliability may require further validation.

2.5. Predictions in the Sports Industry

Machine learning and NLP have found diverse applications in sports analytics, particularly in predicting player transfers and assessing performance. Kim et al. employed traditional performance metrics such as goals and assists in predicting transfer fees in the EPL [28]. Dimov introduced a rule-based framework designed for human use aimed at identifying different categories of sports misinformation [29]. Cwiklinski et al. [30] utilized ensemble methods such as Random Forest, Naïve Bayes, and AdaBoost for player transfer predictions. Silva employed text analysis on scouting reports, demonstrating the power of fusing text-based and traditional data in predicting NBA player performance [31]. These studies signify a shift toward data-driven decision-making in sports, albeit the need for rigorous data validation remains.

2.6. Critical Discussion of the Literature

While existing studies have made significant strides in using machine learning techniques, particularly deep learning and ensemble methods, for misinformation detection, their applicability to the domain of football transfer news remains underexplored [17,18]. Moreover, the current literature focuses on a broad range of features and human behavior patterns for misinformation detection but lacks a specialized focus on the sports industry [25]. NLP technologies such as large language models (LLMs) have shown promise in data structuring but require further validation for their reliability in specific contexts such as sports [26,27].

The research gap lies in the need for a tailored approach that combines advanced machine learning techniques, feature selection, and NLP technologies for detecting misinformation, specifically in football transfer news. Addressing this gap will contribute to the development of more effective, industry-specific misinformation detection strategies.

3. Research Methodology

In this research, we explore the factors driving football transfer news misinformation and the efficacy of ensemble learning in predicting it. The research objectives involve identifying key misinformation factors and evaluating the performance of an ensemble

learning model. A literature review revealed linguistic-, sentiment-, user behavioral-, and topic-related factors [24,25] as well as the potential of ensemble learning for predicting misinformation and the growing use of LLMs for data preparation [26]; therefore, this study aimed to use similar methods. This study's significance lies in facilitating informed decisions for media, clubs, and fans and curbing false information. The mixed-methods approach, blending quantitative and qualitative data, is exploited to provide comprehensive, robust analysis, deciphering the factors of football transfer news misinformation and generating reliable predictive models.

3.1. Research Design

The research design for this study adopts an innovative exploratory sequential mixed-methods approach to investigate the factors driving football transfer news misinformation and evaluate the efficacy of ensemble learning techniques in predicting it. This design facilitates a comprehensive understanding of the research problem by combining qualitative insights and quantitative data analysis to identify misinformation factors and assess ensemble learning model performance.

Opting solely for a qualitative or quantitative approach would limit the study's scope, as each method provides unique insights. Qualitative approaches uncover themes and patterns through unstructured data analysis [32,33], while quantitative approaches measure and generalize relationships among variables using statistical methods [34]. By integrating both methods, the study captures a richer understanding of the phenomenon.

In the first phase, a qualitative approach is employed to analyze football transfer news articles from sources such as the BBC football gossip column and transfermarkt.com. An LLM is used to extract structured information from unstructured text data, such as player names and clubs mentioned. However, instead of solely relying on an LLM to reveal common themes and patterns related to misinformation, this study leverages the football knowledge of researchers to select factors that might have an impact on the predictability of misinformation in transfer news. The chosen factors include the following:

- Age—the player's age in years.
- Time to the start/end of the transfer window—the number of days until the transfer window starts or if it is in the middle of a transfer window, how many days until the end.
- Market value—the player's market value in millions.
- Source—the news source in which the rumor was initiated.
- Position—the player's position.
- Nationality—the player's nationality.
- Clubs mentioned—the football clubs which are mentioned in the rumor.

Adopting this approach ensures that the research is based on domain-specific knowledge, which improves the quality and relevance of the selected factors. It was also too costly to use the GPT-3 model to assess all of the data and to suggest relevant features, which was the original plan.

The second phase adopts a quantitative approach, applying machine learning algorithms to predict misinformation in football transfer news. The dataset from the qualitative phase is used for training and testing various ML algorithms, with an ensemble learning approach implemented to enhance the prediction performance. This phase emphasizes measuring relationships between variables and generalizing results using statistical methods. The entire pipeline for the data can be seen at a high level in Figure 1.

To ensure data reliability and model robustness, k-fold cross-validation is utilized, partitioning the dataset into k subsets and training the model on k-1 subsets while testing it on the remaining subset. Cycling through each of the k subsets to act as the test set exactly once, the process is repeated k times, thereby providing an impartial assessment of how the model is likely to perform on future data [35].

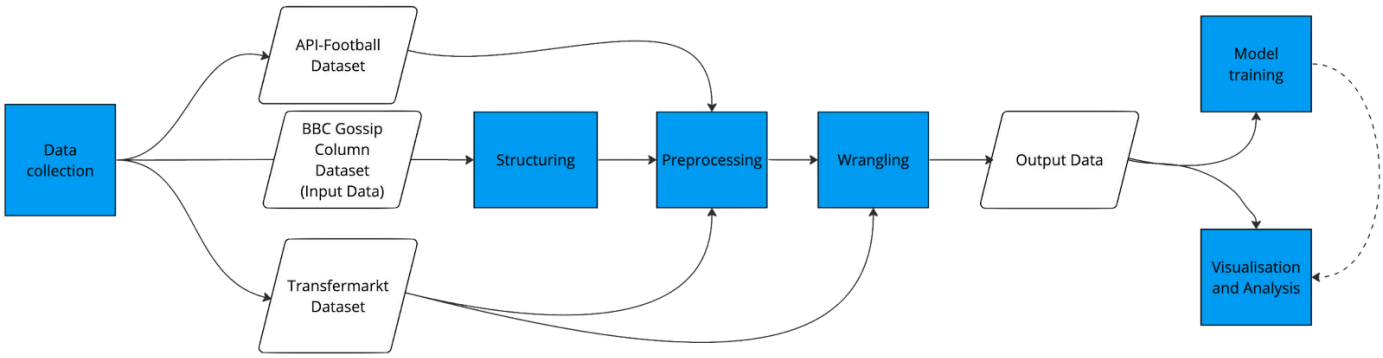

Figure 1. Data pipeline used in this research.

The exploratory sequential mixed-methods design enables a holistic analysis of football transfer news misinformation, supporting key factor identification and the effectiveness of ensemble learning techniques. By incorporating both qualitative and quantitative approaches, this study contributes to a more robust understanding of misinformation in football transfer news, informing future efforts to mitigate its spread.

3.2. Data Collection and Preprocessing

3.2.1. Data Collection

The data collection for this study primarily relies on secondary data sources to analyze the presence of misinformation in football transfer news articles. Secondary data are data previously collected by others for purposes other than one's research, as opposed to primary data, which are collected directly by the researcher for their specific study [36]. Three secondary sources of data are used, the BBC gossip column, transfermarkt.com (a reliable source for factual football transfer data, scores, statistics, and fixtures (available online at https://www.transfermarkt.com/, accessed on 23 May 2023) and API-Football (a restful API for football data, available online at https://www.api-football.com/, accessed on 23 May 2023), which provide a rich dataset of football transfer news and factual transfer data. It is preferable to use secondary data in this study because the research question is aimed at studying misinformation in news reports, and it may take too long to gather factual transfer data without using a secondary data source, as there were 686 transfers in the EPL in the relevant season [37].

Step 1: Data Collection from the BBC Gossip Column

The BBC gossip column serves as an aggregator of football transfer news from various news outlets, presenting a daily summary of the latest rumors and gossip related to player transfers [38,39]. This study focused on the 2021/22 season, during which 304 days of gossip columns were extracted from 1 February 2021 to 31 January 2022. These data cover 3 transfer windows over a whole year, producing 5982 lines of football transfer news. A football transfer window refers to a period during which football clubs are allowed to transfer players between clubs [40].

A Python script is used to scrape the BBC gossip column, and the results are stored in a CSV file. Each line in the CSV represents a separate news report from the BBC gossip columns.

Step 2: Data Structuring Using GPT-3

Subsequently, the GPT-3 language model is used to structure the data in the CSV, extracting features such as player names and football clubs [38]. The additional factors suggested by the study, such as the player's age and nationality, are extracted from the API-Football dataset and the transfermarkt.com dataset.

Step 3: Data Verification with Transfermarkt and API-Football

With the data structured, the study verifies whether each transfer actually occurred by referring to transfermarkt.com, a reliable source for factual football transfer data [39]. Data wrangling techniques are employed to match the dataset from transfermarkt.com with the structured BBC gossip column data using factors such as player names and football clubs [37]. The study scrapes transfermarkt.com and uses the API from API-Football to retrieve EPL football squads, which covers relevant transfer windows.

Step 4: Labeling Data as True or False

By comparing the data from the two sources, the study labels each news report as true or false using an automated matching process indicating misinformation in football transfer news articles. This process is crucial for training and evaluating the machine learning algorithms used later in the research.

3.2.2. Data Preprocessing

The dataset from the BBC gossip column underwent a series of preprocessing steps to ensure its suitability for subsequent analysis. These steps are collectively illustrated in Figure 2. The BBC gossip column contains numerous stories not only about football players but also about managers. To address this issue, the data are cleansed by cross-referencing the football squad datasets used in this study to verify the presence of the mentioned individuals. If they are not found within the datasets, the corresponding news items are discarded.

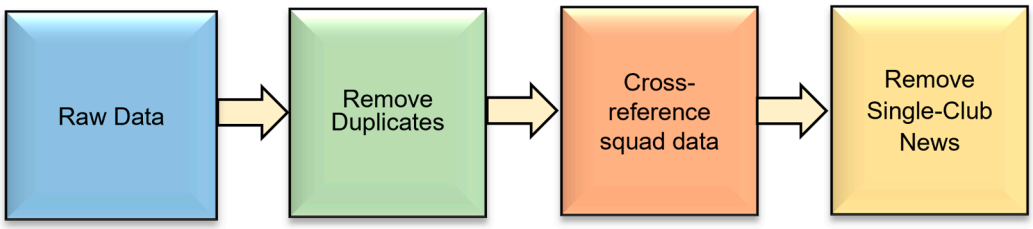

Figure 2. Data preprocessing steps.

Furthermore, certain news items may mention only one club, making it difficult to determine the nature of the news, as it could involve rumors about a player's contract with their current club. To handle such cases, this study implements a rule to discard rumors from the dataset if only one club is mentioned. By implementing these steps, the risk of false positives and negatives is reduced. False positives occur when something is marked as present but not present, while false negatives occur when something is marked as absent but is actually present [41].

Dealing with missing or null values in a dataset is a crucial aspect of data preprocessing known as missing value handling. In the dataset, missing values may arise due to the unavailability of data such as market value, nationality, position, and age from factual sources such as transfermarket.com and API-Football. Additionally, missing data may occur if the player's name is not found in the transfermarkt.com data, rendering the veracity column empty for the corresponding data row.

If factual data are unavailable, this study endeavors to retrieve data from the raw text of rumors using pattern matching and NLP methods like Named Entity Recognition (NER), which extracts entities from text such as people, locations, and organizations [42]. In the event of persistent missing data, the model employs mean imputation, where the mean of each column is used to replace the missing values. Mean imputation may introduce bias in the results [43], but the technique is still chosen since the maximum amount of missing data in this study is 21%, and research shows that unbiased results can still be obtained even for up to 90% missing data [44]. Therefore, mean imputation is a worthwhile selection due to its simplicity and ease of implementation.

After the data are preprocessed, the data are reduced from 5982 to 2480 news units for use in model training and feature analysis. It must be noted that there may be multiple

news units for a line of football transfer news because it may contain multiple rumors; for example, it might be mentioned that a club is trying to buy 2 players. Once the data are labeled true or false, we are left with 542 (21.77%) true rumors versus 1948 (78.23%) false rumors, which presents an obvious class imbalance.

3.2.3. Using GPT-3 for Data Structuring

In this study, the feature extraction process leverages the advanced capabilities of the large language model (LLM) GPT-3 to transform textual data into structured data for machine learning models [45]. Developed by OpenAI, GPT-3 is a cutting-edge LLM with 175 billion parameters, enabling it to learn intricate language patterns, context, and semantic relationships [45]. As an autoregressive model, it predicts the next word in a sequence by considering previous words, resulting in coherent and context-aware text generation [45].

GPT-3 has been successfully applied to various applications, such as translation, summarization, and question answering [45]. Although recent advancements include InstructGPT, GPT-3 remains the better choice for feature extraction in this study's context according to Agrawal [27].

To use GPT-3 for feature extraction and data structuring in this study, the raw text data from the BBC gossip column dataset are fed into GPT-3 as input. Using GPT-3 for this purpose is a form of transfer learning [46]. In this case, a model has been trained on vast amounts of data, and the study is used to enhance feature extraction related to the specific study of football transfer news. This not only speeds up the feature extraction process but also improves the accuracy with which these features are identified, which is critical for subsequent misinformation analysis. This is specifically a form of feature-based transfer learning.

Proper input formatting is crucial for maximizing the effectiveness of an LLM. GPT-3 processes the input data, identifying relevant information and features within the text, such as player names and team names. The data are then structured into a format suitable for machine learning models. These structured data are produced by GPT-3 and can be used to match factual datasets and as input for machine learning models assessing the veracity of football transfer news. The feature extraction process used in this study is visually summarized in Figure 3.

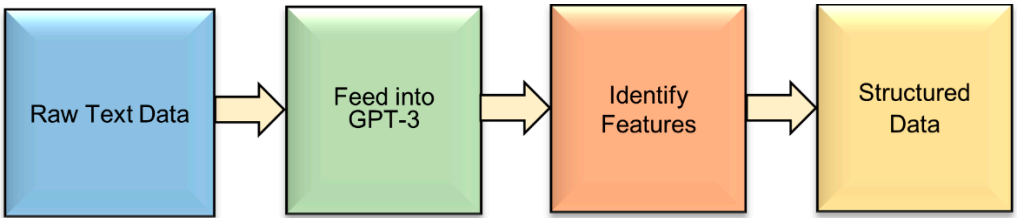

Figure 3. Feature extraction steps.

The authors in [21] removed stopwords such as 'I', 'me', 'we', and others, which do not add significant semantic meaning and can normally be removed to streamline the data processing steps. However, due to the use of GPT-3 in this study, stopwords are left alone because they provide context for the LLM, making it more effective at organizing the data.

3.3. Data Analysis

The data analysis section of the research methodology details the methods and techniques employed to answer the research question and meet the objectives. The data analysis process involves developing and evaluating ensemble learning models for predicting misinformation in football transfer news.

3.3.1. Model Development

Ensemble learning techniques, including Random Forest, AdaBoost, and XGBoost, are employed to develop a model for predicting misinformation in football transfer news. These techniques were chosen based on their proven effectiveness in the literature review and their ability to improve prediction performance by combining multiple base models [21]. These methods can be categorized into two main approaches: bagging and boosting.

Bagging involves constructing multiple decision trees and aggregating their predictions to reduce variance and improve overall performance [21]. The Random Forest algorithm is an example of this technique.

Given a training sample $D_n = ((X_1, Y_1), ..., (X_n, Y_n))$, the Random Forest predictor comprises a collection of M randomized regression trees. Each tree provides a prediction $m_n(x; \theta_n, D_n)$, where $\theta_1, \ldots, \theta_m$ are independent random variables [47]. These variables are utilized for resampling the training dataset and selecting successive directions for node splitting. The forest estimate is given by Equation (1).

$$m_{\infty,n}(x) = E_\theta[m_n(x; \theta, D_n)] \tag{1}$$

Boosting iteratively adjusts the weights of misclassified instances and combines weak learners to produce a strong classifier [21]. AdaBoost and XGBoost are examples of boosting techniques and have proven successful in predicting misinformation [21].

In AdaBoost, the final prediction $F(x)$ is a weighted sum of the weak learners $f_i(x)$ [48], as shown in Formula 3.2, where α_i is the weight assigned to the i^{th} weak learner based on its error rate.

$$F(x) = \sum_{i=1}^{T} a_i f_i(x) \tag{2}$$

In XGBoost, the algorithm optimizes an objective function Obj by combining a loss term $l(y_i, f(x_i))$, which minimizes prediction errors, and a regularization term $\Omega(f)$ to avoid model complexity and overfitting. Together, these form a strong classifier by effectively aggregating multiple weak learners [49]. This is represented by Equation (3).

$$Obj = \sum_{i=1}^{T} l(y_{i,} f(x_i)) + \Omega(f) \tag{3}$$

In addition to the ensemble learning methods mentioned earlier, randomness is widely applied in research to create variation and avoid overfitting. For instance, during the model development phase, a randomness parameter is utilized for model initialization, ensuring consistent random number sequences for reproducible results [50]. Randomness is also employed in cross-validation and preprocessing. In the data preprocessing phase, randomness is used to shuffle the training examples' order, reducing potential bias from their arrangement.

3.3.2. Model Evaluation

The ensemble learning model's performance is evaluated by employing a range of assessment criteria, including accuracy, precision, recall, and the F1-score. In the subsequent equations, TP signifies the count of correctly identified positive instances, TN stands for the count of correctly identified negative instances, FP corresponds to the count of incorrectly identified positive instances, and FN denotes the count of incorrectly identified negative instances.

Accuracy: The ratio of accurately classified instances to total instances, which offers a broad gauge of the model's effectiveness, is computed using Equation (4).

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{4}$$

Precision: The precision ratio of true positive instances to the sum of true positive and false positive instances indicates the model's capacity to precisely identify pertinent instances. It is calculated using Equation (5).

$$Precision = \frac{TP}{(TP + FP)} \tag{5}$$

Recall: The ratio of true positive instances to the sum of true positive and false negative instances illustrating the model's ability to detect all pertinent instances. It is calculated using Equation (6).

$$Recall = \frac{TP}{(TP + FN)} \tag{6}$$

F1-score: The harmonic mean of precision and recall offers a well-balanced assessment of the model's performance, which is particularly valuable when handling imbalanced datasets. It is calculated using Equation (7).

$$F1Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \tag{7}$$

As previously mentioned, cross-validation ensures the model's reliability and generalizability. After cross-validation, the study calculates a cross-validated accuracy score, offering a more dependable and unbiased performance estimate than a single test–train process's accuracy score.

3.4. Implementation Details

This study implemented a comprehensive methodology to collect, process, and analyze data related to football transfer rumors. The implementation details for each stage of the research were seamlessly integrated to ensure a coherent workflow. The source code and additional resources for this study can be found on GitHub at BBC Gossip Column Predictor, showing how the methodologies were applied practically.

Python was used for data collection, processing, and analysis in this study. As a versatile language, Python can handle various tasks, including scientific and numeric computing. It also offers numerous libraries, some of which were used in this research and are discussed further.

This research collected data from various sources using APIs and web scraping techniques. Google's custom search (JSON) API was used to search for relevant articles, and API-Football was used to gather player and team information. Additionally, the Beautiful Soup library was employed to scrape data from websites such as Transfermarkt and BBC Sport.

This study used OpenAI's GPT-3 to extract and structure relevant data from the raw text, focusing on identifying player names and clubs mentioned while excluding international teams and football managers. GPT-3 was prompted to generate a structured JSON response, which was parsed and incorporated into a new CSV file containing the processed information.

For data manipulation and cleaning, this research employed the Pandas library. It leveraged the Python library, theFuzz, for approximate string matching, which uses the Levenshtein distance algorithm that calculates the differences between sequences [32]. The NER Locationtagger library was used for extracting locations from the raw text, which helped assign a nationality to a player if it could not be found in other datasets. If the age and position of the player were not available in other datasets, regex was employed to extract this information from the raw text. The market value of a player was extracted from the datasets and cleaned by converting it to a numerical value and removing currency symbols and units.

To further enrich the feature set and capture any emotional nuances from the transfer rumors, sentiment analysis was incorporated using the sentiment module in the NLTK

library (for sentiment analysis, this study utilizes the NLTK library, a leading platform for building Python programs to work with human language data, https://www.nltk.org/, accessed on 25 May 2023). This involved calculating sentiment scores for each of the text extracts to quantify the emotional tone, ranging from −1 (most negative) to 1 (most positive). These scores help us to understand how sentiment influences the veracity of rumors, as shown in previous studies [24,25].

Another critical preprocessing step involved converting the 'clubs_mentioned' column into a list of clubs and mapping them to the corresponding Transfermarkt club names. One-hot encoding was applied to categorical columns such as 'clubs_mentioned', 'nationality', 'position', and 'source'. These encoded columns were then concatenated with the original data.

To analyze the cleaned and structured data, this research used the Scikit-learn library, a commonly used ML Python library for predictive data analytics. Machine learning algorithms, such as RandomForestClassifier, AdaBoostClassifier, and XGBoostClassifier, were applied to predict the veracity of the transfer rumors.

To ensure a thorough evaluation, a 5-fold cross-validation approach was employed. By dividing the data into 5 equal parts, each fold was used as a testing set once, while the remaining folds were used as the training set to assess the model's performance using accuracy scores and classification reports. Furthermore, random oversampling was used to maintain the class distribution across folds, ensuring a more reliable evaluation. This method involved artificially increasing the number of instances in the smaller group by duplicating them, thus equalizing the class distribution before training. Finally, the most important features that contributed to the model's predictions were identified, providing insights into the factors that played a significant role in the veracity of transfer rumors.

In addition to the aforementioned methodologies, various visualization techniques and statistical analysis methods were utilized to present and interpret the results of the machine learning models. Boxplots and bar charts were generated to illustrate the distribution of classes (true and false) across the features, providing a clear visual representation of the data. Confusion matrices and classification reports were used to assess the performance of each model, providing insights into the accuracy and precision of the predictions. To delve deeper into the relationships between variables, point-biserial correlation coefficients were calculated to measure the strength and direction of associations between continuous variables and veracity.

This study provided an accessible and interactive way for users to run and engage with the research pipeline. By executing the pipeline.py script, users can navigate through the pipeline's different stages, such as data collection, structuring, preprocessing, wrangling, model training and evaluation, and visualization and analysis. Users had the option to run all steps, run steps interactively, or run a single step, offering flexibility in how they interacted with the research. This approach enabled users to better understand the research's inner workings and gain insights into the factors influencing football transfer rumors' veracity. Figure 4 displays an example of the interactive tool used to run the pipeline.

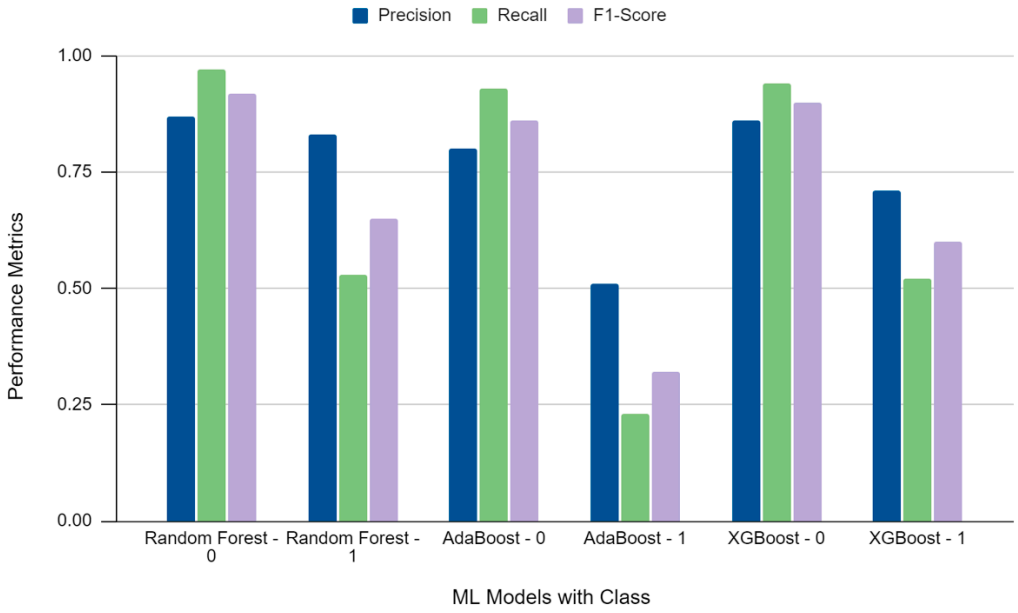

Figure 4. Classification output.

4. Results and Analysis

This study aimed to evaluate the performance of three ensemble learning models, namely, Random Forest, AdaBoost, and XGBoost, in predicting the veracity of football transfer rumors. Furthermore, the relationships between specific features and rumor veracity were investigated. The results are based on the cross-validated accuracy, accuracy of the test set, precision, recall, and F1-score for each model. The relationships between a player's age, time to the start/end of the transfer window, and market value and veracity were examined using correlation coefficients and p values, while categorical variables such as nationality, position, and source were analyzed based on their proportion of true rumors.

4.1. Model Performance

The Random Forest model achieved a cross-validated accuracy of 86.89% and a test set accuracy of 86.59%. It showed high precision and recall for class 0 (false rumors) and moderate precision and recall for class 1 (true rumors). The results indicate that the model effectively predicted false rumors and had a reasonable success rate in identifying true rumors, although it struggled to recognize all of them. The results in general indicate that while the model can be very effective in filtering misinformation, it is still limited in identifying true rumors even if it has the best performance (considering precision, recall, and F1-score) of all three models for class 1. The results for all the examined ML models are shown in Figures 4 and 5.

AdaBoost had a cross-validated accuracy of 80.04% and a test set accuracy of 76.81%. The model's performance for class 0 was reasonable, but it had more false positives than the Random Forest model did. For class 1, the model's precision and recall were lower than Random Forest's, indicating difficulty in identifying true rumors and a greater number of both false positives and false negatives. The model's low F1-score suggests that it has difficulty accurately identifying true rumors, which may result in genuine transfer-related information being overlooked. This could lead to a failure to take appropriate action on actual transfers if they are not correctly identified. XGBoost achieved a cross-validated accuracy of 84.81% and a test set accuracy of 83.57%. The model's performance in predicting false rumors was close to that of the Random Forest model, but it had more false positives for true rumors. This suggests that the model identified a fair number of true rumors but encountered challenges similar to those faced by the Random Forest model. This observation is supported by the model's marginally lower F1-score of 0.6 compared to the Random Forest's score of 0.65. This discrepancy may present challenges, particularly in

studies where accurately identifying the correct number of true rumors is crucial. Among the three models, Random Forest exhibited the highest overall performance, with both the highest cross-validated and test set accuracies, and although all the models produced mediocre results for true predictions, Random Forest still performed the best. This was in contrast to the findings of Singh, who suggested that AdaBoost and XGBoost were superior in detecting misinformation but for different contexts [21].

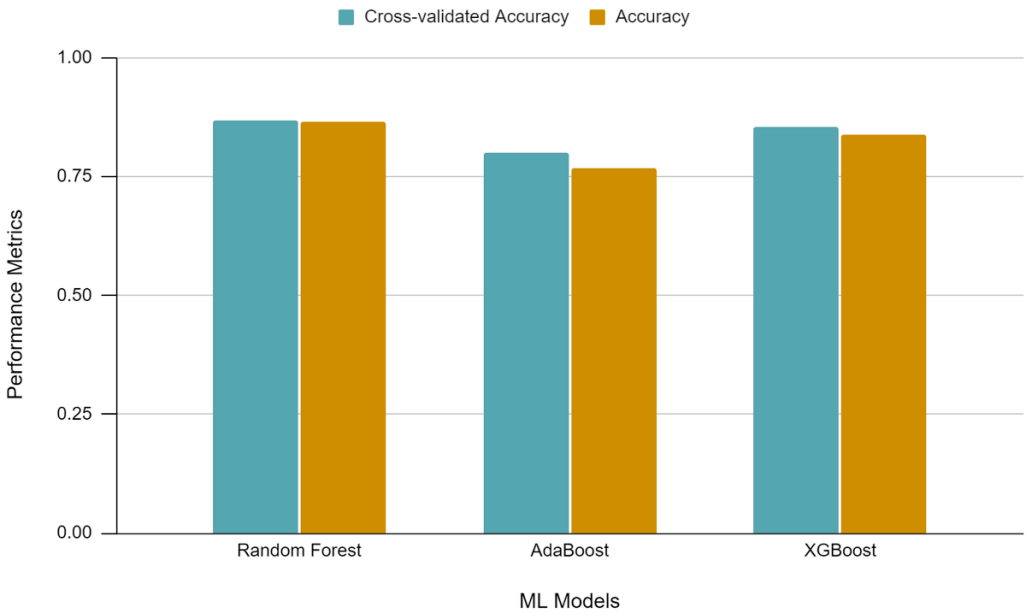

Figure 5. Accuracy outcome.

The classification reports for the three models seemed to suggest that there could be an imbalance in the dataset due to lower precision, recall, and F1-scores in class 1 for all three models. This could only be confirmed by further analysis of the dataset by analyzing the class distribution. The class distribution confirmed that there was an imbalance, with approximately 80% of class 0 and 20% of class 1 instances, as shown in Figure 6.

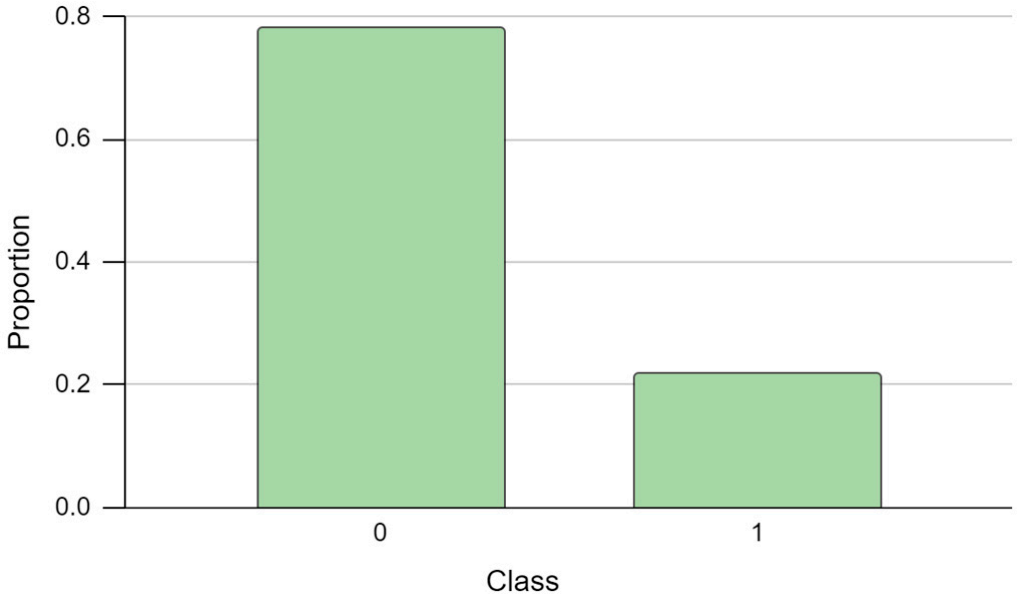

Figure 6. Class distribution in the training set (class 0: false rumors and class 1: true rumors).

After confirming the class imbalance, this study balanced the dataset using Random Oversampling (ROS). ROS duplicates existing minority class samples; in this case, class 1. After applying ROS, the models showed notable improvement in the minority class (class 1) performance. For instance, the Random Forest model's class 1 recall increased from 0.5345 to 0.6250, and the F1-score increased from 0.6509 to 0.6824, while the precision slightly decreased from 0.8322 to 0.7513. However, the test set accuracy slightly decreased for the Random Forest and XGBoost models and significantly decreased for the AdaBoost model, which could suggest potential overfitting or that ROS might not be the most suitable choice. However, Random Forest predicted the correct results 95.54% of the time, suggesting that the model was very effective overall after correcting imbalances. The effects of oversampling the data on all the models are shown in Figures 7 and 8.

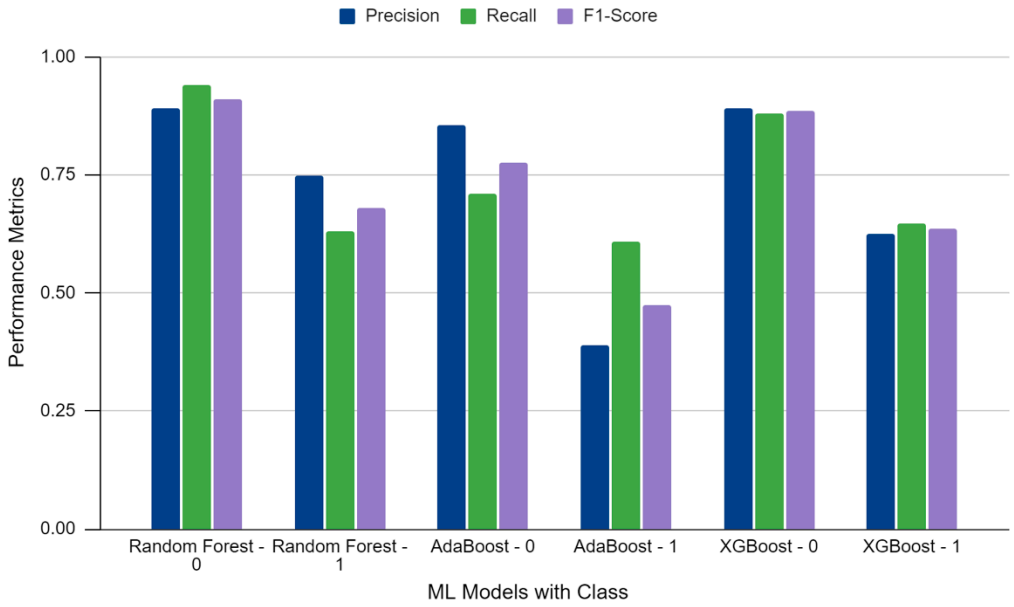

Figure 7. Classification results after ROS.

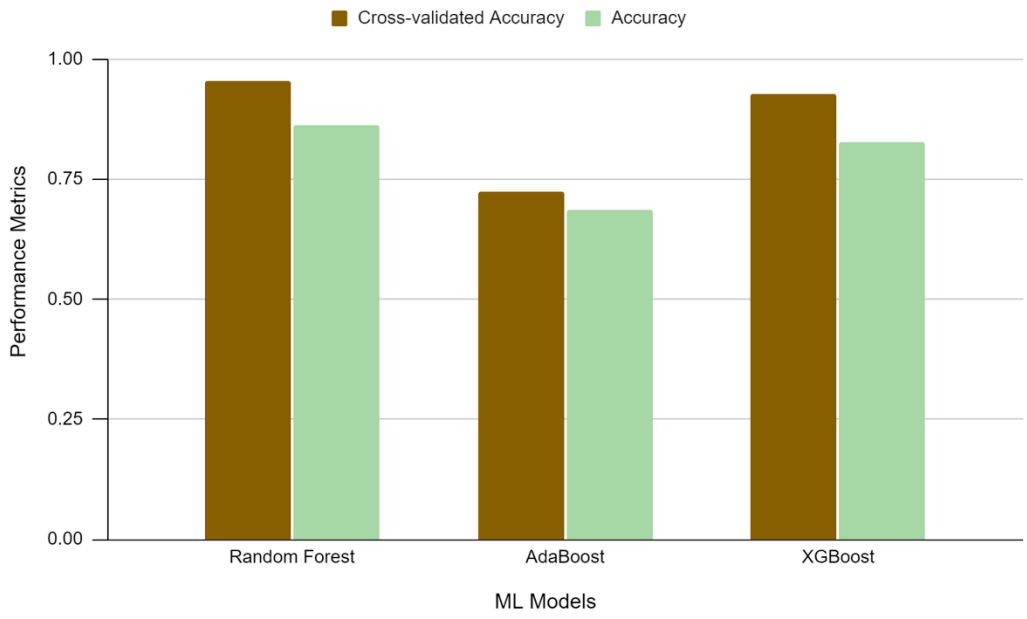

Figure 8. Accuracy results after ROS.

4.2. Important Features and Relationships

This study analyzed a total of 487 features to gain further insights into the factors that influence the veracity of football transfer rumors. The feature importance scores were calculated based on how much each feature contributed to improving the model's performance. A higher value suggests that the feature has a stronger impact on the model's predictions. As the RF model produced the best results from the evaluation, it was the only model used to obtain the most important features.

The main features were determined by our best-performing model, Random Forest. The Random Forest model suggested that market value, time to the start/end of the transfer window, and a player's age were the most important factors, with importance scores of 8.54%, 8.52%, and 7.61%, respectively. The significance of these features can be seen in Figure 9. To assess the strength of the relationship between the features and rumor veracity, the research calculated the correlation coefficient, specifically the point-biserial correlation. To determine whether these correlations were statistically significant, p values were calculated to measure the probability of this being the case.

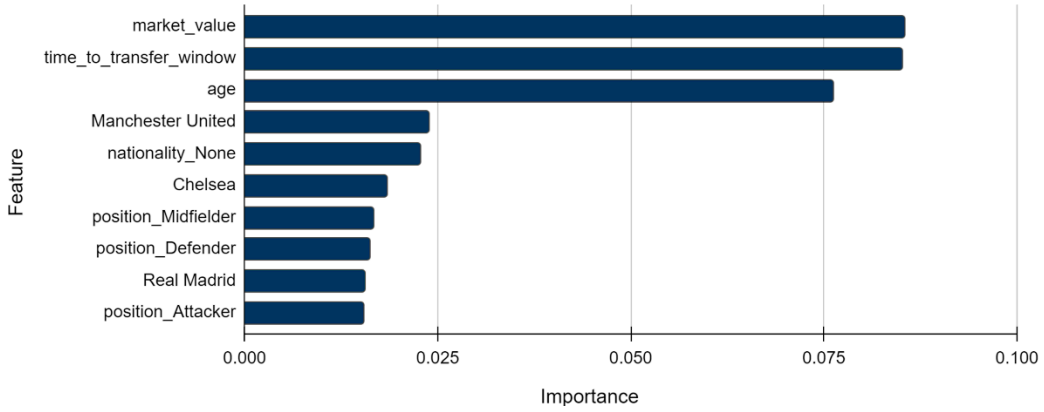

Figure 9. Random Forest feature importance.

For a player's age and time to start/end of a transfer window, both variables exhibited weak negative associations with rumor veracity (correlations of −0.0832 and −0.0917, respectively), and p values less than 0.05 indicated statistically significant relationships. This implies that as the age of a player increases and as the time to the transfer window increases, the likelihood of the rumor being true decreases slightly. The boxplots for age and time to start/end of a transfer window, time_to_transfer_window, (Figures 10 and 11) show lower mean values for true rumors than for false rumors, supporting the weak negative relationships.

Regarding market value, the correlation of 0.0096 reveals a weak positive association with rumor veracity. Nevertheless, the p value exceeded 0.05, implying that there was no statistically significant relationship between the two variables. This suggests that a player's market value does not considerably impact the accuracy of transfer rumors. The boxplot for market value (Figure 12) displays comparable mean market values for both true and false rumors, signifying that market value does not play a crucial role in determining the authenticity of a rumor.

An essential critique of machine learning (ML) is its depiction as a black box, wherein we can obtain effective models without comprehending their inner workings. Addressing this challenge, the utilization of the SHapley Additive exPlanation (SHAP) [51] signifies a significant stride in ML model interpretation. The cutting-edge Python library is commonly integrated into the feature engineering phase of ML projects. This facilitates an in-depth understanding of the outcomes produced by ML models [52]. SHAP value analysis was conducted to interpret the Random Forest model's use of features in predicting the veracity of football transfer rumors. SHAP is a method used in machine learning for understanding the impact of each feature on the model's predictions. Calculating SHAP values precisely

presents challenges. Nonetheless, through amalgamating insights from existing additive feature attribution methods, we can estimate them [51]. Given the extensive total of 487 features, attempting to visualize each feature would prove impractical. Thus, we selected the seven most crucial features for examination. Additionally, the presence of categorical features further complicates their interpretation within SHAP value diagrams, limiting the insights they can provide. Consequently, our focus on these selected features for further analysis is in line with our initial findings and offers the most meaningful approach. Through scatter plots, we uncover how specific features influence the predictions of our model.

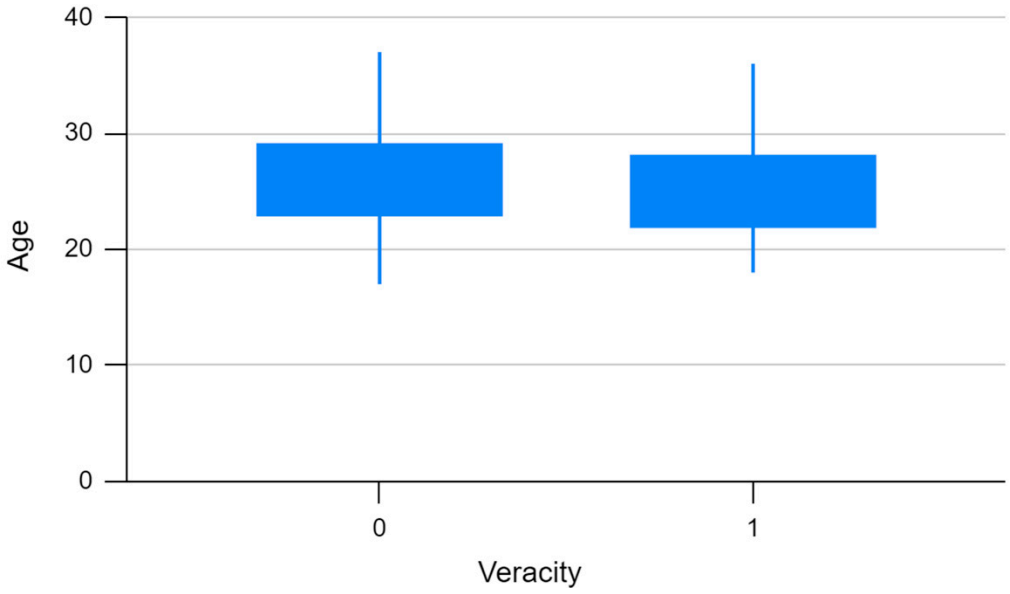

Figure 10. Distribution of age against veracity.

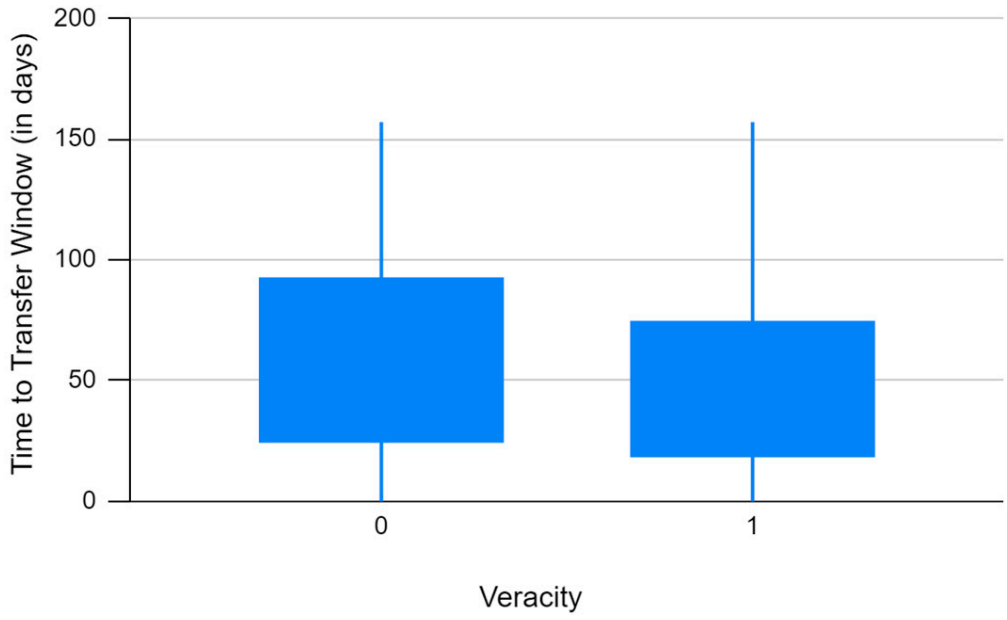

Figure 11. Distribution of time_to_transfer_window against veracity.

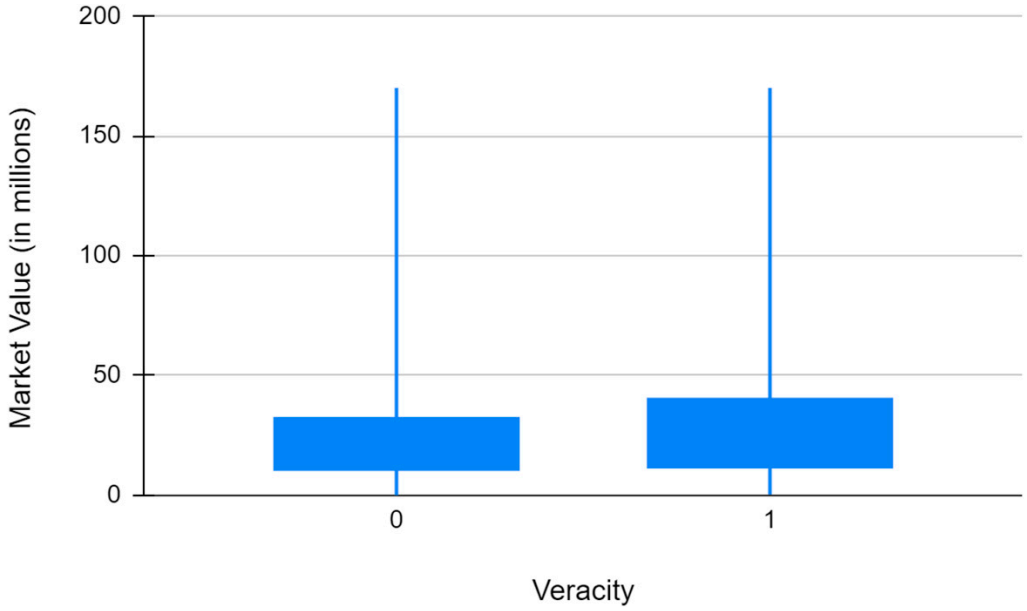

Figure 12. Distribution of market value against veracity.

The *Y*-axis depicts feature names arranged in descending order of importance, with the most significant features positioned at the top. On the *X*-axis, the SHAP value illustrates the extent of the change in log odds. Each point's color on the graph corresponds to the respective feature value, with red indicating high values and blue indicating low values. Every point represents a row of data from the original dataset.

As seen in Figure 13, for false rumors, when the SHAP value of a feature "Market Value" is +0.3 for a prediction, it means that, on average, an increase in market value tends to increase the model's prediction by 0.3 units. Notably, our results underscore the significant impact of "Position," "Market Value," "Nationality," and "Age" in determining the presence of false rumors.

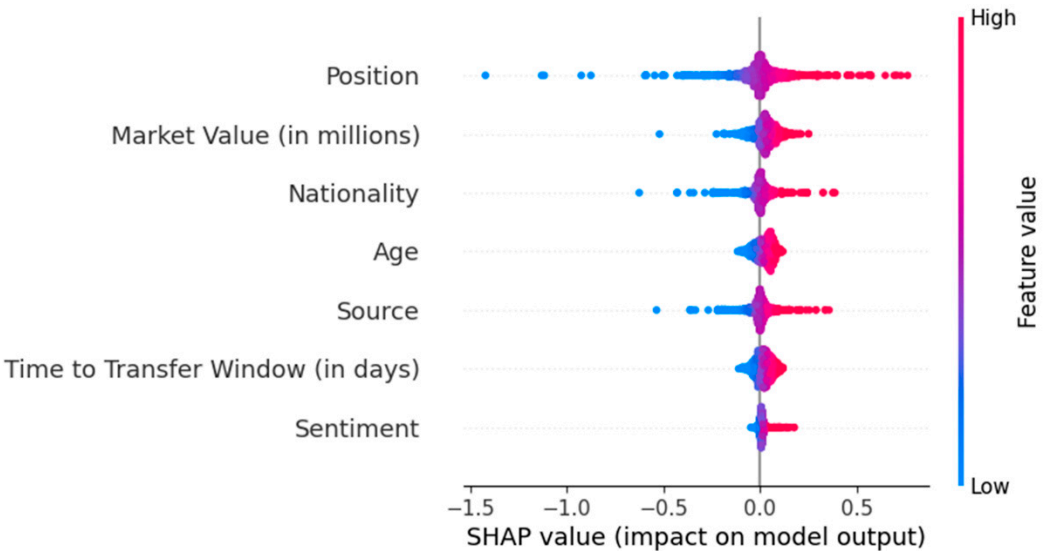

Figure 13. Distribution of SHAP values for class 0 (false rumors) across selected features.

On the other hand, in Figure 14 concerning true rumors, a SHAP value of −0.2 for "Market Value" indicates that, on average, a lower market value tends to decrease the model's prediction by 0.2 units. Notably, Figure 14 reveals a shift in the positions of the features. Nevertheless, three previously mentioned features remain paramount: "Market

Value," "Age," and "Position." Additionally, "Time to Transfer Window" emerges as a noteworthy feature in the predictions. Both age and time to transfer exhibit significant variability in their SHAP values across the axis, underscoring their nuanced impact on rumor veracity, which may be influenced by other factors. This variability implies that neither factor in isolation consistently determines the truthfulness of a rumor, as depicted in Figures 13 and 14.

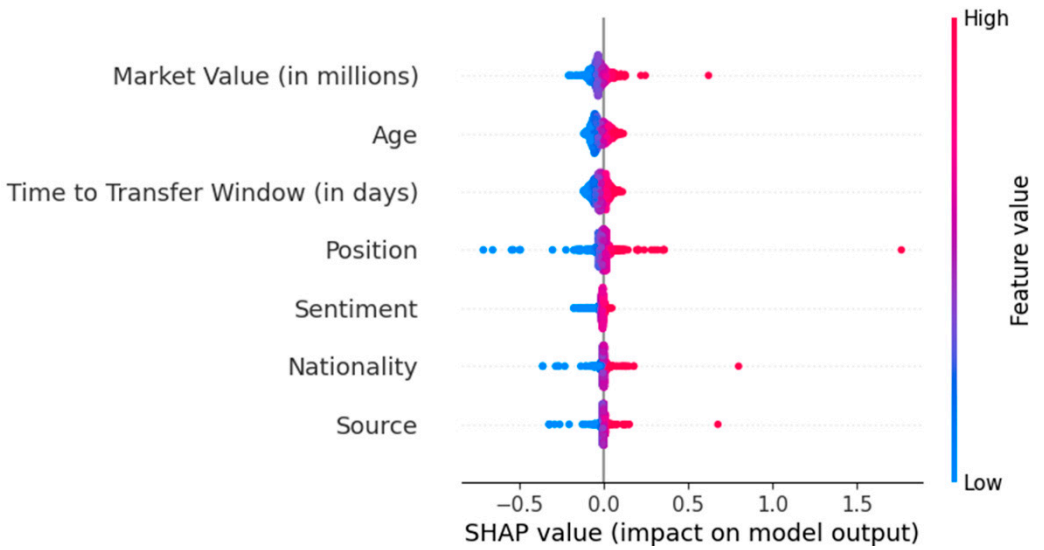

Figure 14. Distribution of SHAP values for class 1 (true rumors) across selected features.

Conversely, the market value of players exhibits a more uniform influence on predicting rumors as true, as evidenced by positive SHAP values. This consistency reinforces our findings that proximity to the transfer window enhances the accuracy of rumors. These scatter plots provide more evidence in support of the explanations laid out earlier in this section (see Figure 9).

Notably, when sentiment scores are added to the feature set, the feature becomes a significant predictor in the Random Forest model, with an importance score of 5.14%, but it is still behind the three top aforementioned predictors. It marginally enhanced the cross-validated accuracy, elevating it to 95.59% from the previously recorded 95.54%. There is a slight improvement in the precision and recall of true rumors (class 1), while the model's initial weak performance on false rumors persists. These adjustments signify a noteworthy, albeit subtle, advancement in identifying true rumors.

For categorical variables such as nationality, position, and source, the research calculated the proportion of true rumors for each category to identify any potential relationships with rumor veracity. To ensure a robust analysis, the study only considered categories with a rumor count greater than ten. This threshold was chosen to avoid drawing conclusions from categories with a small sample size, which could lead to unreliable or misleading results due to the influence of outliers or random variations.

A breakdown of the proportions of true rumors across categorical data is shown in Figures 15–17. A striking contrast emerges between French players, who lead the pack with 32.66% true rumors, and their counterparts from Côte d'Ivoire, who trail at a mere 4.35%. Surprisingly, despite the majority of news sources being UK-based, English players only had a 12.56% chance of a rumor being accurate.

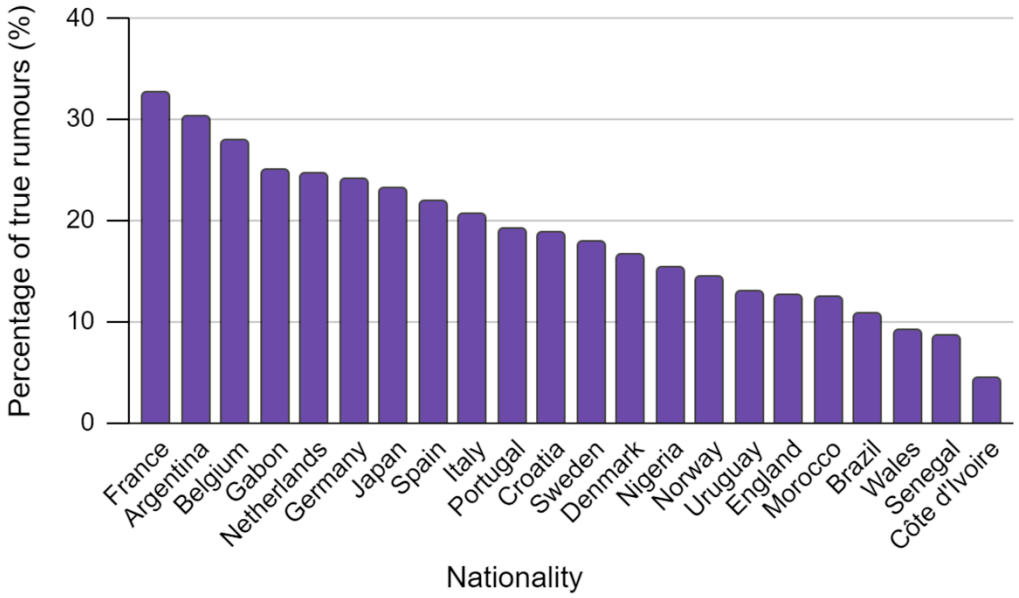

Figure 15. Proportion of true rumors per nationality.

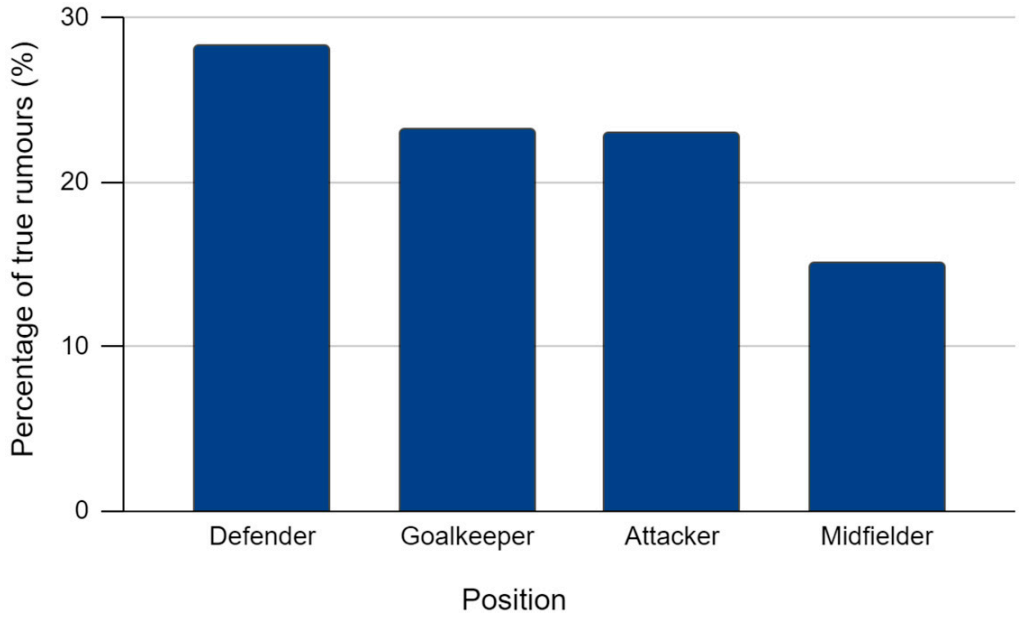

Figure 16. Proportion of true rumors per player position.

When considering positions, defenders topped the charts with 28.25% true rumors, while midfielders lagged behind at 15.07%. Among news sources, www.chroniclelive.co.uk outperformed the others with a 41.18% rate of true rumors, whereas www.fichajes.net languished at the bottom with 4.17%. Intriguingly, Twitter, often blamed for propagating misinformation, registered a relatively high percentage of true rumors, contradicting some research findings [2,25]. These insights underscore the substantial impact that source credibility has on rumor accuracy, with some sources proving more dependable than others.

This study evaluated the performance of three ensemble learning models—Random Forest, AdaBoost, and XGBoost—in predicting the veracity of football transfer rumors. Among these models, Random Forest demonstrated the highest overall performance. Furthermore, the analysis revealed that age, transfer window timeline, and market value were among the most important features for predicting rumor veracity. A player's age and transfer window timeline showed weak but statistically significant negative relationships

with rumor veracity, while a player's market value did not have a significant influence. In the case of categorical variables such as nationality, position, and source, this study found varying proportions of true rumors across different categories, indicating that specific variables could influence the probability of a rumor being true.

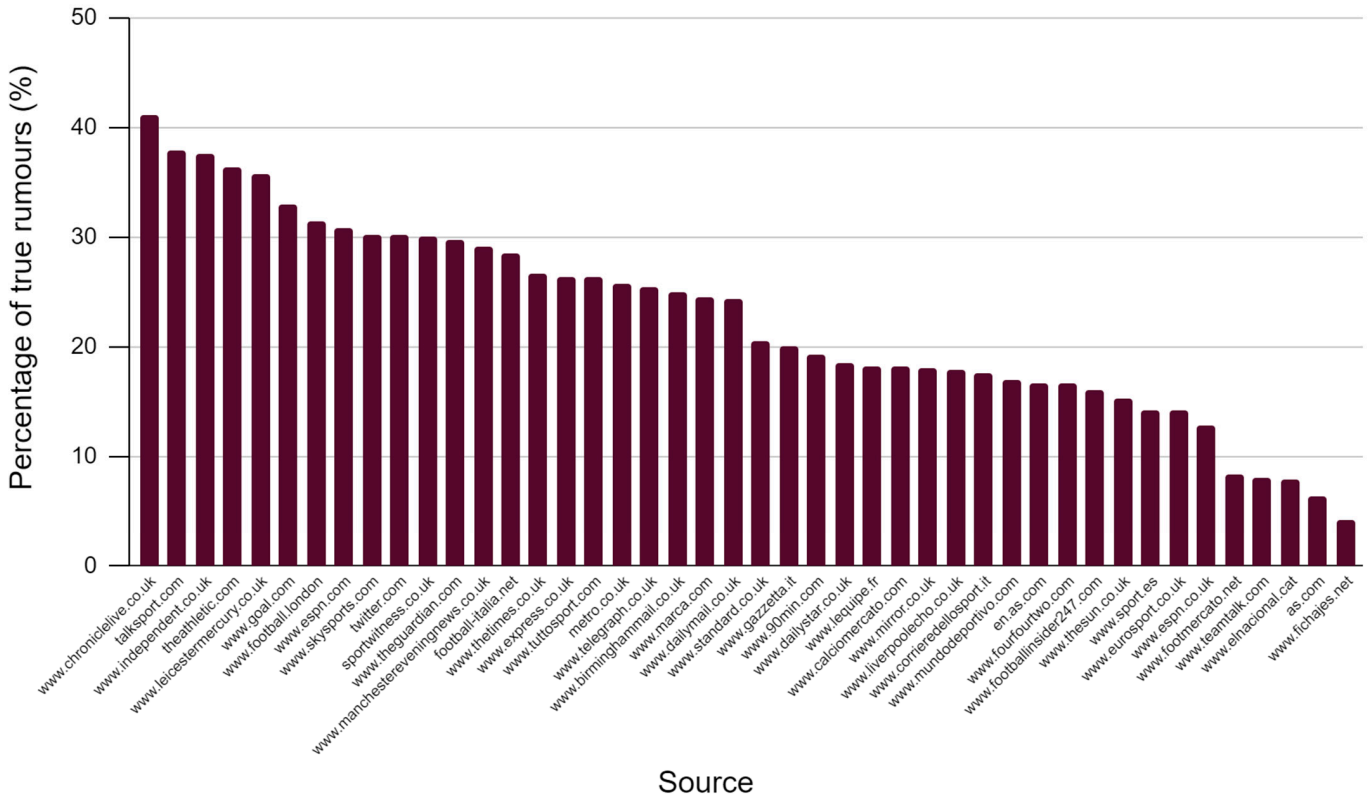

Figure 17. Proportion of true rumors per source.

4.3. Discussion

4.3.1. Strengths

In alignment with existing research, especially the studies by Hansrajh et al. [20] and Singh et al. [21], the use of ensemble learning techniques in this study exhibited robust capabilities for detecting misinformation. In particular, the use of Random Forest as a top-performing model not only adds a new dimension to the ensemble learning discussion but also provides actionable insights. This approach is particularly beneficial in the misinformation industry, where insights into factors can enhance technology and policy changes [3,4]. However, the absence of deep learning exploration leaves open questions, particularly given the high performance of such techniques, as shown in [17,18].

A noteworthy divergence occurs in feature selection. Unlike prevalent research that focuses on behavioral or textual features [23,24], the significance of factors such as a player's age and time to the start/end of the transfer window emerged. These novel features not only create the ability for more nuanced models but also hold real-world significance by enabling stakeholders such as news agencies to develop misinformation detection algorithms that can utilize this context. The integration of natural language processing for data structuring extends the body of work that emphasizes the role of NLP in misinformation detection, resonating with studies by Dunn et al. [26] and Agrawal et al. [27].

4.3.2. Limitations

The analysis is limited to football transfer news articles from the BBC gossip column. These sources may not be representative of all football transfer news sources. However, the BBC gossip column is a collection of news articles from across the press, so there should be a

diverse set of articles; however, there will not be every article about transfers in the relevant EPL seasons, so this might not present the full image of what factors affect misinformation in the EPL.

The study is restricted to transfer windows in the years between 2021 and 2023 in the EPL, and the results may not be applicable to other leagues or transfer windows. The sample size could be larger to create more robust conclusions. In addition, the lack of focus on other leagues may mean that the study might not apply to other leagues with different transfer dynamics and media landscapes. Although the focus is on three transfer windows, the researchers believe that they do provide high relevance to the interpretation of the results by potential stakeholders, as they will have less concern about how media news reporting will change.

Another limitation is related to the differences in club names across various datasets, such as "Man Utd" vs. "Manchester United." This issue is addressed by using the library "theFuzz" to calculate similarity scores between club names. The study sets a threshold of 75 for determining whether the club names are similar enough. Club names with similarity scores between 50 and 75 were manually checked and aligned based on the researcher's football knowledge. However, this approach may still result in some discrepancies in the data.

Although the GPT-3 LLM is a powerful tool for data preparation and feature extraction, it is not infallible. It may cause misinterpretation, mislabeled features, or even hallucinations [26], leading to inaccurate data processing and poor model performance. This risk is mitigated by checking the structured data from GPT-3 against the factual data from the transfermarkt.com and the API-Football datasets.

4.3.3. Future Research Directions

To address some limitations, future work could focus on improving model performance by refining feature selection using the most important features suggested by this study, exploring more advanced ML algorithms, or incorporating additional data sources, as suggested by the literature [17,18,21], e.g., from different football leagues or more historical data. Moreover, the findings from this study can be applied to other domains beyond football, such as politics, finance, or healthcare, where misinformation detection is critical. In addition, other types of features, such as linguistic features, have been used to predict misinformation [23], and in future studies, these features could be included in this same context to determine whether there are improvements in these results.

5. Conclusions

Informed by the literature review, which highlighted the potential advantages of ensemble learning models in detecting misinformation [20,21] and the importance of NLP techniques in structuring data [26,27], this study investigated three ensemble learning models (Random Forest, AdaBoost, and XGBoost) and identified key factors that influence the veracity of football transfer rumors.

Our findings resonate with the literature on ensemble learning and NLP methods. A player's age and time to the start/end of the transfer window showed weak negative relationships with rumor veracity, indicating that for older players and times further from the transfer window, rumors are slightly less likely to be true. These relationships might be explained by the fact that older players tend to have fewer potential transfers remaining in their careers and that rumors closer to the transfer window might be more accurate due to the immediacy of the event. The findings of this study align with those of previous studies that demonstrated the importance of diverse features in predicting misinformation [23–25]. In contrast, a player's market value did not have a statistically significant relationship with rumor veracity, suggesting that other factors may overshadow its effect on rumor veracity, as observed in studies focused on predicting sports transfers [28,30].

Despite the promising results, the study can be expanded using other sport media datasets or news in different countries and languages to improve the applicability of our ensemble learning approach.

By building upon the insights from the literature review and the findings of this study, researchers can develop new combined approaches for detecting and predicting misinformation in various contexts by employing advanced randomized learner models (e.g., stochastic configuration networks) with ensemble learning.

Author Contributions: All authors had an equal contribution in preparing and finalizing the manuscript. Conceptualization: I.R. and M.L.; methodology, I.R., M.L. and M.A.; validation: I.R., M.L., M.A. and J.H.; formal analysis: I.R., M.L., M.A. and J.H.; investigation: I.R., M.L., M.A. and J.H.; data curation: I.R.; writing—original draft preparation: I.R. and M.L.; writing—review and editing: I.R., M.L., M.A. and J.H.; supervision: M.L., M.A. and J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Joshi, A.M.L.; Data Analytics & Artificial Intelligence: What It Means for Your Business and Society. IMD business School for Management and Leadership Courses, 05-Dec-2022. Available online: https://www.imd.org/research-knowledge/articles/artificial-intelligence-real-world-impact-on-business-and-society/ (accessed on 4 January 2023).
2. Wu, L.; Morstatter, F.; Carley, K.M.; Liu, H. Misinformation in social media. *ACM SIGKDD Explor. Newsl.* **2019**, *21*, 80–90. [CrossRef]
3. Allen, J.; Howland, B.; Mobius, M.M.; Rothschild, D.M.; Watts, D. Evaluating the fake news problem at the scale of the information ecosystem. *Sci. Adv.* **2020**, *6*, eaay3539. [CrossRef] [PubMed]
4. Cavazos, R.; CHEQ. The Economic Cost of Bad Actors on the Internet. 2020. Available online: https://info.cheq.ai/hubfs/Research/Economic-Cost-BAD-ACTORS-ON-THE-INTERNET-Ad-Fraud-2020.pdf (accessed on 6 March 2023).
5. Postiglione, A.; Postiglione, G. Football: Between Esports, Crypto, NFT and Metaverse. Rome Business School. 12 December 2022. Available online: https://romebusinessschool.com/research-center/football-is-the-most-profitable-sport-with-global-revenue-of-47-billion/ (accessed on 7 March 2023).
6. Merten, B. The Impact of Transfer Spending in Expediting Improvement of On-Field Performance of English Premier League Clubs. Bachelor's Thesis, University of South Carolina, Columbia, SC, USA, 2022; p. 4.
7. Rojas Torrijos, J.L.; Mello, M.S. Football misinformation matrix: A comparative study of 2020 Winter transfer news in four European sports media outlets. *J. Media* **2021**, *2*, 625–640. [CrossRef]
8. Bridge, T. Records Tumble as Premier League Clubs Spend £815m. Deloitte United Kingdom. 1 February 2023. Available online: https://www2.deloitte.com/uk/en/pages/press-releases/articles/records-tumble-as-premier-league-clubs-spend.html (accessed on 7 March 2023).
9. Bright, S.; Subedar, A. 'Rooney to China'?: The Real Impact of Fake Football News. BBC News. 14 July 2017. Available online: https://www.bbc.com/news/blogs-trending-40574049 (accessed on 4 January 2023).
10. Economic Benefits of Premier League Confirmed by Report. Premier League Football News, Fixtures, Scores & Results. 21 April 2022. Available online: https://www.premierleague.com/news/2434933 (accessed on 7 March 2023).
11. Evans, S. Premier League celebrates 30 year rise to global dominance. Reuters. 16 August 2022. Available online: https://www.reuters.com/lifestyle/sports/premier-league-celebrates-30-year-rise-global-dominance-2022-08-16/ (accessed on 25 March 2023).
12. Brown, S. Machine Learning, explained. MIT Sloan. 21 April 2021. Available online: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained (accessed on 7 March 2023).
13. Thriving in the era of pervasive AI. The Wall Street Journal. 21 July 2020. Available online: https://deloitte.wsj.com/articles/thriving-in-the-era-of-pervasive-ai-01595358164 (accessed on 7 March 2023).
14. Accenture. Accenture Report: Artificial Intelligence Has Potential to Increase Corporate Profitability in 16 Industries by an Average of 38 Percent by 2035. Newsroom. 20 June 1970. Available online: https://newsroom.accenture.com/news/accenture-report-artificial-intelligence-has-potential-to-increase-corporate-profitability-in-16-industries-by-an-average-of-38-percent-by-2035.htm (accessed on 7 March 2023).
15. Ognyanova, K.; Lazer, D.; Robertson, R.E.; Wilson, C. Misinformation in action: Fake news exposure is linked to lower trust in Media, Higher Trust in government when your side is in power. *Harv. Kennedy Sch. Misinformation Rev.* **2020**, *1*, 1–19. [CrossRef]
16. Muhammed, T.S.; Mathew, S.K. The disaster of misinformation: A Review of Research in social media. *Int. J. Data Sci. Anal.* **2022**, *13*, 271–285. [CrossRef] [PubMed]

17. Alghamdi, J.; Lin, Y.; Luo, S. A comparative study of machine learning and Deep Learning techniques for fake news detection. *Information* **2022**, *13*, 576. [CrossRef]
18. Chen, M.-Y.; Lai, Y.-W.; Lian, J.-W. Using deep learning models to detect fake news about COVID-19. *ACM Trans. Internet Technol.* **2022**, *23*, 1–23. [CrossRef]
19. Liu, Y.; Wu, Y.-F. Early detection of fake news on social media through propagation path classification with recurrent and Convolutional Networks. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
20. Hansrajh, A.; Adeliyi, T.T.; Wing, J. Detection of online fake news using blending ensemble learning. *Sci. Program.* **2021**, *2021*, 3434458. [CrossRef]
21. Singh, G.; Selva, K. A comparative study of hybrid machine learning approaches for fake news detection that combine multi-stage ensemble learning and NLP-based framework. *TechRxiv* **2023**. [CrossRef]
22. Sahithi, G.L.; Roshmi, V.; Sameera, Y.V.; Pradeepini, G. Credit card fraud detection using ensemble methods in machine learning. In Proceedings of the 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 28–30 April 2022. [CrossRef]
23. Zhao, Y.; Da, J.; Yan, J. Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches. *Inf. Process. Manag.* **2021**, *58*, 102390. [CrossRef]
24. Buzea, M.C.; Trausan-Matu, S.; Rebedea, T. Automatic fake news detection for Romanian Online News. *Information* **2022**, *13*, 151. [CrossRef]
25. Vosoughi, S.; Roy, D.; Aral, S. The spread of true and false news online. *Science* **2018**, *359*, 1146–1151. [CrossRef] [PubMed]
26. Dunn, A.; Dagdelen, J.; Walker, N.; Lee, S.; Rosen, A.S.; Ceder, G.; Persson, K.; Jain, A. Structured information extraction from complex scientific text with fine-tuned large language models. *arXiv* **2022**. [CrossRef]
27. Agrawal, M.; Hegselmann, S.; Lang, H.; Kim, Y.; Sontag, D. Large Language Models are Few-Shot Clinical Information Extractors. *arXiv* **2022**. [CrossRef]
28. Kim, Y.; Bui, K.H.N.; Jung, J.J. Data-driven exploratory approach on player valuation in football transfer market. *Concurr. Comput. Pract. Exp.* **2019**, *33*, e5353. [CrossRef]
29. Dimov, P. Recognition of fake news in sports. *Strateg. Policy Sci. Educ.-Strateg. Na Obraz. I Nauchnata Polit.* **2021**, *29*, 18–27. [CrossRef]
30. Ćwiklinski, B.; Giełczyk, A.; Choraś, M. Who will score? A machine learning approach to supporting football team building and transfers. *Entropy* **2021**, *23*, 90. [CrossRef] [PubMed]
31. Silva, F.; SAS Voices. Going beyond the Box Score: Text Analysis in Sports. SAS Voices. 8 June 2020. Available online: https://blogs.sas.com/content/sascom/2020/06/08/going-beyond-the-box-score-text-analysis-in-sports/ (accessed on 7 January 2023).
32. Levenshtein Distance. Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 26 September 2023. Available online: https://en.wikipedia.org/wiki/Levenshtein_distance (accessed on 29 June 2023).
33. Aspers, P.; Corte, U. What is qualitative in qualitative research. *Qual. Sociol.* **2019**, *42*, 139–160. [CrossRef] [PubMed]
34. Gorard, S. *Quantitative Methods in Educational Research the Role of Numbers Made Easy*; Continuum: London, UK, 2007.
35. Refaeilzadeh, P.; Tang, L.; Liu, H. Cross-validation. In *Encyclopedia of Database Systems*; Springer: Boston, MA, USA, 2009; pp. 532–538.
36. Bouchrika, I. Primary Research vs Secondary Research: Definitions, Differences, and Examples. Research.com. 9 December 2022. Available online: https://research.com/research/primary-research-vs-secondary-research (accessed on 25 March 2023).
37. Premier League-Transfers 21/22. Transfermarkt. Available online: https://www.transfermarkt.com/premier-league/transfers/wettbewerb/GB1/saison_id/2021 (accessed on 25 March 2023).
38. Saturday's Transfer Gossip: Nagelsmann, Mendy, Kovacic, Pochettino, Paqueta, Sangare. BBC Sport. Available online: https://www.bbc.com/sport/football/gossip (accessed on 25 March 2023).
39. Transfermarkt. Wikipedia. 24 March 2023. Available online: https://en.wikipedia.org/wiki/Transfermarkt (accessed on 25 March 2023).
40. Banerjee, R. Transfer Window Terminology Explained: What Do Football's Deadline Day Phrases Mean? Goal.com. 21 February 2023. Available online: https://www.goal.com/en/news/transfer-window-terminology-explained-football-deadline-day-phrases-mean/blta171749901f75e05 (accessed on 25 March 2023).
41. Classification: True vs. False and Positive vs. Negative | Machine Learning | Google Developers. Google. Available online: https://developers.google.com/machine-learning/crash-course/classification/true-false-positive-negative (accessed on 14 April 2023).
42. Rao, C.R.; Wegman, E.J.; Solka, J.L. Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications. In *Handbook of Statistics*; Elsevier North Holland: Amsterdam, The Netherlands, 2005; pp. 403–428.
43. Jamshidian, M.; Mata, M. Advances in Analysis of Mean and Covariance Structure when Data are Incomplete. In *Handbook of Latent Variable and Related Models*; Elsevier: Amsterdam, The Netherlands, 2008; pp. 21–44.
44. Madley-Dowd, P.; Hughes, R.; Tilling, K.; Heron, J. The proportion of missing data should not be used to guide decisions on multiple imputation. *J. Clin. Epidemiol.* **2019**, *110*, 63–73. [CrossRef]
45. Brown, T.B.; Amodei, D.; Sutskever, I.; Radford, A.; McCandlish, S.; Berner, C.; Clark, J.; Chess, B.; Gray, S.; Litwin, M.; et al. Language Models are Few-Shot Learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.

46. Ali, A.H.; Yaseen, M.G.; Aljanabi, M.; Abed, S.A. Transfer learning: A new promising techniques. *Mesopotamian J. Big Data* **2023**, *2023*, 29–30. [CrossRef]
47. Biau, G.; Scornet, E. A Random Forest Guided Tour. *Test* **2016**, *25*, 197–227. [CrossRef]
48. Hornyák, O.; Iantovics, L.B. AdaBoost algorithm could lead to weak results for data with certain characteristics. *Mathematics* **2023**, *11*, 1801. [CrossRef]
49. Bentéjac, C.; Csörgő, A.; Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* **2020**, *54*, 1937–1967. [CrossRef]
50. Nair, A. Harnessing Randomness in Machine Learning. Medium. 21 February 2022. Available online: https://towardsdatascience.com/harnessing-randomness-in-machine-learning-59e26e82fdfc (accessed on 14 April 2023).
51. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. Neural Information Processing Systems. 2017. Available online: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html (accessed on 15 July 2023).
52. Scavuzzo, C.M.; Scavuzzo, J.M.; Campero, M.N.; Anegagrie, M.; Aramendia, A.A.; Benito, A.; Periago, V. Feature importance: Opening a soil-transmitted helminth machine learning model via SHAP. *Infect. Dis. Model.* **2022**, *7*, 262–276. [CrossRef]