FROM RUMORS TO FACTS: EVALUATING FOOTBALL TRANSFER NEWS

Chen Ziyi
Department of Electronic Engineering
2022010503

Wen Shenghan
Weixian College
2022012636

Yang Hongyi
Department of English Literature and Linguistics
2023011638

January 5, 2025

ABSTRACT

In this study, we analyzed 13,251 football transfer rumors collected from `www.transfermarkt.co.uk`. We clustered the news sources using metrics such as coverage, accuracy, and a new "EarlyBird" metric that measures how much earlier a transfer is reported before it happens. This allowed us to identify common strategies used by these sources. Additionally, we extracted key features from the rumors and developed a predictive model to estimate the likelihood of a reported transfer occurring. Our model achieved an accuracy of 73.77%, demonstrating the effectiveness of our approach and providing valuable insights into how football transfer news is reported.

1 Introduction

Football has become the universal language, transcending ethnicities and age groups worldwide. Its growing global popularity has significantly increased the commercial value of the sport's key figures—the players. Over the past decade, the football transfer market, where clubs pay fees to acquire players from other teams, has surged to nearly $10 billion. This monumental market attracts immense global attention, making news about potential player moves, commonly referred to as transfer rumors, a focal point of discussion. These rumors influence fans, analysts, and even club strategies, highlighting their substantial impact on the football ecosystem.

However, transfer rumors exhibit considerable variability in accuracy and reliability. Some sources are renowned for providing near-confirmatory updates with phrases like "here we go," while others often announce unexpected breaking news that ultimately proves unfounded. This inconsistency underscores the necessity for a systematic analysis of the patterns and predictors that determine the reliability of these rumors. Understanding these dynamics can help stakeholders navigate the complex landscape of football transfers with greater insight and confidence.

To address this need, we utilized the comprehensive database Transfermarkt, a highly reputable source known for its extensive archive of transfer records and associated rumors contributed by football fans worldwide. We scraped 13,251 transfer rumors related to footballers who transferred between 2005 and 2023, using their actual transfer records as ground truth for validation. Recognizing that different news sources may employ distinct strategies tailored to their specific audiences and objectives, we clustered these sources based on metrics such as coverage, accuracy, and a novel "EarlyBird" metric. The EarlyBird metric quantifies the lead time before a transfer materializes, allowing us to identify common strategies and characteristics among the sources.

Furthermore, we extracted key features from the rumors and developed a predictive model to assess the likelihood of a reported transfer occurring. Our model achieved an accuracy of 73.77%, demonstrating the effectiveness of our approach and providing valuable insights into the dynamics of football transfer reporting. This study not only sheds light on the reliability of various news sources but also offers a methodological framework for analyzing and predicting the outcomes of transfer rumors.

The remainder of this paper is structured as follows: section 2 reviews related literature on football transfer analysis and rumor credibility; section 3 details our data collection, pre-processing, clustering techniques, and predictive modeling

approach; section 4 presents our experimental setup and results, while section 5 concludes with a discussion of our findings, limitations, and suggestions for future research.

2 Related Work

Identifying misinformation in news reports has been a important topic of data science. In the field of football transfer news, [Runsewe et al., 2024] focused on from the BBC gossip column, extracting 304 days of gossip columns that resulted in 5982 lines of transfer news in the 2021/22 season. We followed the idea of prediction based analysis using machine learning models including Random Forest, but extended it to rumors of various sources, and highlighted the feature engineering process, in which we discussed how different factors of rumors may impact their credibility.

To extract accurately the attitude of the rumor post to a specific potential transfer, inspired by the work [Zhou and Yu, 2023] on biomedical relations, we called the OpenAI API by prompts integrated with the content of rumor posts to acquire a semi-quantitative score of confidence.

3 Method

3.1 Data Acquisition

The pipelines for our data acquisition are illustrated in Figure 1. We start with the transfer records webpages on Transfermarkt, collecting 500 records for each year from 2005 to 2023. Using the Rcrawler package to loop through URLs across years and pages, we extract the main pages of all players associated with these records. From the player pages, Rcrawler is then used to fetch the URLs of rumor pages, which contain all relevant rumor posts about the player. By identifying the corresponding nodes with Selector Gadget and using the rvest package, we scrape the rumor source, the posting date, and the user-summarized content of each rumor. For ground truth data, we dynamically request the player's transfer records page to obtain JSON-formatted transfer histories. Using the jsonlite package, we convert this data into a data frame for further alignment and analysis.

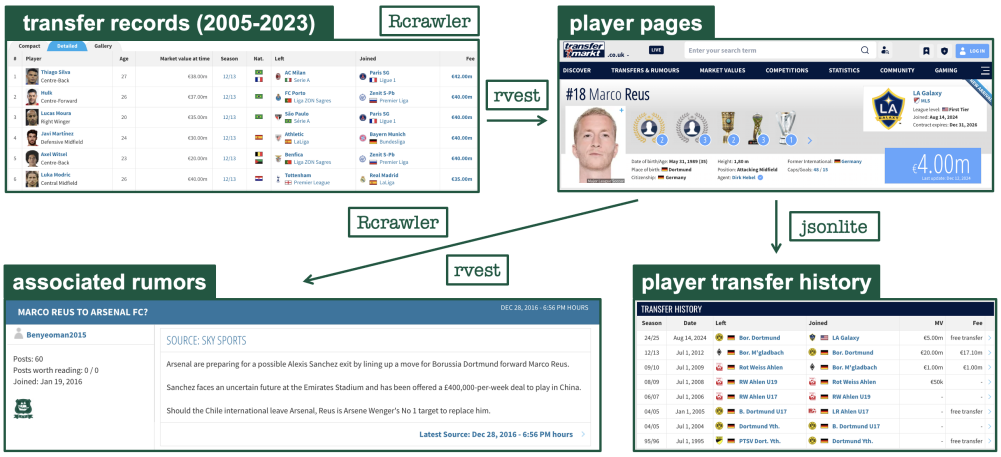

Figure 1: Pipelines for Data Acquisition

3.2 Text Processing

To evaluate the confidence expressed in rumor posts about potential transfers, we aimed to create a quantified metric (Figure 2). Initially, we used the `get_sentiment()` function from the Syuzhet package in R, a dictionary-based method that detects words with specific attitudes and assigns a positivity score, which is then summed for the entire paragraph. However, this method is unable to accurately assess the confidence regarding a specific player's transfer to a particular club and is overly influenced by irrelevant or redundant parts of the posts. To address these limitations, we integrated the ChatGPT 3.5 turbo API. By structuring the post content into a designed prompt, we utilized the language model to generate a graded numerical evaluation of confidence and return in JSON format.

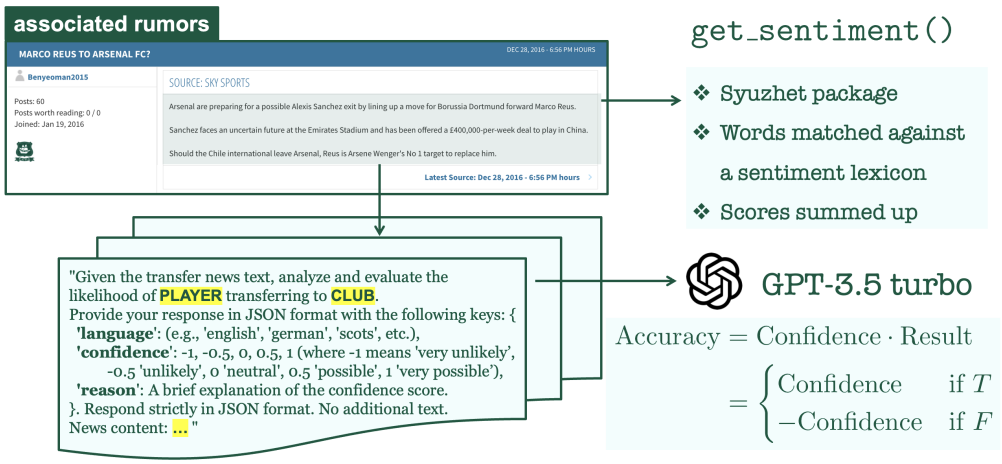

Figure 2: Framework for Analyzing Confidence in Transfer Rumor Contents

3.3 Data Alignment

3.3.1 Source Name Alignment

The source names we get from rvest are not standardized and mainly consists of two types: URLs (e.g. `www.theguardian.com`, `www1.bbc.com`) and specific names (e.g. Guardian, L´Equipe).

For specific names, we used Google to search for keywords that showed up in the names. By using the First Organic Search Result as our target answer, we are able to get reliable and consistent standardized URL results that had undergone manual double-check.

For URLs, the main problem is that the same rumor source can have different URLs and domain names (e.g. m.bbc.com, www.bbc.co.uk). Our first step is to match the URL start (https://) and delete subdomain name (e.g. www./en./m.) that appear before the actual information. Then we match and delete top domain name that appear at the end of the URL. All URLs are processed using regex expressions in R.

The final result we get for rumor source are formatted as https://xxx and is easy for further data processing.

3.3.2 Rumor-Fact Alignment

The alignment of rumors and facts is determined based on two main criteria: the player's name and the timing of the event, as illustrated in Figure 3. A rumor and fact are considered aligned if and only if the player's name matches and the event occurs within a specified time frame (e.g., within a few months following the publication of the rumor). If the player's name matches but the target club differs, the rumor is deemed incorrect. Similarly, if no match is found between the rumor and the fact, the rumor is also classified as incorrect, indicating that the player did not actually transfer.

First, the player's name serves as a unique identifier, ensuring consistency between the rumor and the corresponding fact for that entry. A direct match is obtained by standardizing the name, which involves converting it to lowercase and removing special characters (such as hyphens). If the resulting names are identical, they are considered to match.

In contrast, club names exhibit greater variability and inconsistency, necessitating a more intricate rectification process. This process includes removing common suffixes that denote clubs or associations, addressing regional name abbreviations, and restoring Latin transliterations. After applying this rectification, a match between clubs indicates that the club associated with the rumor is accurate.

Regarding timing, the rumor and corresponding fact must satisfy the condition that the fact occurs after the rumor. Furthermore, if a rumor is associated with multiple facts, it may suggest that the player underwent multiple transfers. In such cases, the most recent fact, in terms of its proximity to the rumor, is selected as the corresponding fact.

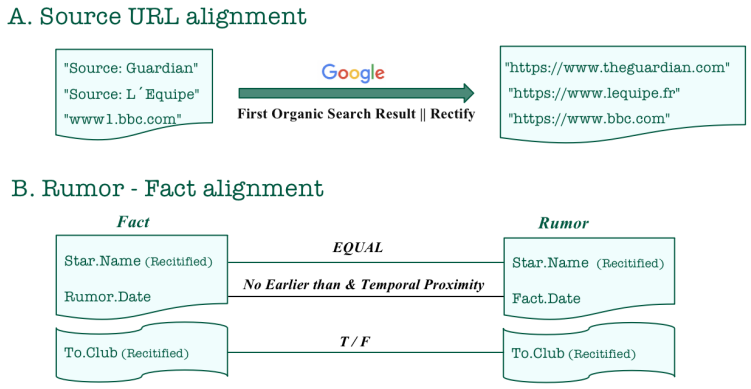

Figure 3: Methods for alignment

3.4 Feature Engineering

In this section, we will analyze data features such as time, reliability indicators, content, and source, and select those that are more useful for prediction and offer greater interpretability.

3.4.1 Time

Time encompasses various units, including month, week, hour, and minute. As illustrated in Figure 4, a significant nonlinear relationship between the month and accuracy was observed.

Additionally, by constructing a random forest model to evaluate the contribution of each feature to the accuracy (i.e., the share of each feature in the model's predictive power), as shown in Figure 5, it is evident that the month significantly outperforms other features in terms of its impact on prediction accuracy. This further supports the hypothesis that the month is a key factor in determining the accuracy of transfer rumors.

This is likely due to the cyclical nature of transfer seasons, wherein rumors occurring farther from the actual transfer period tend to be less accurate.

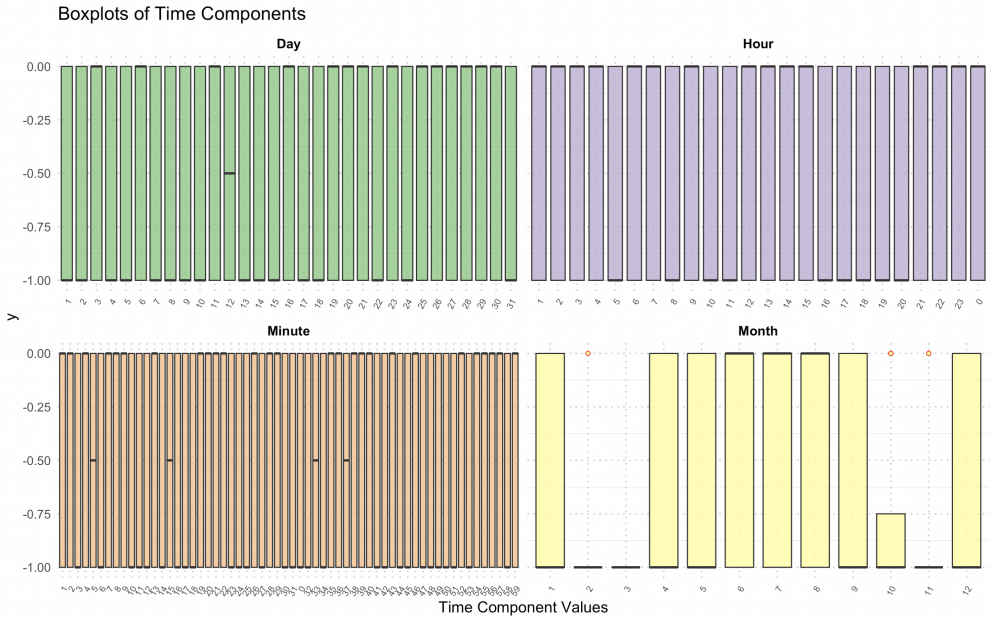

Figure 4: Time Boxplot

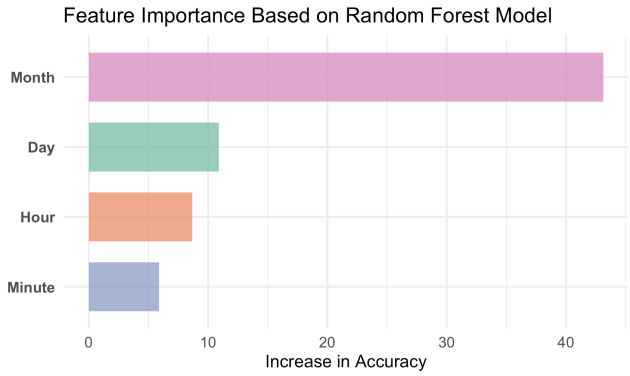

Figure 5: Increased accuracy for Time

3.4.2 Reliability

Variable Correlation:

- **GPT.confidence**: 0.24611174
- **Rumor.Sentiment**: -0.03115098

GPT.confidence exhibits a significantly stronger correlation with the truthfulness of rumors compared to Rumor.Sentiment.

Further analysis, as depicted in Figure 6, shows that Rumor.Sentiment lacks any discernible linear or nonlinear relationship with the truthfulness of the rumor. In contrast, GPT.confidence displays a clear positive correlation.

This disparity may arise because Rumor.Sentiment primarily reflects local semantics, focusing on the sentiment within the immediate context of the rumor. On the other hand, GPT.confidence incorporates global semantics, capturing broader contextual information. Additionally, GPT's pre-training likely provides it with a richer base of prior knowledge, which could enhance its ability to assess the truthfulness of rumors.

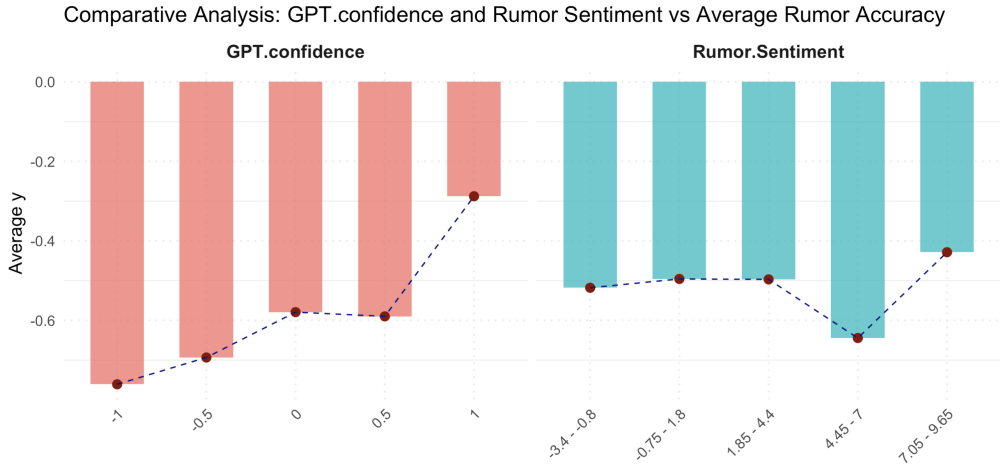

Figure 6: Correlation between reliability and truthfulness

3.4.3 Rumor Content and Source

When using only the rumor content as a feature, we construct a random forest model to evaluate its contribution to increased accuracy (i.e., the proportion of each feature's influence on model accuracy), as shown in Figure 7.

Interpretative analysis:

- **united / manchester / chelsea / arsenal / liverpool / city:** These terms refer to well-known football clubs (e.g., Manchester United, Manchester City, Chelsea, Arsenal, etc.). The frequency of their occurrence in transfer rumors may be correlated with the prominence of these teams. Rumors involving popular clubs often attract more attention, leading to higher likelihoods of verification and appearing in more credible news sources.
- **deal / transfer / loan / fee:** These words describe essential aspects of transfer transactions (such as deal, transfer, loan, transfer fee). Rumors that include specific details about transfers tend to be perceived as more credible, as the presence of concrete transfer-related information suggests a closer connection to factual events.
- **summer / season:** These terms are associated with the timing of the transfer. The cyclical nature of football transfers, particularly during the summer transfer window, makes rumors mentioning specific periods (such as "summer") more plausible and credible.
- **striker / club:** These words indicate the player's role or the club involved. Specific references to player roles (e.g., striker) or particular clubs can lend professionalism and specificity to the rumor. Rumors that mention detailed roles or positions are often more credible than those using vague or generalized terms.
- **sky / according:** These are source-related terms. For example, "Sky" may refer to a media outlet (e.g., Sky Sports), while "according" indicates attribution to a source. The credibility of the source is crucial for evaluating the truthfulness of a rumor, as reputable sources are more likely to disseminate accurate information.
- **sign / move / will:** These action verbs describe potential player actions (such as signing, moving, or intending to move). Clear and decisive verbs enhance the specificity of the rumor, suggesting it may be based on tangible developments, thereby increasing its credibility.
- **yearold:** This typically refers to the player's age (e.g., "25-year-old"). Rumors that include an age reference are often more credible, as age is a commonly discussed detail in transfer-related contexts and is less likely to be speculative.

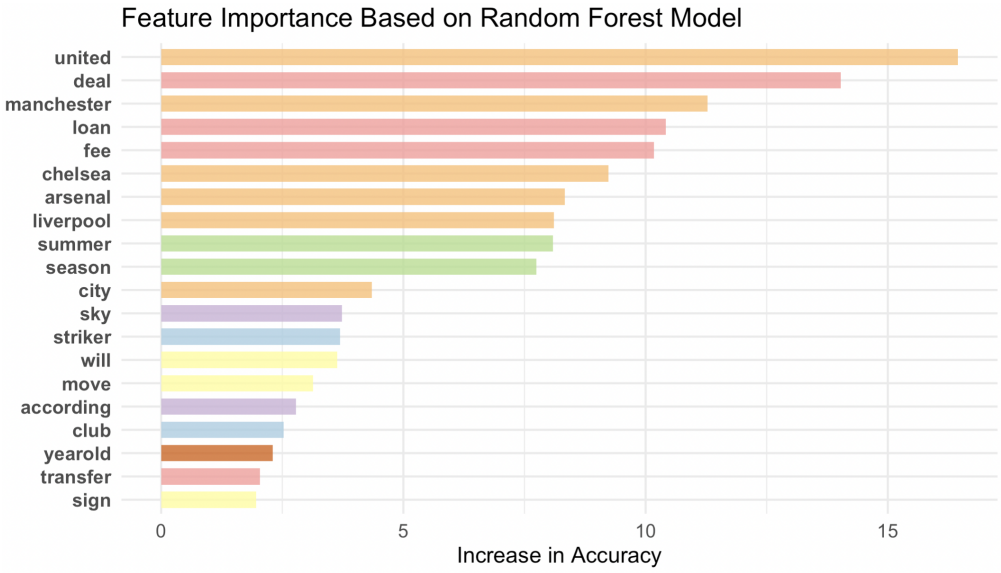

Figure 7: Increased accuracy for content

When using only the rumor source as a feature, we construct a random forest model to evaluate its contribution to increased accuracy (i.e., the proportion of each feature's influence on model accuracy), as shown in Figure 8.

The reliability of media sources tends to be polarized, with some sources being highly credible and others being notably unreliable. This distinct variation in source reliability makes it a more effective predictor for rumor accuracy.

3.4.4 Others

Through experimentation, we observed that both the nationality of the rumor publisher and the player's position also contribute to the accuracy of subsequent predictions. As a result, we have incorporated these two features into the model.

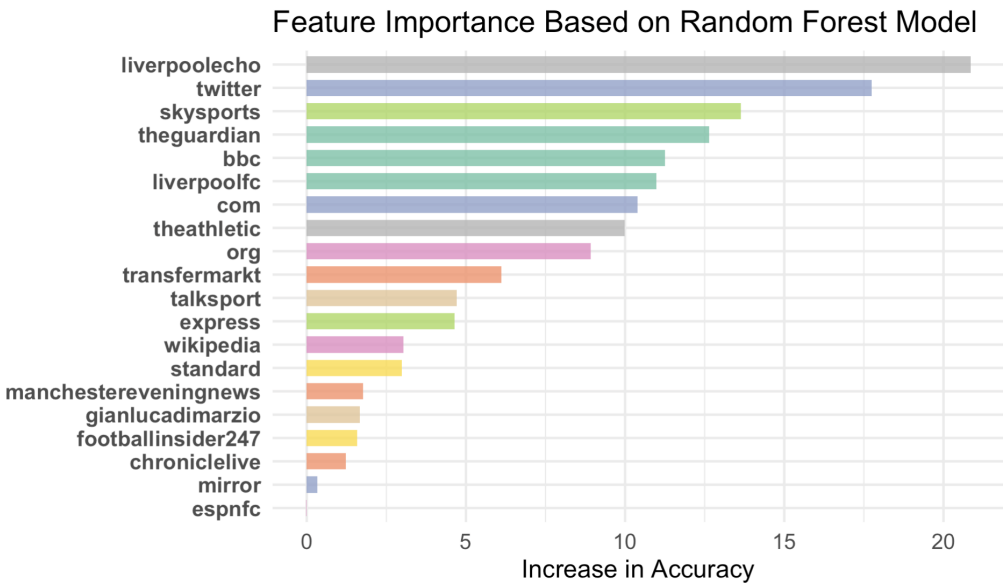

Figure 8: Increased accuracy for source

3.5 Prediction

Figure 9 illustrates the prediction pipeline. After aligning the data, the features are categorized into three types based on their data formats:

- **String text features:** These include the rumor content and the rumor source. These features undergo preprocessing, such as stopword removal, and are then transformed into feature matrices. Sparse entries are eliminated to reduce model complexity and accelerate the training process.
- **Numeric features:** These consist of attributes such as the rumor's publication month and GPT-confidence. These features are directly used in their numeric form for model training.
- **Categorical features:** These include categorical variables such as the nationality of the rumor publisher and the player's position. Depending on the model, these features may be treated as categorical variables (e.g., in random forest models) or transformed into matrices (e.g., in SVM or ElasticNet models).

Considering the data imbalance, we performed downsampling to balance the number of positive and negative samples. A model without training would achieve an accuracy of around 50%.

Finally, we randomly partitioned the data into training and testing sets, with a 7:3 ratio.

3.5.1 SVM

SVM is a supervised machine learning algorithm used for classification and regression tasks. It aims to find the optimal hyperplane that maximizes the margin between different classes in the feature space. The problem can be formulated as the following optimization problem:

$$\max_{\boldsymbol{\alpha}} \left[\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \right],$$

$$\alpha_i \geq 0, \quad \sum_{i=1}^{N} \alpha_i y_i = 0, \quad \alpha_i \leq C, \forall i.$$

Key Parameters:

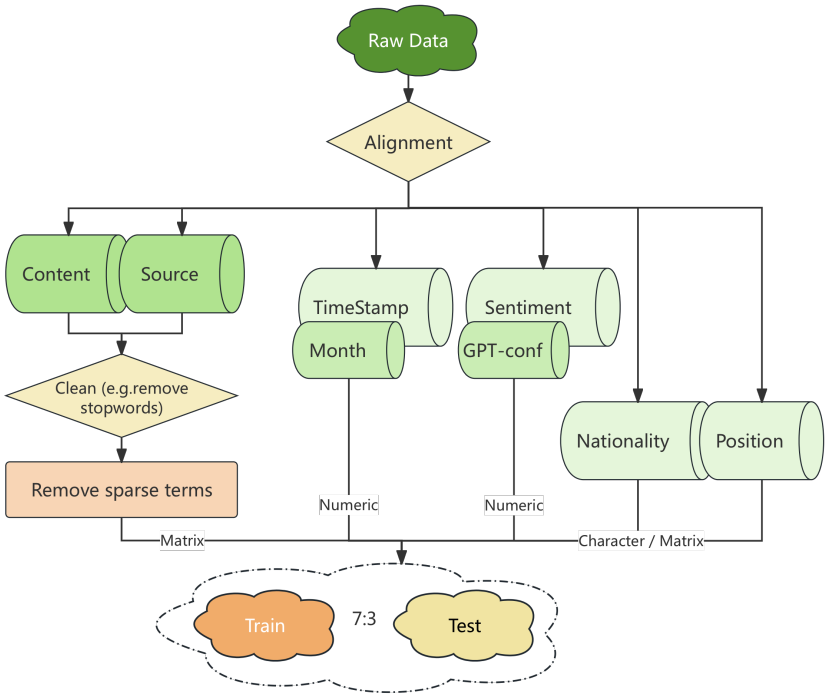

Figure 9: Pipeline for prediction

- **Kernel Function: radical** - The kernel function $K(x_i, x_j)$ is used to capture non-linear relationships between the data points. In this case, the Radial Basis Function (RBF) kernel is employed, which maps the data into a higher-dimensional space where it becomes easier to find a linear hyperplane.
- **Cost C: 3** - The cost parameter C controls the trade-off between achieving a low margin error and ensuring that the decision boundary is as smooth as possible. A value of $C = 3$ was selected based on parameter tuning, which aims to balance overfitting and underfitting.

3.5.2 ElasticNet

Logistic regression is a statistical method used for binary classification, which models the probability that a given input belongs to a particular class. The objective function, including regularization, is given by:

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} \left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right] + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + \frac{1}{2}(1 - \alpha) \sum_{j=1}^{p} \beta_j^2 \right).$$

Here, p_i represents the predicted probability for each data point, y_i is the actual label (either 0 or 1), and β_i are the model parameters. The regularization term ensures that the model doesn't overfit the training data.

Key parameters:

- **Family: binomial** — indicating a binary classification problem.
- **Lambda: auto** — automatically selected during model training, often via cross-validation.
- **Alpha: 0.5** — a parameter that controls the mix between L1 regularization (Lasso) and L2 regularization (Ridge), set to 0.5 based on tuning for optimal performance.

The model aims to balance between minimizing the log-likelihood of the logistic regression and the regularization term, which helps prevent overfitting by penalizing large model coefficients.

3.5.3 Random Forest

Random Forest is an ensemble learning method used for classification and regression tasks. It builds multiple decision trees and merges them together to produce a more accurate and stable prediction. Each tree in the forest is trained on a random subset of the data, with each node split based on a random subset of features, which helps reduce overfitting and increases generalization.

Key parameters:

- **NTree: 300** — The number of trees in the forest. In this case, 300 trees are used to improve the model's performance and stability. Increasing the number of trees typically reduces variance and enhances accuracy, but comes with a computational cost.

4 Result

4.1 Visualization

We developed a RShiny app for interactive visualization and exploratory data analysis in source-specific features. The app uses kmeans method to cluster sources using selected features. Also, the app utilizes Principal Component Analysis (PCA) analysis as a method to reduce dimension while retaining most significant information. Together, PCA1 and PCA2 provided another coordinate to analyze the effect of source grouping in a more visual-friendly way.

For RShiny part, the app uses multiple `observeEvent()` functions to update data grouping and the graph output in real-time. Details can be found in the source code provided.

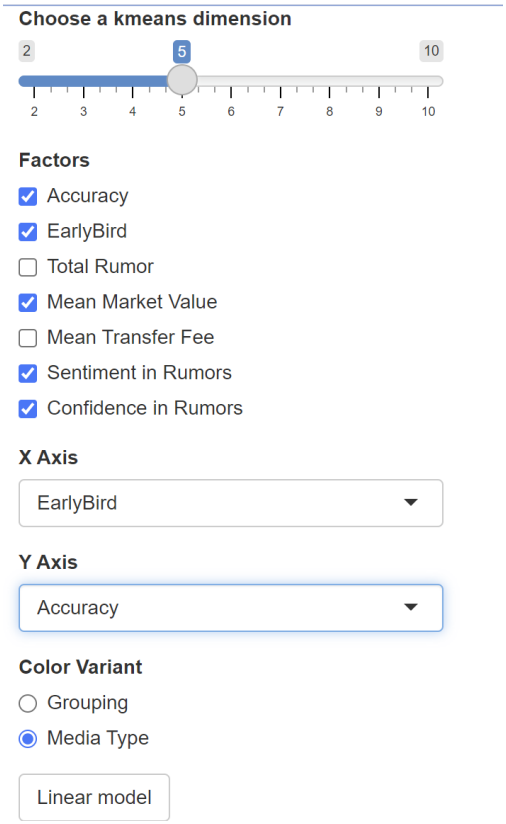

Figure 10: User interface of RShiny app

As is shown in Figure 10 from top to bottom, the user interface includes:

- Dimension used for kmeans grouping
- Features used in kmeans grouping

- X axis: parameter used to draw the x axis (EarlyBird / Confidence / PCA1)
- Y axis: parameter used to draw the y axis (Accuracy / Sentiment / PCA2)
- Color variant: parameter used to color different rumor sources (kmeans Grouping / Media type)
- Linear model: adds a linear regression line to the chart

The output includes:

- k-means Cluster Assignment Matrix: shows kmeans grouping result in each media type
- Scatter plot: Each point(or bubble) represents a rumor source we fetched. The size illustrates the quantity of rumors posted by the source. Colored by grouping, with legend.

The RShiny app can be utilized to discover more about the source-specific characteristics. Besides, to reveal more about properties of the sources we introduced a manually labelled `media type` label for each rumor source, which categorizes sources into five categories. They are:

- `Club Specific Website`: Website dedicated to a specific club. Example: Liverpool, Chelsea, Arsenal
- `Mainstream Media`: Mainstream media that cover all sorts of news including football transfers. Example: BBC, The Guardian, The Telegraph
- `Social Media`: Social media platforms with user generated content. Example: Facebook, X, Twitter
- `Sport Website`: Website dedicated to sports news only. Example: Skysports, 90 MIN
- `Other`: Source that can not be placed into categories above. Example: Youtube

The result we get as shown in Figure 11, which suggests features of rumor source can be an indication of its actual media type. (e.g. club specific website have a long EarlyBird, as well as high confidence and accuracy, which may be explained by their inside news and long-term focus on surrounding a specific club; social media have a preference to cover transfer rumors of high-market-value football stars and have a high accuracy, which might be biased by confirmation posts; see more in detail in Table 1)

Table 1: Features in each media type, mean value

Rumor_Label	accuracy	coverage	earlyBird	total_rumor	marketvalue	fee	confidence
Club Specific Website	0.1909089	*0.01221372*	**121.33506**	*29.00000*	*13558507*	*20085741*	**0.8503613**
Social Media	**0.27028308**	0.1127948	*45.32854*	257.25000	**25059916**	**31561311**	0.6008637
Mainstream Media	0.13433864	**0.18828220**	56.81000	**271.03371**	16448348	22086556	0.5833447
Sport Website	*0.09570226*	0.02412944	61.45449	67.09091	17020123	23037226	*0.5522350*
Other	0.15206255	0.03894491	60.56025	88.83333	21040859	26598533	0.5762029

4.2 Prediction

4.2.1 SVM

Accuracy and 95% CI:

- **Accuracy:** 71.80%
- **95% CI:** (0.6919, 0.7431)

The following confusion matrix shows the performance of the Support Vector Machine (SVM) classifier:

Prediction / GT	FALSE	TRUE
FALSE	429	163
TRUE	181	447

Table 2: Confusion Matrix for SVM

```
                      1  2  3  4  5
Club Specific Website 0  0  0  5  2
Mainstream Media      2  0  7 11  8
Other                 1  2  0  2  1
Social Media          1  1  2  0  0
Sport Website         7  1  6 10  9
```

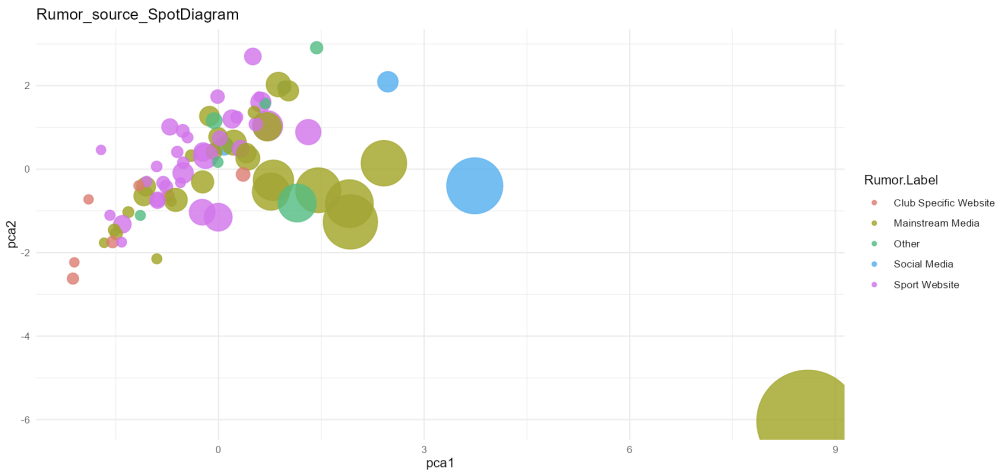

Figure 11: k-means Cluster Assignment Matrix and Rumor Source Scatter plot

4.2.2 ElasticNet

Accuracy and 95% CI:

- **Accuracy:** 73.61%
- **95% CI:** (0.7104, 0.7606)

ElasticNet performs better than the SVM model. The confusion matrix is shown as below:

Prediction / GT	FALSE	TRUE
FALSE	438	150
TRUE	172	460

Table 3: Confusion Matrix for ElasticNet

4.2.3 Random Forest

Accuracy and 95% CI:

- **Accuracy:** 73.77%
- **95% CI:** (0.7121, 0.7622)

The Random Forest classifier achieved the highest accuracy of 73.77%, a bit higher than ElasticNet. The confusion matrix for this model is as follows:

Prediction / GT	FALSE	TRUE
FALSE	478	188
TRUE	132	422

Table 4: Confusion Matrix for Random Forest

The feature importance in Random Forest was measured using the Gini index. The top 30 features with the highest Gini index are shown in the figure12.

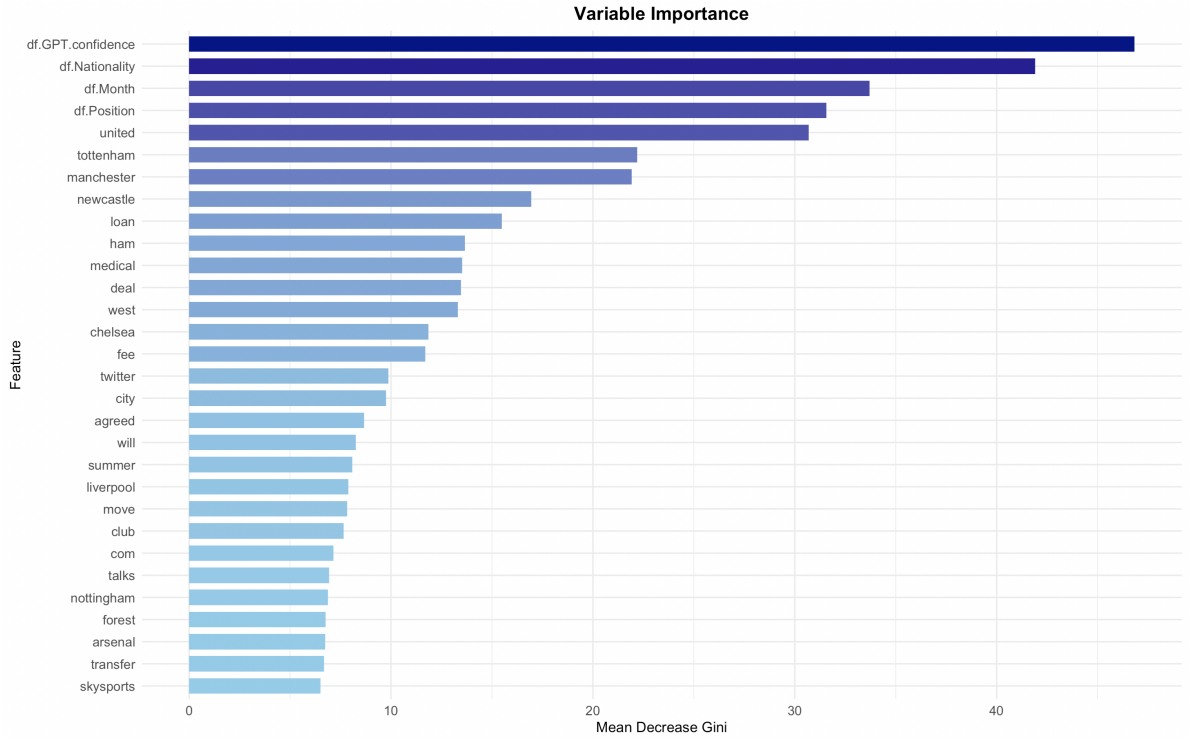

Figure 12: Top 30 Features Based on Gini Index in Random Forest

5 Discussion

5.1 Results

This study aimed to evaluate football transfer rumors by systematically analyzing their features and building predictive models. By leveraging 13,251 rumored transfers from Transfermarkt, we explored the relationship between rumor characteristics and their eventual outcomes. We employed clustering techniques to group news sources based on metrics such as accuracy, coverage, and reporting tendencies, and developed predictive models to assess the likelihood of a transfer materializing, achieving a peak accuracy of 73.77% using the Random Forest algorithm.

Our analysis highlighted several critical features that contribute to predicting transfer outcomes. Notably, the timing of rumors, measured by the publication month relative to the transfer window, emerged as a key factor, reflecting the cyclical nature of transfer markets. Additionally, the credibility of rumor sources, derived from historical accuracy rates, played a significant role in prediction. Features extracted from rumor content, such as specificity in naming clubs and player details, further enhanced the model's performance by emphasizing the importance of concrete, context-rich reporting.

The inclusion of GPT-generated confidence scores provided an additional layer of interpretability, capturing the level of certainty expressed in rumors. These scores demonstrated a meaningful correlation with rumor accuracy, validating their usefulness as a feature. Through feature importance analysis, we found that combining temporal, content-based, and source-based attributes allowed our model to uncover nuanced patterns underlying rumor reliability.

5.2 Limitations

Despite strengths mentioned above, there are several limitations to our study. The dataset, while extensive, is restricted to Transfermarkt and may not represent all global transfer rumors. Furthermore, the significant imbalance between true and false rumors in the dataset posed challenges for the machine learning models, necessitating sampling techniques to mitigate its impact. Expanding data sources and applying advanced methods to address class imbalance could further enhance predictive accuracy.

Future research could explore real-time prediction of emerging rumors, integrate social media data for a more holistic view, and apply advanced natural language processing techniques to capture richer semantic features. By quantifying the

reliability of rumor sources and extracting predictive features, we contribute to a deeper understanding of the dynamics of transfer news. Consequently, the methods introduced in this work could be extended to study rumor dynamics in other domains, such as for use in analyzing news in financial markets or political news.

5.3 Conclusion

In conclusion, our study effectively evaluated football transfer rumors by analyzing their features and developing predictive models. Utilizing the dataset collected from Transfermarkt, we identified key factors such as timing, source credibility, and content specificity that significantly influence rumor outcomes. The Random Forest algorithm achieved a peak accuracy of 73.77%, with GPT-generated confidence scores enhancing interpretability. Despite limitations like dataset restrictions and class imbalance, our findings offer valuable insights into rumor dynamics. In addition, we developed a small application using RShiny to facilitate interactive data exploration and visualization. Finally, we discussed both what we have achieved and the limitations of our work, providing insights for future research endeavors.

6 Contributions

Wen Shenghan: Introduction, Related Works, Data Acquisition, Text Processing

Yang Hongyi: Data Acquisition, Text Processing, Data cleaning and pre-processing, Rshiny

Chen Ziyi: Data cleaning and pre-processing, Data Alignment, Feature Engineering, Prediction

References

[Runsewe et al., 2024] Runsewe, I., Latifi, M., Ahsan, M., and Haider, J. (2024). Machine learning for predicting key factors to identify misinformation in football transfer news. *Computers*, 13(6):127.

[Zhou and Yu, 2023] Zhou, S. and Yu, S. (2023). High-throughput biomedical relation extraction for semi-structured web articles empowered by large language models. *arXiv preprint arXiv:2312.08274*.