FTR-18: Collecting rumours on football transfer news

Danielle Caled, Mário J. Silva

INESC-ID Lisboa, Instituto Superior Técnico, Universidade de Lisboa
Lisbon, Portugal
{dcaled, mjs}@inesc-id.pt

Abstract

This paper describes ongoing work on the creation of a multilingual rumour dataset on football transfer news, FTR-18. Transfer rumours are continuously published by sports media. They can both harm the image of player or a club or increase the player's market value. The proposed dataset includes transfer articles written in English, Spanish and Portuguese. It also comprises Twitter reactions related to the transfer rumours. FTR-18 is suited for rumour classification tasks and allows the research on the linguistic patterns used in sports journalism.

1 Introduction

In the last years, the way news are spread has changed. Social networks, blogs/micro-blogs and other untrusted online news sources have gained popularity, thus allowing any user to produce unverified content. This modern kind of media enables real-time proliferation of news stories, and, as consequence, increases the diffusion of rumours, hoaxes and misinformation to a global audience [7]. As news content is continuously published online, the speed in which it is disseminated hinders human fact-checking activity. A piece of information whose "veracity status is yet to be verified at the time of posting" is called rumour [14]. News articles, scientific researches and conspiracy-theory stories can float in the veracity spectrum. They are characterised by distinct stylistic dimensions and these features enable their spread, but are orthogonal to their truthfulness [9].

The football transfer market offers a fertile ground for rumour dissemination because rumours and misinformation spreads may be motivated by personal profit or public harm. Releasing false or misleading news about alleged moves of players between clubs emerges as a strategy to increase a player's market value and transfer fees. Negligent transfer announcements are assigned to high sums directly related to some transfers [4]. In addition, news organisations also benefit from announcements of alleged transfer moves, attracting public attention by selling contents and advertisements. Maia showed that most of the football transfer news published by the three biggest Portuguese sports diaries during the Summer of 2015 were not confirmed [4]. Table 1 provides an example of a news article about a transfer to Real Madrid that was latter denied by the club in an official announcement. Both the transfer news and the announcement were released on July 4, 2018.

According to the Union of European Football Associations' (UEFA) annual report[1], the European transfer market reached a record revenue of €5.6 billion during the Summer of 2017, a 6% increase in spending during this period relatively to the highest value recorded in the last ten years. Economical side effects of rumours on football clubs shares are also reported, like the rumour of Cristiano Ronaldo joining Juventus FC, which made the club shares jump almost 10% on a single day[2] before any official announcement.

Motivated by the extensive attention given by sports media to transfer news and the high amounts involved, we propose the creation of Football Transfer Rumours 2018 (FTR-18), a transfer rumours dataset for researching the linguistic patterns in their text and propagation mechanisms. FTR-18 is designed as a multilingual collection of articles published by the relevant news organisations either in English, Spanish or Portuguese languages. Besides news articles, our proposed dataset also comprises Twitter posts associated to the transfer rumours. The present work describes the creation process of FTR-18 dataset and discusses

[1] https://goo.gl/FoAv2g
[2] https://goo.gl/3wKhj3

Table 1: Example of a transfer rumour.

Topic: Real Madrid agreed a deal for PSG's Kylian Mbappé.
Agreeing news source: www.msn.com[a] **Headline**: Real Madrid agree stunning €272m deal to sign Kylian Mbappe **Extract**: *Real Madrid have reportedly agreed an eye-catching €272 million deal with Paris Saint-Germain for Kylian Mbappe.* *According to reports coming out of France from Baptiste Ripart, the French champions are set to lose Mbappe, with Madrid agreeing to pay the fee over four instalments.*
Evidence: Real Madrid official announcement[b] **Extract**: *Given the information published in the last few hours regarding an alleged agreement between Real Madrid C.F. and PSG for the player Kylian Mbappé, Real Madrid would like to state that it is completely false.* *Real Madrid has not made any offer to PSG or the player and condemns the spreading of this type of information that has not been proven by the parties concerned.*

[a] https://goo.gl/HhRafF
[b] https://goo.gl/7oCqBy

the intended usage of the collection.

2 Related Work

Collections of rumours from different natures have been assembled before. However, these datasets mainly focused on political and social issues [5, 6, 10]. Some of these datasets were built using rumours collected from social media platforms [5, 12, 13], such as Twitter, or with data extracted from fact-checking websites [6, 10], like PolitiFact[3] and Snopes[4], while others were crafted using manually created claims based on Wikipedia articles [8].

The first large-scale dataset on rumour tracking and classification was proposed by Qazvinian et al. [5]. This dataset comprises manually annotated Twitter posts (tweets) from five political and social controversial topics, with tweets marked as *related* or *unrelated* to the rumour. It also provides annotations about users' beliefs towards rumours, i.e., users who endorse versus users who refute or question a given rumour. Twitter posts were also used in the construction of two datasets under the PHEME project [1]: PHEME dataset of rumours and non-rumours (PHEME-RNR) and PHEME rumour scheme dataset (PHEME-RSD). These two collections are composed of tweets from rumourous conversations associated with newsworthy events mainly about crisis situations. PHEME-RNR contains stories manually annotated as *rumour* or *non-rumour*, hence being suited for the rumour detection task [12]. On the other hand, PHEME-RSD was annotated for stance and veracity, following crowd-sourcing guidelines [11], and tracked three dimensions of interaction expressed by users: support, certainty and evidentiality [13]. PHEME-RSD contains conversations threads in English and in German, however this data is extremely imbalanced as less than 6% of the tweets are written in German.

Silverman developed a dataset for the analysis of how online media handles rumours and unverified information [6]. The resulting Emergent database is a collection of online rumours comprising topics about war conflicts, politics and business/technology. Emergent data was generated with the help of automated tools to identify and capture unconfirmed reports early in their life-cycle. Then, rumours were classified according to their headline and body text stances and monitored with respect to social network shares (Twitter, Facebook and Google Plus) and version changes. In a posterior work, Ferreira and Vlachos leveraged Emergent into a dataset focusing veracity estimation and stance classification [3]. This modified dataset consists of claims and associated news articles headlines. Claims were labelled with respect to their veracity, while the news articles headlines were categorised according to their stances towards the claim: *supporting*, *denying* or *observing*, if the article does not make assessments about the veracity the claim.

LIAR is another dataset that could be employed in fact-checking and stance classification tasks [10]. LIAR includes manually labelled short statements from PolitiFact. All the statements were obtained from a political context, either extracted from debates, campaign speeches, social media posts, news releases or interviews. Each statement was evaluated for its truthfulness and received a veracity label accompanied by the corresponding evidences.

Thorne et al. developed FEVER (Fact Extraction and VERification), a large-scale manually annotated dataset focusing on verification of claims against textual sources [8]. FEVER is composed of claims generated from modifications in sentences extracted from introductory sections of Wikipedia pages. The claims were manually classified as *supported*, *refuted* or *not enough info* (if no information in Wikipedia can support or refute the claim). Additional Wikipedia justification evidences concerning supporting and refuting sentences were also recorded.

The PHEME-RSD and FEVER datasets were employed in stance detection and veracity prediction

[3] http://www.politifact.com/
[4] https://www.snopes.com/

shared tasks. The PHEME-RSD was employed in the Semantic Evaluation (SemEval) 2017 competition, Task 8 RumourEval[5] [2]. In the first sub-task (Sub-task A: Stance Classification), participants were asked to analyse how social media users reacted to rumourous stories, while for the second sub-task (Subtask B: Veracity prediction), participants were asked to predict the veracity of a given rumour. In the FEVER shared task[6], participants should build a system to extract textual evidences either supporting or refuting the claim.

Despite the variety of existing datasets for rumour analysis, none of them addresses football rumours. Besides, rumours in multilingual environments like the European Football market are yet to be explored by the academic research community. For our purpose, however, it is extremely important to track how news about transfer rumours are covered by sports media in different languages, as UEFA transfers generally occur between distinct European countries. Hence, our proposal differs from the previous work as it introduces a new dataset comprising football transfer rumours in three different languages: English, Spanish and Portuguese.

3 Dataset building

The data included in FTR-18 was harvested during the 2018 Summer Transfer Window (STW18). Transfer windows are periods during the year in which football clubs can make international transfers. Every league has the right to choose two annual transfer windows, with the opening and closing dates being defined by the football associations of the league.

The collected material includes both transfer news and rumours and the corresponding reactions to these rumours on Twitter. As we develop a multilingual dataset containing news and comments in English, Spanish and Portuguese involving UEFA clubs, the harvesting is performed during a common period including the transfer windows of England, Spain and Portugal. Accordingly, data is being collected from June 24 until August 31, 2018.

FTR-18 is built in three stages. First, we browsed the media and social networks for an initial assessment of transfer rumours involving top UEFA clubs. Next, we performed both a semi-automated news articles harvesting and a Twitter crawling, selecting news and tweets related to the rumours, respectively. Finally, as transfer moves are confirmed (or not), we annotate rumour veracity.

[5]http://alt.qcri.org/semeval2017/task8/
[6]http://fever.ai/task.html

3.1 Football clubs selection

The first step in the creation of the FTR-18 is the selection of the clubs to follow during the Summer Transfer Window of 2018. We picked English, Spanish and Portuguese clubs that played the group stage in 2017-2018 UEFA Champions League. This decision was made based on the languages we are more familiar with. With this restriction, we decided to monitor the following clubs:

England: Chelsea, Liverpool, Manchester City, Manchester United, Tottenham Hotspur

Spain: Atlético Madrid, Barcelona, Real Madrid, Sevilla

Portugal: Benfica, Porto, Sporting CP

3.2 Transfer rumours selection

The next stage of the FTR-18 generation is the collection of transfer news. We selected transfer rumours involving the monitored clubs. The rumours included narratives of permanence or transfer moves concerning football players, that is, whether players are leaving or being hired by the monitored clubs. For consistency reasons, we restricted the rumour selection to players affiliated to clubs from countries having English, Spanish or Portuguese as their majority language. The rumours were manually tracked and extracted from many sports news sources.

For each rumour, we annotated the *target* player, the target's club by the opening of the 2018 Summer Transfer Window (denoted as *source*), and the rumoured *destination* club (See Figure 1). We considered rumours either about the transference of a target to another club (source $\neq$ destination) or about the permanence of the player in the current club (source = destination).

3.3 News articles harvesting

News harvesting involved searching and collecting news articles, published by different news organisations, related to the identified rumours in the lan-

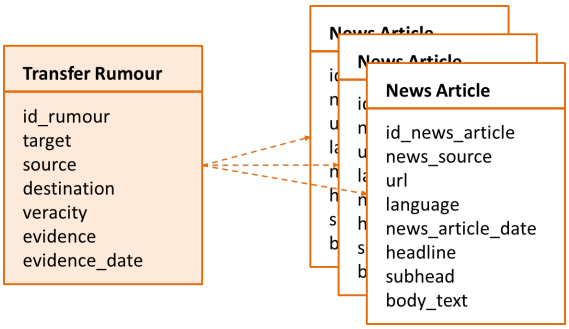

Figure 1: Transfer rumour and news articles attributes.

guages of the dataset. It was performed on a daily basis during the STW18. Each selected news article is associated with a single transfer move, and it might report a rumour or a verified information. We restricted our harvesting to news containing headlines with clear transfer reports about a single target, even if they report secondary moves in the body text (e.g. another player's alleged move, or interest manifested by other clubs). News with headlines pointing potential transfer of a player to multiple clubs were discarded.

For the news collection, we extracted content and meta-data from each article. Content information includes *headline*, *subhead*, and *body text*, while for the meta-data we registered the *article source*, *language*, *url* and the *publication date* (See Figure 1). The news harvesting task is semi-automated, as we could not build scrapers suited for all news sources due to their poor CSS structure.

3.4 Rumour reactions crawler

Once a rumour is identified, we build a crawler using the Twitter's streaming API[7] to collect posts related to the rumour. In compliance to the established language restrictions, we only harvest tweets written in the languages of the FTR-18 dataset. We record both the *message content*, *tweet creation date* and the data associated to the user who posted the message (*id*, *screen name*, *description*, *location*, *friends* and *followers count*, *verification account status*). If the post is a retweet, we also register the *retweet status*, the data associated to the user who posted the original message and the *retweet* and *favourite counts*.

Besides collecting tweets related to rumours, we also track Twitter account meta-data and followers statistics either for the players, for the monitored clubs and for the clubs involved in the transfer rumours.

3.5 Rumour veracity annotation

As the rumour unfolds, we annotate the rumour veracity by adding to the transfer rumour meta-data the evidences that support or refute the rumour and the publication date of these evidences. Rumour veracity is to be inferred until the end of 2018 Summer Transfer Window. Thus, for the case when source $\neq$ destination, the rumour veracity is stated as *True*, if the transfer is confirmed, or is assigned as *False*, if the target remains in the source club or if the target is transferred to a different destination during STW18. Similarly, for rumours in which source = destination, the rumour veracity is inferred as *False* if the player is transferred, or *True* otherwise.

[7] https://developer.twitter.com/en/docs/tweets/filter-realtime/overview.html

Table 2: Current status of FTR-18 dataset.

	News articles	Tweets
English	1,517	1,130K
Spanish	747	677K
Portuguese	781	257K
Total	3,045	2,064K

Despite the challenges of manually determining the credibility of a rumour [14], UEFA registration after STW18 will provide ground truth data for the annotation of the rumours' veracity.

4 Initial Analysis

Currently, the FTR-18 dataset comprises 3,045 news articles and more than 2,064K tweets (See Table 2). The news articles subset considers transfer news covered by online sports media written in English (1,517), Spanish (747) or Portuguese (781) by 96 different news organisations. This collection includes 304 transfer moves associated with 175 target football players. The Twitter subset contains original messages and retweets written in English (1,130K), Spanish (677K) or Portuguese (257K). The Twitter posts are related to 112 claimed transfer moves involving 84 different football players. The FTR-18 dataset meta-data and corresponding news source scrapers are available at `https://github.com/dcaled/FTR-18`.

In a preliminary analysis of the collected news articles, we noticed many cross-references between different news sources. This pattern has already been identified by Silverman [6] and Maia [4]. We observed a constant appearance of attribution formulations like "*source S reported*" or "*according to source S*", followed or not by the link to the referred news source. Other structures used to conceal the origin of the rumour are "*according to reports/sources*", "*are said*", "*player P linked to club C*", "*reports suggest*". Expressions like these appear in most of our collected news articles. Another common strategy is reporting an unverified information using news headline as a question [6] ("*Ronaldo out, Neymar in at Real Madrid?*[8]"). Although less common, this last example is often adopted, mainly by Spanish media.

Another interesting observation concerning transfer news is the cascading move chain associated to a rumour. For example, news sources condition the acquisition of a new player on the sale of another player by the same club ("*Cristiano Ronaldo open to Juventus move as Real Madrid consider selling to fund move for*

[8] https://www.theguardian.com/football/2018/jul/03/football-transfer-rumours-cristiano-ronaldo-real-madrid-juventus-neymar

Table 3: Datasets comparison (RD = rumour detection, RT = rumour tracking, SC = stance classification, VC = veracity classification, ER = evidence retrieval).

Dataset	Data source	Publicly Available	Multi-lingual	Usage				
				RD	RT	SC	VC	ER
Qazvinian's [5]	Social media	✗	✗		✓	✓		
Emergent [3, 6]	News & Rumour sites & Social media	✓	✗			✓	✓	
PHEME-RNR [12]	Social media	✓	✗	✓				
PHEME-RSD [13]	Social media	✓	✓	✓		✓	✓	
FEVER [8]	Wikipedia	✓	✗				✓	✓
LIAR [10]	Rumour sites	✓	✗			✓	✓	
FTR-18	News sites & Social media	✓	✓	✓	✓	✓	✓	

Kylian Mbappe or Neymar[9]"). This pattern drives the reader to expect a continuation of the rumourous story and its unfoldings, thus keeping audience's attention for follow-up stories.

5 Discussion

FTR-18 is suited for most of the steps involved in a rumour classification process, as presented by Zubiaga et al. [14]. Table 3 displays a comparison of FTR-18 against existing datasets, showing used data sources, availability, multilingual characteristics and possible usage. At this first phase of the dataset development, we perform annotations on rumours veracity. This task can be finished by the official closing date of STW18 (August 31, 2018), when all football transfers must be concluded. Any subsequent transfers are frozen until the next Winter Transfer Window (January 2019). Once each rumour is labelled on its truthfulness, rumour veracity assessment could be performed on the FTR-18 dataset.

In a second annotation phase, we will manually label our news articles subset, identifying which articles report rumours and which report verified information. This annotation will make FTR-18 suited for a rumour detection task. Stance detection is another possible application for the FTR-18 dataset. The collected content allows the classification of how the news headlines orient towards a given transfer rumour. Besides news' stances, we can evaluate the attitudes manifested by the audience with respect to a transfer move using reactions to Twitter posts. Both news articles labelling and stance annotation depend on human judgement.

The collected data also encourages analysis of rumour tracking. This research topic is scarcely explored and consists of identifying content associated with a rumour that is currently monitored [14]. We intend to label the harvested news and Twitter posts a posteriori as related or unrelated to the rumour, following Qazvinian et al's approach [5]. As result, FTR-18 could be used as a rumour dataset suited for binary classification of posts in relation to a rumour.

Finally, the FTR-18 dataset will also allow the analysis on how football transfer news are reported by the sports media. The news articles collection will serve as input for identifying the linguistics structures employed by journalists in transfer coverage. We expect to conduct an investigative work on the propagation patterns present in football transfer news, such as the recurrent conditional transfer moves and the echo effect caused by repetitions of unverified stories published by third-party news organisations. Further improvements in the FTR-18 dataset include the addition of new football leagues (e.g. CONMEBOL), the expansion of monitored clubs set and the collection of data from future transfer windows.

[9]https://www.independent.co.uk/sport/football/transfers/cristiano-ronaldo-transfer-news-latest-juventus-real-madrid-neymar-kylian-mbappe-unveil-move-a8434171.html

Acknowledgements

This work was supported by FCT, under grant No. UID/CEC/50021/ 2013.

References

[1] L. Derczynski and K. Bontcheva. Pheme: Veracity in Digital Social Networks. In *Proceedings of the 10th Joint ACL ISO Workshop on Interoperable Semantic Annotation (ISA)*, 2014.

[2] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. W. S. Hoi, and A. Zubiaga. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of SemEval*, 2017.

[3] W. Ferreira and A. Vlachos. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016.

[4] P. M. Maia. Jornalismo desportivo: mercado de transferências - relação entre jornalistas e fontes de informação. Master's thesis, Universidade Nova de Lisboa, Lisboa, Portugal, 2016.

[5] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011.

[6] C. Silverman. Lies, damn lies and viral content. 2015.

[7] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*, 2017.

[8] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018.

[9] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 2018.

[10] W. Y. Wang. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017.

[11] A. Zubiaga, M. Liakata, R. Procter, K. Bontcheva, and P. Tolmie. Crowdsourcing the annotation of rumourous conversations in social media. In *Proceedings of the 24th International Conference on World Wide Web*, 2015.

[12] A. Zubiaga, M. Liakata, and R. Procter. Learning reporting dynamics during breaking news for rumour detection in social media. *arXiv preprint arXiv:1610.07363*, 2016.

[13] A. Zubiaga, M. Liakata, R. Procter, G. W. S. Hoi, and P. Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 2016.

[14] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys*, 2018.