# Fact Checking: Detecting and Verifying Facts

## *Fact Checking: Detectando y Verificando Hechos*

**Bilal Ghanem**
Universitat Politècnica de València
bigha@doctor.upv.es

**Abstract:** With the uncontrolled increasing of fake news, untruthful claims, and rumors over the web, recently different approaches have been proposed to address the problem. To distinguish false claims from truthful ones may positively affect the society in different aspects. In this paper we describe the motivations towards fake news research topic in the recent years, we present similar and related research topics, and we show our preliminary work and future plans.

**Keywords:** Fact checking, rumors, fake news, disinformation

**Resumen:** Debido a su descrontrolado incremento, recientemente se ha comenzado a investigar y a proponer aproximaciones a la detección de rumores y noticias y afirmaciones falsas. Distinguir proposiciones falsas de las verdaderas puede afectar positivamente a la sociedad en diferentes aspectos. En este artículo describimos la motivación actual para investigar en el tema de las noticias falsas, presentamos temáticas similares y que están relacionadas, y mostramos nuestro trabajo preliminar y nuestros planes de futuro.

**Palabras clave:** Fact checking, rumores, noticias falsas, desinformación

## 1 Fake News

Fake news is an important topic so the research community become interested in its identification necessity. Fake news as a topic has been defined as: "An inaccurate, sometimes sensationalistic report that is created to gain attention, mislead, deceive or damage a reputation"[1]. Unlike misinformation, in fake news, authors have previous intention to pose a misleading sentence. Whereas in misinformation, authors have inaccurate or confused information about specific topic.

The spreading of this phenomenon has increased in a massive way in online sites. The openness of the web has led to increase the number of online news agencies, social media networks, and online blogs. These platforms allow certain organizations or individuals to pose fake news, because they guarantee them several attractive factors, such as privacy, free-access, availability, and large audience. In the same time, a large number of untrusted news agencies have appeared. These sites can affect the public opinions about specific issues, where they have political, social, or financial agendas.

In a recent approach (Nelson and Taneja, 2018), the authors investigated the user's behavior when visiting these sites. They examined online visitations data across different Internet devices. During the US elections in 2016, they found that the number of devices that visited trusted news sites was 40 times larger than fake news ones. This observation shows a hypothesis that the Internet users are aware of fake news and they are able to discriminate them from others. Later on, (Bond Jr and DePaulo, 2006) showed that humans can detect lies only 4% better than random chance. They analyzed more than 200 meta-data of trusted sites and showed that their good reputation may contribute to have more visitors. This study revealed that the users are not able to detect fake news and they are affected by the good reputation of online sites.

Improving the search engines ranks of untrusted sites might make them more popular, and maybe, truthful. Therefore, fake news sites' admins employ website spoofing or au-

[1] https://whatis.techtarget.com/definition/fake-news, visited in May 2018.

Lloret, E.; Saquete, E.; Martínez-Barco, P.; Moreno, I. (eds.) *Proceedings of the Doctoral Symposium of the XXXIV International Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*, p. 19–23 Sevilla, Spain, September 19th 2018.

thentic news styling technique to mimic the hight reputation of the authentic sites in an attempt to make their sites closely similar. The last US election has drawn a real fear in the American nation about fake news (Nelson and Taneja, 2018). The fast spreading of fake news motivated the owner of Wikipedia encyclopedia to create a news site called WikiTribune[2] to promote Evidence-based journalism.

## 2 *Rumors*

A similar research topic that has attracted the research community attention recently is rumors detection. Despite the different definitions of rumors in the literature, the most accepted is: "Unverified and instrumentally relevant information statements in circulation" (DiFonzo and Bordia, 2007). Both, rumors and fake news are similar research topics, although they have been tackled independently. Unlike fake news, rumors may turn out to be true, false, or partly true, wherein the time of posting the veracity is unknown. Rumors normally remain in circulation (ex. retweeted by Twitter users) until a trusted destination uncover or verify the truth. Another difference has been shown by (Zubiaga et al., 2018): rumors cannot just be classified by their veracity type (true, false, half-true), but also by the credibility degree (high or low). Previously, (Allport and Postman, 1947) studied rumors from a psychological perspective. They were interested in answering why people spread rumors in their environment. In that time (1947), it was difficult and complex to find a clear answer. But in the recent years, a renewed interests conducted to answer the question: people spread rumors when there is uncertainty, when they feel anxiety, and when the information is important.

According to (Zubiaga et al., 2018), rumors in literature have been studied from different perspectives: rumors detection (rumor or not), tracking (in social media, detecting posts dealt with these rumors), stance classification (how each user or post is oriented towards a rumor's veracity) and veracity prediction (true, false, or unverified). In rumourEval shared task at SemEval-2017 (Derczynski et al., 2017), the organizers proposed two different subtasks: stance classification, and veracity prediction. (Enayet and El-Beltagy, 2017) have achieved the highest result in veracity prediction. They used the percentage of replying tweets, the existence of hashtags, and the existence of URLs as features for a classification model. According to the previous tasks, the veracity prediction was the one more related to fake news detection.

## 3 *Fact Checking in Presidential Debates*

Fake news gained attention in the 2016 US presidential elections, where both Democrats and Republicans blamed each other for spreading false information. According to a public poll, many people after the election believed that fake news had affected the election results. Even during the presidential debate, journalists found that there were many false claims between candidates. Their goal was to weaken or to build bad reputation of the contenders. These false claims could draw real effect on the elections results. Especially, presidential debates are long by their nature and to detect false claims may need more time than the needed to spread these false claims among the public. Also, these debates contain large number of sentences accused by each candidate and most of these claims are un-factual claims (opinions). These opinions composed another challenge to the journalists to filter them from the factual claims. This real issue has taken attention by CLEF-Lab 2018. In this lab, two shared tasks have been proposed: check worthiness and factuality. In the following, we will give brief review about these two tasks, and show our preliminary approach

### 3.1 CLEF-2018 Check That Lab

As mentioned above, two different tasks have been proposed at CLEF-2018 Check That lab (Nakov et al., 2018). The first task was similar to rumor detection. The idea is to detect claims that are worthy for checking. The other task is complementary, where the factuality of these factual claims is needed to be checked.

**Task 1 - Check-Worthiness:** A set of presidential debates from the US presidential election is presented for the task, where each claim in the debate text has been tagged manually as worth to be checked or not. The full text of the debates is used in the task to

[2] https://www.wikitribune.com, visited in May 2018.

allow participants to exploit contextual features in the debates. The task goal is to detect claims that are worthy of checking and to rank them from the most worthy one for checking to the lowest. Our preliminary approach for this task (Ghanem et al., 2018b) was inspired from previous works proposed by (Granados et al., 2011) and Stamatatos (2017). The authors in the former work have used a text distortion technique to enhance thematic text clustering by maintaining the words that have a low frequency in documents. Similarly, in the latter work, the same text distortion technique was used for authorship attribution. The authors maintained the words that had the highest frequency in the documents to detect the author from his/her writing style. In this work, we used the same text distortion technique to detect worthy claims. We believed that this type of tasks is more thematic than stylistic, where the writing style is not as important as the thematic words. We have maintained the thematic words (that have the lowest frequency) using a ratio of C; the higher value of C is, the more thematic words are maintained. Also, we maintained a set of linguistic cue words (LC) that were used previously by (Mukherjee and Weikum, 2015) to infer the news credibility. Additionally, we maintained also named entities from being distorted, such as: Iraq, Trump, America. Through the manually checking of the claims, we found the checking worthy claims tended to list different types of named entities. After applying the distortion process, the new version of the text was used by Bag-of-Chars using the Tf-Idf weighting scheme. The new distorted text became less biased by the high frequency words, such as stopwords. For the ranking purpose, we used a method inspired from the K-Nearest Neighbor (KNN) classifier to rank these worthy claims based on the distance to the nearest neighbor. For the classification process, we used KNN classifier. During the training phase of the task, the Average Precision @N was used. In the testing phase, a set of testing files were provided by the organizers and the Mean Average Precision (MAP) measure was employed. In Table 1 the results for the task are presented. It is worth to mention that the task has been organized also in Arabic, where the English claims were translated manually. In the English part of the task, our approach (UPV-INAOE) has achieved the third position among seven teams. In the Arabic part, only two teams have submitted their results. Similarly to English, the results are close and there is not a big difference among them. We believed that the low result of our approach in the Arabic part is because we translated the LC lexicons automatically and a manual translation might have been more reliable.

| Team | English | Arabic |
|---|---|---|
| Prise de Fer | 0.1332 | - |
| Copenhagen | 0.1152 | - |
| UPV-INAOE | 0.1130 | 0.0585 |
| bigIR | 0.1120 | 0.0899 |
| Fragarach | 0.0812 | - |
| blue | 0.0801 | - |
| RNCC | 0.0632 | - |

Table 1: Official results for the Task1, released using the MAP measure

**Task 2 - Factuality:** As we mentioned above, this task concerns with detecting the factuality of the claims from the US presidential election. The claims that are unworthy for checking have not been annotated and kept in the debates to maintain the context. Factual claims have been tagged as True, False, and Half-True. These debates are provided in two languages, English and Arabic, similarly to the previous task. The macro F1 score was used as the performance measure.

Our approach for this task was based on the hypothesis that factual claims have been discussed and mentioned in online news agencies. In our approach (Ghanem et al., 2018a), we used the distribution of these claims in the search engines results[3]. Furthermore, we supposed that truthful claims have been mentioned more by trusted web news agencies than untruthful ones. Thus, our approach depended on modeling the returned results from search engines using similarity measures with the reliabilities of the sources. Our feature set consists of two types: dependent and independent features. For the dependent features, we used cosine over embedding between a claim query and each of the first N results from the search engines. We used the main sentence components to built the sentences embeddings, discarding stop-

[3] We used in our experiments both Google and Bing search engines.

| Team | English | Arabic |
|---|---|---|
| Copenhagen | 0.705 | - |
| FACTR | 0.913 | 0.657 |
| UPV-INAOE | 0.949 | 0.820 |
| bigIR | 0.964 | - |
| Check It Out | 0.964 | - |

Table 2: Official results for the Task2, released using the MAE measure

words. Also, for each result, we extracted the AlexaRank value for its site, to capture the reliability of the result source. Finally, another text similarity feature was used to measure the similarity using the full sentences, without using embeddings and discarding stopwords. From these set of features, we built other dependent features that capture the distribution of some of these previous ones. We used Standard Deviation and Average of the previous cosine similar values, and in a similar manner, for AlexaRank values.

The official results of task 2 are shown in Table 2. In general, the low results of both complex tasks can give an intuition of how much this research topic is. Further work is needed.

## 4 Future Work

Fact checking became recently an even more interesting research topic. We think that detecting worthy claims should be the first step in fake news identification. We will address these issues both in political debates and social media. To the best of our knowledge, the previous works on fact checking only concentrated on validating the facts using external resources. We will work on investigating facts from different aspects: linguistically, structurally, and semantically. In this vein, SemEval-2019 lab[4] proposes two tasks for the next year: first to determine rumor veracity and support for rumors, which is similar to the one that was proposed previously in SemEval-2017. Secondly, fact checking in community question answering forums, which is a new environment for investigating facts veracity. This shows the high interest of the research community in these two research topics. Therefore, participating in these tasks is one of our future plans.

[4] http://alt.qcri.org/semeval2019/index.php?id=tasks, visited on May 2018

## References

Allport, G. W. and L. Postman. 1947. The psychology of rumor. *Oxford, England: Henry Holt.*

Bond Jr, C. F. and B. M. DePaulo. 2006. Accuracy of deception judgments. *Personality and social psychology Review*, 10(3):214–234.

Derczynski, L., K. Bontcheva, M. Liakata, R. Procter, G. W. S. Hoi, and A. Zubiaga. 2017. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. *arXiv preprint arXiv:1704.05972.*

DiFonzo, N. and P. Bordia. 2007. Rumor, gossip and urban legends. *Diogenes*, 54(1):19–35.

Enayet, O. and S. R. El-Beltagy. 2017. Niletmrg at semeval-2017 task 8: Determining rumour and veracity support for rumours on twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 470–474.

Ghanem, B., M. Montes-y Gòmez, F. Rangel, and P. Rosso. 2018a. Upv-inaoe - check that: An approach based on external sources to detect claims credibility. In *Working Notes of CLEF 2018 – Conference and Labs of the Evaluation Forum*, CLEF '18, Avignon, France, September.

Ghanem, B., M. Montes-y Gòmez, F. Rangel, and P. Rosso. 2018b. Upv-inaoe - check that: Preliminary approach for checking worthiness of claims. In *Working Notes of CLEF 2018 – Conference and Labs of the Evaluation Forum*, CLEF '18, Avignon, France, September.

Granados, A., M. Cebrian, D. Camacho, and F. de Borja Rodriguez. 2011. Reducing the loss of information through annealing text distortion. *IEEE Transactions on Knowledge and Data Engineering*, 23(7):1090–1102.

Mukherjee, S. and G. Weikum. 2015. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 353–362. ACM.

Nakov, P., A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, L. Màrquez, W. Zaghouani, P. Atanasova, S. Kyuchukov, and G. Da San Martino. 2018. Overview of the CLEF-2018 CheckThat! Lab on automatic identification and verification of political claims. In P. Bellot, C. Trabelsi, J. Mothe, F. Murtagh, J. Nie, L. Soulier, E. Sanjuan, L. Cappellato, and N. Ferro, editors, *Proceedings of the Ninth International Conference of the CLEF Association: Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science, Avignon, France, September. Springer.

Nelson, J. L. and H. Taneja. 2018. The small, disloyal fake news audience: The role of audience availability in fake news consumption. *new media & society*, page 1461444818758715.

Stamatatos, E. 2017. Authorship attribution using text distortion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 1138–1149.

Zubiaga, A., A. Aker, K. Bontcheva, M. Liakata, and R. Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):32.