**FacTeR-Check: Semi-Automated Fact-Checking through Semantic Similarity and Natural Language Inference**

**Alejandro Martín**  
Universidad Politécnica de Madrid  
Madrid, Spain  
alejandro.martin@upm.es

**Javier Huertas-Tato**  
Universidad Politécnica de Madrid  
Madrid, Spain  
javier.huertas.tato@upm.es

**Álvaro Huertas-García**  
Universidad Politécnica de Madrid  
Madrid, Spain  
University Rey Juan Carlos de Madrid  
Madrid, Spain  
alvaro.huertas@upm.es

**Guillermo Villar-Rodríguez**  
Universidad Politécnica de Madrid  
Madrid, Spain  
guillermo.villar@upm.es

**David Camacho**  
Universidad Politécnica de Madrid  
Madrid, Spain  
david.camacho@upm.es

**Abstract**

Our society produces and shares overwhelming amounts of information through Online Social Networks (OSNs). Within this environment, misinformation and disinformation have proliferated, becoming a public safety concern in most countries. Allowing the public and professionals to efficiently find reliable evidences about the factual veracity of a claim is a crucial step to mitigate this harmful spread. To this end, we propose FacTeR-Check, a multilingual architecture for semi-automated fact-checking that can be used for either applications designed for the general public and by fact-checking organisations. FacTeR-Check enables retrieving fact-checked information, unchecked claims verification and tracking dangerous information over social media. This architectures involves several modules developed to evaluate semantic similarity, to calculate natural language inference and to retrieve information from Online Social Networks. The union of all these components builds a semi-automated fact-checking tool able of verifying new claims, to extract related evidence, and to track the evolution of a hoax on a OSN. While individual modules are validated on related benchmarks (mainly MSTS and SICK), the complete architecture is validated using a new dataset called NLI19-SP that is publicly released with COVID-19 related hoaxes and tweets from Spanish social media. Our results show state-of-the-art performance on the individual benchmarks, as well as producing a useful analysis of the evolution over time of 61 different hoaxes.

**Keywords** Misinformation, Transformers, COVID-19, Hoax, Natural Language Inference, Semantic Similarity

1 Introduction

Misinformation and disinformation are two terms that resound since a long time. Inaccurate information has been largely used for varied purposes for decades and centuries. However, the emergence of Internet, Online Social Networks and Instant Messaging Services has undoubtedly facilitated its rapid creation and diffusion. These two terms reflect
a problem that continues to expand and which involves an increasing concern to society. Yet, there are important differences between both terms: while misinformation involves inaccurate information propagated without knowing it is false, disinformation involves disseminating deliberately false information in order to deceive people.

The COVID-19 pandemic has undoubtedly drawn attention to this problem, when misinformation and disinformation meet health and affect public safety. From the initiation of this pandemic, an incessant repetition of falsehoods has been generated and propagated, undermining the work of health authorities in the fight against COVID-19. False reports about its origin, its death rate, or about vaccines have been a constant threat to control this virus.

Fact-checking organisations are on the forefront combating the propagation of false claims, where intensive work is done to deny hoaxes that circulate through different channels, such as Online Social Networks (OSNs), Instant Messaging Services or Mass Media. The verification process conducted by these companies is mostly carried out by hand, however, it is barely reflected in OSNs. Users of these platforms share fake information without even realising it is indeed a falsehood or deliberately posting false claims without further consequences.

Recent advances in Natural Language Processing, such as the Transformer architecture [1], allow to deal with complex human language for a plethora of tasks, such as summarization, translation, sequence classification, question answering or context-aware sentences similarity evaluation. The embeddings generated by this type of models for a piece of text, a vector representation composed of hundreds of

In this research, we leverage the most recent advances in Natural Language Processing to develop a semantic-aware multilingual Transformer-based architecture for semantic similarity evaluation, semi-automated fact-checking and tracking of information pieces in Online Social Networks. We present an architecture that, on the one hand, can help general public in checking the veracity of a claim (i.e. a tweet) through context-aware automated comparison against a databases of hoaxes. On the other hand, our proposal aims at providing useful tools for fact-checking organisations to track and monitor hoaxes circulating in OSNs.

In contrast to previous approaches previously proposed, our tool relies on a semi-automated fact-checking process, using fact-checkers databases as source of verified claims. This ensures the quality of the predictions of the model, instead of relying on training sets of false data that severely limit the capacity of the model to detect the most recent. Another major difference lies in the context-aware and multilingual capacities we introduce due to the use of the Transformer architecture, a very important advance to deal with human language understanding and to allow comparisons between different languages without translation. The multilingual capacity will help to do fact check no matter the language of the candidate claim and the verified facts is. Finally, we also integrate a tracking module to analyse the whole propagation cascade of the hoax, a very valuable tool to explore its whole story in a social network.

To validate and to show the capabilities of the architecture proposed, we use the COVID-19 pandemic scenario in Spanish speaking countries. We manually selected 61 hoaxes related to Covid-19 and extracted related tweets using Twitter API. Our architecture allows to label the degree of entailment of these tweets with a hoax, providing a useful insight of the propagation of hoaxes in Spanish on Twitter throughout one year.

In summary, this research presents the following contributions:

Figure 1: Diagram showing the two possible usage flows of FacTeR-Check.
FacTeR-Check

- A labelled dataset of Spanish tweets IDs with a degree of entailment against a list of 61 hoaxes.
- A context-aware multilingual semantic similarity method for searching hoaxes with high similarity to a given query.
- A Natural Language Inference model for semi-automated fact-checking. This model allows to check if there is an entailment, contradiction or neutral relation between two statements.
- A deep insight of misinformation and disinformation circulating on Twitter related to Covid-19 in Spanish speaking countries during one year.

The remaining sections of this manuscript are organised as follows: Section 2 summarises a series of background concepts and the most relevant state-of-the-art works. Section 3 presents the whole architecture designed for semi-automated fact-checking. Section 4 reports the experiments conducted to evaluate the different modules that compose the FacTeR-Check architecture. Section 5 presents the dataset built in this research of hoaxes found in Twitter and publicly released in this research. Section 6 provides a deep analysis of the propagation of hoaxes related to Covid-19 in Spanish in Twitter.

2 Background and related work

In this section, a short selection of some relevant background work is presented together with an overview of the state-of-the-art literature. The section provides some recent contributions and works on transformers architectures (Sec 2.1), semantic (textual-based) similarity methods (Sec 2.2), natural language inference tasks (Sec 2.3), automated fact-checking (Sec 2.4), and misinformation tracking in OSN (Sec 2.5).

2.1 The Transformer architecture

In 2017, a group of researchers working at Google presented the Transformer [1], a novel network architecture based on the concept of attention to deal with complex tasks involving human language, such as translation. This architecture revolutionised the Natural Language Processing field, allowing to train models to address highly complex tasks efficiently. From then, an uncounted number of applications, architectures, and models have been published to address tasks such as sentiment analysis [2], text generation [3] or question answering [4]. However, the attention concept was also soon exported to other domains such as music generation [5] or image generation [6].

One of the most important characteristics of these architectures in the Natural Language Understanding field lies in their context-aware capabilities, enabling to perform tasks such as question answering with high performance. While in previous NLP statistical-based approaches words were treated independently without considering the existing relations between them in a sentence or a text, the attention-based mechanism of the Transformer architecture allows to consider these relations and to establish deep connections.

As in the case of other deep architectures such as Recurrent Neural Networks (RNNs) or Convolutional Neural Networks (CNNs), the Transformer involves a series of encoder and decoder layers that operate sequentially over the input. The goal of this architecture of this architecture is to obtain a vector representation called embedding of the input sentence as comprehensive as possible to later be used in specific tasks. For instance, BERT is a specific implementation of the Transformer architecture where the output for a given input is an embedding of 768 positions that define multiple characteristics of the input. Due to the large amount of data, execution time and computational resources required to train this kind of models, researchers usually employ pre-trained architectures that are later fine-tuned to solve specific tasks.

A plethora of architectures have been proposed implementing the attention-based mechanism since it was proposed. Models such as BERT [7], Roberta [8], XML [9] or XLM-RoBERTa [10] are being used in a large number of NLP tasks with great success.

2.2 Semantic Textual Similarity

Measuring the degree of similarity between a pair of texts is a problem that has attracted the attention of many researchers for many years from the natural language processing and information retrieval fields. The complexity of this task has resulted in a variety of approaches to obtain similarity measures able to consider the higher number possible of characteristics. Classical approaches relying on lexical based information have been largely used for this task, however, they are extremely limited, since they do not allow to compare the real semantic value [11]. These methods fail to detect similarity between synonyms and they do not consider the existing relations between words of a sentence. Gomaa and Fahmy [12] proposed a taxonomy of similarity methods. String-based similarity methods operate with string and
2.4 Automated Fact-Checking

Automated Fact-Checking (AFC) involves different tasks and issues, such as extracting check-worthy claims from a speech or a large text, building fact-checking tools based on previously checked facts or to evaluate at what level a claim can be considered true. These AFC methods typically integrate Machine Learning techniques, however, researches have also highlighted the limitations of these approaches due to the training set used or the detection of paraphrasing. Nevertheless, recent advancements in this field, mainly because of the development of architectures using the attention-based mechanism, have led to important progress in the area.

Typically, Automated Fact-Checking is usually conducted through NLP models. There are different approaches to address this task according to the inputs. One possibility is to derive the veracity of a claim without further knowledge or context, an approach highly unreliable. Similarly, a multi-source approach has been proposed to combine
different information sources [34]. Other researchers leverage knowledge to reach more reliable decisions. FEVER is a dataset of claims extracted from Wikipedia and manually labelled as Supported, Refuted or NotEnoughInfo [35]. Hanselowski et al. [36] made public another dataset for automated fact-checking, with validated claims and documents annotated. WikiFactCheck-English [37] contains claims, context information and evidence documents. A comparative transformer-based approaches for misinformation detection is presented by Huertas et al. [38].

These datasets are usually employed to train machine learning-based tools for AFC to later classify news claims without considering recent knowledge [39]. From another point of view, literature can also be organised according to how technology helps fact-checkers. An analysis study by Nakov et al. [40] identifies several tasks: searching for check-worthy claims, identifying already fact-checked claims, searching for evidences or providing automated fact-checking services.

In terms of specific implementations for AFC, Naderi and Hirst [41] use linguistic features and a classifier in a statement multi-class classification task. Karadzhoj et al. propose the use of LSTM networks to classify claims in combination with relevant fragments of text from external sources [42]. Kotonya et al. [43] provide a broad analysis of the state-of-the-art literature of automated fact-checking approaches that are focused on explainability. Other important implementation is ClaimBuster [44], which monitors live discourses and detects claims that are present in a repository, however limited details are provided regarding its implementation and there is no mention to the use of context-aware semantic models. More recent approaches have made use of the Transformer architecture. Stummbach and Ash [45] use GPT-3 to generate a summary of evidences for a fact check decision. The attention-based mechanism has been also used for the identification of check-worthy statements [46]. BERT has been also used for veracity prediction and explanation generation in a public health domain [43].

2.5 Misinformation tracking in OSNs

Online Social Networks (OSNs) are the perfect environment for a fast and uncontrolled growth of misinformation and disinformation. The effects produced by the complex opinion dynamics that occur in these platforms such as *polarisation*, *echo-chambers*, *peer pressure* or *social influence* hinder the process of analysing the propagation of a false claim. Monti et al. [48] propose the use of Geometric Deep Learning to detect false claims in Online Social Networks, an approach which allows to take into consideration the propagation as a graph. A similar approach is followed by FakeDetector [49], in this case using a graph neural network and explicit and latent features to represent both text, creators and subjects. With a different objective, researchers have propose the use of transformers for profiling hate speech in Twitter [50].

The fight against misinformation in Online Social Networks has also been explored from an author perspective, modelling user profiles and their characteristics according to the probability to trust or distrust false claims [51 52].

3 Fighting misinformation through Semantic Similarity and Natural Language Inference

FacTeR-Check aims at helping in the whole verification process, analysis and tracking of false claims mainly circulating on social networks. Our tool implements an interconnected architecture with multilingual and deep human language understanding capabilities, substantially differing from previous completely automated but limited methods proposed in the literature relying on an initial immutable knowledge base. These methods used to train a machine learning classifier which fail when zero-shot prediction is performed, that is to say, when a claim which has never been verified by fact-checkers is presented. Instead, given the undeniable need to provide answers based on updated information sources, FacTeR-Check leverages the work already being conducted by fact-checking organisations to validate new claims. This semi-automated fact-checking process implies a close joint working between computational intelligence experts and fact-checking organisations.

Besides, FacTeR-Check not only helps during the fact-checking process, but also in the collection and analysis of the whole history of a hoax, automatising the process of obtaining a broad oversight of its propagation over time. This is a powerful instrument to fight against mis- and disinformation spreading on social networks. FacTeR-Check provides four different main functionalities (see Fig. 2):

1. **Multilingual Semantic similarity evaluation**: For each new claim received, the architecture searches for semantically-similar hoaxes verified by fact-checkers in a database constantly updated. We make use of an ensemble of Transformer models to generate a representation embedding for each claim present in the database and for the one received as input. Then, a similarity distance is used to calculate the most similar hoaxes.

2. **Multilingual Natural Language Inference**: Once a selection of similar hoaxes is presented, an NLI modules calculates the entailment probability with the input claim. If a coincidence is found (an entailment probability
exceeds a certain threshold), the input claim is consider as false information. This module also allows to detect if the input claim denies or contradicts the hoax.

3. **OSN automated retrieval**: In order to study the level of spread and presence of the hoax on a particular Online Social Network, a query containing a series of relevant keywords is created and send it to the API of the OSN. This enables to collect posts or tweets of users related to a false claim to be tracked. This step includes two transformer-based models for keyword extraction and Named Entity Recognition.

4. **Misinformation tracking in OSNs**: Based on the three previous functionalities, it is possible to extract a pool of claims from OSNs and to filter those which replicate and support a false claim used as input. This module allows to analyse a large set of posts or tweets according to their creation date, user or other metadata.

The four functionalities described enable two different workflows, as shown in Fig. 2. One is intended to provide a useful mechanism for a semi-automated fact verification, checking claims against a database of facts verified by fact-checking organisations. This workflow requires a semantic similarity module for filtering facts according to a certain degree of similarity and a second step of Natural Language Inference, to detect if there is textual entailment.

The second workflow is designed to aid them in the process of monitoring and tracking the life of a false claim in an Online Social Network. This involves to extract relevant keywords and named entities from the claim to build a search query which is sent to the API of the OSN in order to extract tweets or posts presenting content related to the input claim. The semantic similarity and NLI modules allow then to filter all the data to keep tweets or claims actually supporting the false claim. Next subsections describe in detail each functionality.

### 3.1 Semantic Similarity

Semantic is the level of language that deals with the meaning of a sentence by focusing on word-level interactions. This research aims to infer and understand information from texts in order to tackle misinformation by comparing sentence embeddings that condense the semantic level of language. In contrast to previous approaches focused on statistical natural language processing, FacTeR-Check implements semantic and context-aware semantic similarity evaluation. Through the use of semantic-aware and context-aware models, the goal is to evaluate the degree of similarity between a new claim against and a database of fact-checked claims. The result will be a subset of fact-checked claims ensuring a certain degree of similarity.

To measure the semantic similarity between texts, the cosine similarity function can be used. This metric takes advantage of the text representation as a vector in a high-dimensional space to compute the semantic and contextual proximity of the text representation as a vector in a high-dimensional space to compute the semantic and contextual proximity between a pair of texts, an operation which enables to assess their semantic similarity. The cosine distance between two sentence embeddings \( u \) and \( v \) is a variant of the inner product of the vectors normalised by the vectors’ L2 norms, as shown in equation (1):

\[
\text{CosSim}(u, v) = \frac{\sum_{i=1}^{N} u_i v_i}{\sqrt{\sum_{i=1}^{N} u_i^2} \sqrt{\sum_{i=1}^{N} v_i^2}} = \frac{\langle u, v \rangle}{\|u\| \|v\|}
\]

where \( N \) represents the number of dimensions composing the sentence embeddings \( u \) and \( v \), \( \langle u, v \rangle \) is the inner product between the two vectors, and \( \| . \| \) is the L2 norm.

With the goal of building an accurate representation of each sentence, an ensemble approach has been adopted. The potential of this type of method to combine word embeddings has been assessed in the state-of-the-art literature [53, 54].
showing that a mixture of embeddings featuring different characteristics leads to more robust representations and better performance than single embedding-based methods. Besides, a further advantage of ensemble methods is the expansion of vocabulary coverage.

In the ensemble proposed, the output is calculated by concatenating the embeddings of four well-known multilingual models available at Sentence-Transformers\(^2\), all of them fine-tuned on MSTS\(^3\), a multilingual extended version of the Semantic Textual Similarity Benchmark (STSb)\(^4\). Typically, Semantic Textual Similarity (STS) tasks include examples composed of a pair of sentences and a score ranging between 0 and 5 according to the degree of similarity. STS Benchmark\(^4\) comprises a selection of the English datasets used in the STS tasks between 2012 and 2017 from SemEval. In order to work on a multilingual scenario, in this work, the STS Benchmark was extended to 15 different languages using the Google translator library.

---

\(^{2}\)https://www.sbert.net/
\(^{3}\)https://github.com/Huertas97/Multilingual-STSB
\(^{4}\)http://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark

---

**Figure 3:** Architecture for the evaluation of information pieces against hoaxes already identified by fact checkers. A first step allows to retrieve hoaxes that are semantically similar to the input text. In the second step, a Natural Language Inference model measures the degree of entailment against each hoax retrieved in step 1.

**Figure 4:** Ensemble and dimensionality reduction approach proposed. Concatenation of embeddings from four multilingual sentence-transformers models applying PCA dimensionality reduction.
FacTeR-Check

- **paraphrase-xlm-r-multilingual-v1**: Distilled version of RoBERTa trained on large-scale paraphrase data using XLM-R as the student model.
- **stsb-xlm-r-multilingual**: Distilled BERT version trained in NLI and STSb using XLM-R as the student model.
- **paraphrase-multilingual-MiniLM-L12-v2**: Multilingual version of the MiniLM model from Microsoft trained on large-scale paraphrase data.
- **paraphrase-multilingual-mpnet-base-v2**: Distilled version of the MPNet model from Microsoft fine-tuned with large-scale paraphrase data using XLM-R as the student model.

These pre-trained models are fine-tuned on MSTSB using Cosine Similarity Loss from Sentence Transformers. To obtain the best results and avoid overfitting, we optimized the following hyperparameters using the grid search method: learning rate, epochs, batch size, scheduler, and weight decay. The selected hyperparameter values and the resulting model have been published at HuggingFace.

As explained by Sidorov et al. [59], cosine similarity applied to a pair of N-dimensional vectors has both time and memory $O(N)$ complexity. That is, time and memory grow linearly with the number of dimensions of the vectors under comparison. This is the main drawback of the use of ensemble models on semantic search with sentence embedding. To address this issue, the Principle Component Analysis (PCA) is computed and applied to the whole architecture as shown in Fig. 4. This enables to reduce dimensionality, removing redundant information across embeddings while retaining the most relevant information in a new $N$-dimensional space.

In order to maximise efficiency, the embedding of each fact-checked claim is precalculated. When receiving a new fact-checked claim, its embedding representation will be obtained applying the models of the ensemble and the PCA to the concatenated outputs and saved into the fact-checked claims database. This will allow to easily evaluate new claims, calculating the cosine distance to each fact-checked claim stored.

### 3.2 Natural Language Inference

Once a top-k corpus of hoaxes above a specific degree of semantic similarity has been identified, Natural Language Inference is used to infer the relation between the new input statement (hypothesis) and each fact-checked claim (premise). This relation may be entailment, contradiction or neutral. While semantic similarity is unable to detect finer nuances, a NLI model is able to detect a entailment or contradiction relationship given a pair of sentences. If we manage to detect if an statement entails a hoax, we can safely assume that the statement supports the hoax and therefore contains misinformation. Nevertheless, it is important to mention that Language Inference is not aware of the intentionality behind an statement, an issue which is not addressed in this research.

To better describe the NLI task, let $\langle p, h \rangle$ be a sentence pair of hoax and statement. Using language inference we can infer contradiction and neutral probabilities, however, our main focus is on finding the degree of entailment. We formally want to find if $h$, our statement, is a hoax $h_f$ or we are unable to determine the nature of the statement $h_u$. Formally we want to approximate Eq. 2:

$$ f(p, h) \approx P(p|h_f) $$

where $p$ is a hoax or fact-checked claim verified by fact-checkers and we have certainty that involves fake information, $h$ is the verifiable statement found by semantic similarity and $h_f$ is the event in which the statement contains misinformation. Therefore, our purpose is to find a suitable function $f$ that is able to approximate this probability. Finding $P(p|h_f)$ is equivalent to finding the probability of the entailment of $\langle p, h \rangle$. On the other hand we can safely say that $1 - P(p|h_f) = P(p|h_u)$ as the contradiction and neutrality of $\langle p, h \rangle$ does not give a meaningful explanation for $h$.

In order to find $f$, the transformer model XLM-RoBERTa-large is chosen. Transformer models for NLI have problems when transferring to unseen domains [60], so special consideration is given to the fine-tuning process. To train this network, two datasets are used, XNLI and SICK. The inner transformer model XLM-R is fine-tuned first on XNLI. In this case, we used the model available at the Huggingface transformers repository. After this step, a classification head is added to the model, which includes a) a global average pooling of the last hidden state of the transformer model, b) a linear layer with 768 neurons and tanh activation, c) a 10% dropout for training and d) a classifier linear layer with softmax. This classification head is trained on the SICK dataset, freezing the XLM-R weights to preserve previous pre-training. This is optimized using Adam optimizer with 0.001 learning rate. The best weights are decided on the validation subset of SICK.

---

5Fine-tuned models available in [https://huggingface.co/AIDA-UPM](https://huggingface.co/AIDA-UPM)

6[https://huggingface.co/joeddav/xlm-roberta-large-xnli](https://huggingface.co/joeddav/xlm-roberta-large-xnli)
3.3 Semi-automated (S-AFC) fact-checking through Natural Language Inference and Semantic Similarity

In this work, we propose a 2 steps process to perform semi-automated fact-checking (S-AFC). The semantic similarity and Natural Language Inference modules described in the two previous sections (3.1 and 3.2) are the pillars of this S-AFC process. The first step allows to filter an entire database of fact-checked statements or hoaxes, retrieving those that present semantic similarities with the new input claim. As a result, an ordered list by the degree of similarity is obtained, and the top \( k \) results are selected. Then, the NLI module allows to perform language inference between the input claim and each candidate hoax in the top-\( k \) result. If a fact-checked claim is found to correlate the input claim with enough certainty, the new claim is labelled.

This two-step process (see Fig. 1) is highly useful for different purposes. In addition to a semi-automated fact-checking of new claims that need to be checked, the combination of semantic similarity and Natural Language Inference can be used to analyse the evolution and presence of a particular statement in a large amount of data. For instance, in an Online Social Network such as Twitter, it is possible to filter thousands of tweets seeking for those that endorse or reject the statement.

3.4 Automated tracking of hoaxes in Twitter

The massive volume of information present on social media platforms makes it unmanageable to track and monitor hoaxes’ evolution manually. For this reason, we propose an automatic social media tracking method based on the generation of search queries composed of keywords and search operators. These keywords are employed to extract information, such as tweets or posts, related to a given claim from the API of a social network. All the data download will offer an extensive view to study the evolution of a piece of misinformation.

The use of keywords is due to the limitations that the API of these OSN impose. While searching for a given statement will only deliver tweets or posts replicating almost exactly the original input claim, the use of keywords aims to increase this search space and to obtain a wider picture. The method used for automatic keyword extraction is adapted from KeyBERT [62]. KeyBERT is a keyword extraction technique that uses semantic-aware Transformer-based models to compute word and tweet embeddings and cosine similarity to find the most semantically similar words to the tweet. Accordingly, the most similar words are the keywords that best describe the tweet meaning.

Our proposal, named FactTeR-ChecKey, uses our multilingual MSTSb-paraphrase-multilingual-mpnet-base-v2 model as the semantic-aware model. To optimise the multilingual keyword extraction, stopwords are removed by detecting the language with CLD2 and removing the appropriate stop words with the NLTK toolkit [63]. Additionally, the bert-spanish-cased-finetuned-ner from Hugging Face is included as the Name Entity Recognition (NER) model for Spanish. This NER model is applied only in Spanish, so the keyword extraction tool remains multilingual.

![Figure 5: Examples of query building from English and Spanish hoaxes for searching through Twitter API.](image)

4 Evaluation of the FacTeR-Check architecture

In this section, the Semantic Similarity, Natural Language Inference and keyword extraction (FactTeR-ChecKey) modules are evaluated using different benchmark datasets from the state-of-the-art literature. The following subsections describe in detail the results obtained for each task.
FacTeR-Check

(a) Single fine-tuned models

Figure 6: Number of components selection in MSTSB development set. Average Spearman Correlation Coefficient of the single fine-tuned models (a), and ensemble architectures (b) using cosine similarity for the 15 languages as a function of the number of components from the extended STS-Benchmark development set. The average of correlation coefficients is computed by transforming each correlation coefficient to a Fisher’s z value, averaging them, and then back-transforming to a correlation coefficient.

| Model + PCA | Dimensions | EN-EN | EN-ES | ES-ES | Avg |
|-------------|------------|-------|-------|-------|-----|
| Model + PCA | EN-EN | EN-ES | ES-ES | Avg |
| MSTSB_paraphrase-multilingual-MiniLM-L12-v2 | 84.92 | 85.71 | 81.04 | 81.04 | 83.08 | 83.28 | 81.03 | 81.02 |
| MSTSB_stsb-xlm-r-multilingual | 84.35 | 85.11 | 82.84 | 83.17 | 83.39 | 83.89 | 81.85 | 82.08 |
| MSTSB_paraphrase-xlm-r-multilingual-v1 | 84.79 | 85.50 | 82.73 | 82.97 | 83.38 | 83.58 | 82.23 | 82.39 |
| MSTSB-paraphrase-multilingual-mpnet-base-v2 | 86.69 | 87.27 | 84.21 | 84.28 | 84.93 | 85.19 | 83.20 | 83.28 |
| Ensemble 2 | 85.91 | 86.72 | 83.49 | 83.69 | 84.42 | 84.68 | 83.12 | 83.28 |
| Ensemble 3 | 86.64 | 87.55 | 84.50 | 84.80 | 85.24 | 85.72 | 83.85 | 84.21 |
| Ensemble 4 | 86.77 | 87.78 | 85.00 | 85.52 | 85.56 | 86.20 | 84.24 | 84.71 |

Table 2: Spearman ρ and Pearson r correlation coefficient between the sentence representation from multilingual models with PCA dimensionality reduction and the gold labels for STS Benchmark test set.
4.1 Semantic similarity evaluation

The multilingual STS Benchmark (generated with Google Translator) has been used for the evaluation of the semantic similarity module. The overall results in the test sets are shown in Table 1. While the EN-EN column refers to the original STS Benchmark dataset, EN-ES, ES-ES are calculated using the translated version. These results reveal that the best performance is obtained with the fine-tuned MSTSb-paraphrase-multilingual-mpnet-base-v2 model. This table also presents the results obtained with different combinations of the models. The best Ensemble of only 2 models is composed of the concatenation of MSTSb_paraphrase-xlm-r-multilingual-v1 and MSTSb_paraphrase-multilingual-MiniLM-L12-v2. Ensemble 3 adds MSTSb-paraphrase-multilingual-mpnet-base-v2 model while and Ensemble 4 includes all models reaching a maximum of 2688 dimensions. Surprisingly, only Ensemble 3 exceeds the best-fit model at the cost of incorporating more than twice as many dimensions.

As expected, the use of ensemble based approaches increases dramatically the number of dimensions. In order to tackle this problem, Principal Component Analysis (PCA) is used to reduce dimensionality. PCA is a data transformation and dimensionality reduction method that finds a subspace that explains most of the data variance while keeping attractive properties, such as removing linear correlation between dimensions and avoiding irrelevant dimensions with low variance. On the other hand, PCA is an unsupervised method that does not guarantee that the new feature space will be the most appropriate for a supervised task. To cope with this disadvantage, a total of 90K parallel sentences representing 15 languages and extracted from three well-known resources (TED2020, WikiMatrix and OPUS-NewsCommentary) are used to fit the PCA for each model. The relation between performance obtained and reduction size is shown in Fig. 6. As can be seen, both in the case of single fine-tuned models and ensemble architectures, the performance converges with less than 200 principal components, which provides a substantial space reduction. The best PCA space is selected according to the MSTSb development set average performance across languages.

Table 2 shows the results after combining PCA and the ensemble approach, proving that this dimensionality reduction methods leads to better performance, reducing dramatically the number of dimensions. An illuminative example is Ensemble 4, which reduces from 2688 to 429 dimensions after applying PCA with the highest scores across all languages. This method not only reduces up to six times the initial dimensions of the ensemble, but it also requires fewer dimensions than most of the single models. This demonstrates that ensemble approaches in combination with dimensionality reduction techniques allow to build accurate and efficient semantic textual similarity models.

4.2 Performance of the Natural Language Inference module

The NLI module is in charge of determining the relation between two statements (a fact-checked statement) and a new input claim. This relation, which can be either entailment, contradiction or neutral, will be based on different probabilities. Thus, a threshold has to be defined in order to assign the final label. The most likely scenario is one with a large database of fact-checked claims verified by fact-checkers. Once a new claim has to be checked, it will be compared with the NLI module against those verified claims existing in the database above a certain degree of semantic similarity. As result, if enough degree of entailment is found, the new input claim will be labelled according to the verified claim found.

We evaluate our approach using the testing subset provided by SICK. A well-known collection of pairs of sentences with entailment, contradiction and neutral relation. Results are presented on Table 3. For comparison, we include the results of two benchmark methods: GenSen and InferSent. In case of GenSen, it achieves 87.8% accuracy while InferSent reaches 0.863. Our proposed approach reaches 87.7% accuracy while maintaining the multi-lingual capabilities of XLM-RoBERTa, which is useful to contrast information from culturally separated hoaxes. This is represented in the Spanish and interlingual sections of Table 3, where the same metrics are computed. We observe a slight drop in quality, mostly due to SICK being mono-lingual, though Spanish and inter-lingual results are quite robust on their own with 82.9% and 85.3% accuracy respectively. We want to highlight the high accuracy attained by the module when mixing languages, allowing for international tracking of misinformation.

4.3 Performance of the keywords extraction module

In order to evaluate the benefits of FactTeR-CheckKey, our approach is compared against two baseline methods in a general and a Twitter-specific scenario. The two baseline methods selected for this comparison are the statistical

---

1. https://pypi.org/project/pycld2/
2. The languages used in this scenario are: ar, cs, de, en, es, fr, hi, it, ja, nl, pl, pt, ru, tr, zh
3. https://www.ted.com/participate/translate
FacTeR-Check

| Language | Label            | Precision | Recall  | F1-score | Support |
|----------|------------------|-----------|---------|----------|---------|
| english  | CONTRADICTION    | 0.9158    | 0.7486  | 0.8238   | 712     |
|          | ENTAILMENT       | 0.8475    | 0.8946  | 0.8704   | 1404    |
|          | NEUTRAL          | 0.8856    | 0.9022  | 0.8938   | 2790    |
| Summary  | Macro Avg.       | 0.8830    | 0.8484  | 0.8627   | 4906    |
|          | Weighted Avg.    | 0.8791    | 0.8777  | 0.8770   | 4906    |
|          | Accuracy         | 0.8777    | -       | -        | 4906    |
| spanish  | CONTRADICTION    | 0.8511    | 0.7388  | 0.7910   | 712     |
|          | ENTAILMENT       | 0.7446    | 0.9031  | 0.8162   | 1404    |
|          | NEUTRAL          | 0.8797    | 0.8451  | 0.8461   | 2790    |
| Summary  | Macro Avg.       | 0.8251    | 0.8190  | 0.8178   | 4906    |
|          | Weighted Avg.    | 0.8369    | 0.8292  | 0.8296   | 4906    |
|          | Accuracy         | 0.8292    | -       | -        | 4906    |
| inter    | CONTRADICTION    | 0.8825    | 0.8737  | 0.8072   | 1424    |
|          | ENTAILMENT       | 0.7925    | 0.8989  | 0.8423   | 2808    |
|          | NEUTRAL          | 0.8828    | 0.8586  | 0.8705   | 5580    |
| Summary  | Macro Avg.       | 0.8526    | 0.8337  | 0.84     | 9812    |
|          | Weighted Avg.    | 0.8569    | 0.8534  | 0.8533   | 9812    |
|          | Accuracy         | 0.8534    | -       | -        | 9812    |

Table 3: Results for the SICK test set. Spanish results are extracted from machine translations of the SICK test set. Interlingual results are made from pairing interchangeably Spanish and English prompts.

Rapid Automatic Keyword Extraction (RAKE) algorithm and the multilingual version of KeyBERT which use paraphrase-xlm-r-multilingual-v1 as semantic-aware model. RAKE is a well-known statistical method for keyword extraction based on the collocation and co-occurrence of words by eliminating stopwords and punctuation, not taking into account any semantic information for the extraction process. On the other hand, KeyBERT incorporates state-of-the-art Transformer models for keyword extraction. The evaluation task consists on extracting keywords from the 60 Spanish hoaxes used previously in this project. Figure 5 provides an overview of the hoaxes data and the queries built for searching through the Twitter API. The queries are built concatenating the different keywords extracted with the “AND” logical operator. Precision, recall, and F1 score are the metrics used to evaluate the ability to extract keywords compared to manually extracted keywords.

Due to the differences between a general search engine and a the Twitter search API, which entails several restrictions, we have evaluated the performance of FacTeR-ChecKey in both. While a common search engine such as Google allows rich queries and provides flexibility when using verbs as input, the Twitter search API is very restricted and only searches for the exact words used in the input. In the first stage of the project, in which hoax-related information was extracted with manually extracted keywords, it was observed that verbs limited the information retrieved due to these limitations. Therefore, verbs were removed from the Spanish keywords extracted for the Twitter scenario and an additional POS tagging filter was applied to the automatic keyword extracted. The POS tagging filter is performed using Spacy, and the best model is selected from three possible models: small, medium, and large. It is noteworthy to highlight that although the automatic keyword extraction method is only evaluated on Spanish hoaxes, it can be easily extended to other languages.

Table 4: Evaluation of the keywords extraction module in general purpose tasks.

| Keyword Model               | Precision | General Recall | F1-score |
|-----------------------------|-----------|----------------|----------|
| RAKE Spanish                | 0.4683    | 0.9503         | 0.6139   |
| KeyBERT Multilingual        | 0.5694    | 0.8989         | 0.6200   |
| FactTeR-ChecKey Multilingual| 0.6283    | 0.8267         | 0.6988   |

Our technique clearly has an advantage over RAKE and KeyBERT approaches both in general (see Table 4) and Twitter-specific scenarios (see Table 5), where verbs are not considered. One advantage of FactTeR-ChecKey is that the type of information retrieved can be regulated by building queries from more specific to more general. Specific queries include all extracted keywords and gradually become more general as the terms are iteratively excluded from the query

https://twitter.com/search-advanced?lang=en
based on the similarity score. For this reason, our method has many practical applications. From already checked hoaxes, it is possible to extract information related to other hoaxes and to evaluate the check-worthiness of new hoaxes.

5 NLI19-SP: A Spanish Natural Language Inference dataset of hoaxes in Twitter

One of the goals of this research has been to build a dataset of tweets spreading misinformation claims detected and verified by fact-checkers. We have selected Twitter as the target OSN due to its large number of users, the availability of an API and the intensive movement of both information, misinformation and disinformation. Besides, our dataset is focused on misinformation spread in Spanish. To build such dataset, we have followed a four-step process:

1. **Hoaxes collection:** We gathered a pool of 61 hoaxes identified by fact-checker organisations.
2. **Search queries generation:** It is necessary to build representative queries with keywords to retrieve tweets to the hoaxes from Twitter API.
3. **Tweets retrieving:** By using FacTeR-ChecKey, we built a search query for each of the hoaxes in order to download tweets related to them from the Twitter search API.
4. **Filtering by semantic similarity:** We apply the semantic similarity module to filter tweets semantically related to each hoax.
5. **Natural Language Inference labelling:** The NLI module is applied to label the tweets according to their relation with the original hoax, detecting those that support or contradict the false claim.

The result of applying this pipeline is a pool of semantically-similar tweets for each hoax labelled as *entailment*, meaning that the tweet endorses the false claim, *contradiction* or *neutral*.

For the extraction of false claims already identified by fact-checkers we used LatamChequea Coronavirus, a database of misinformation about COVID-19 detected by 35 Spanish-language organisations and coordinated by Chequeado, and based on the global database created by the International Fact-checking Network (IFCN). Among all the indicators in this database, the variable used for our purpose will be the title of each false post registered. Given that the NLI and semantic similarity modules require the false claim to be expressed as clearly as possible, redundant words such as "hoax" or "message" that refer to the hoax itself are discarded.

The second step involves the generation of search queries for each hoax through the FacTeR-ChecKey module. These search queries are then used through the Twitter API to find posts that are sharing that type of disinformation. Each search query generated was later manually enhanced to retrieve the maximum number of tweets spreading that false information. Each resulting query is composed of potential keywords from that falsehood, linked by search operators and the use of parentheses to improve the results.

Furthermore, each set of keywords was optimised by adding synonyms and similar expressions to catch different ways to express the same piece of false information, because a hoax does not have to be propagated with the same words in the social network. This enables the collection of variants of the same hoax from different Hispanic geographical areas and avoids the implementation of a biased search of tweets from a unique Hispanic country.

The third step defines the automated search on Twitter API by using the search queries generated. This search is limited to the time period between the 1st of January 2020 to the 14th of March 2021. Moreover, replying tweets matching

11 [https://chequeado.com/latamcoronavirus/](https://chequeado.com/latamcoronavirus/)
the query have not been excluded, since they can also misinformation. The result of this process comprises 61 queries selected for the automated search from reported hoaxes and tweets collected through them thanks to Twitter API. Appendix I shows the hoaxes in Spanish but and the English translation.

In the next step the dataset has been curated using the semantic similarity module to filter tweets that actually present semantic similarity with the identified hoax. Finally, the Natural Language Inference component is applied to label each tweet as entailment, contradiction or neutral according to the relation with the original hoax statement as presented by the fact-checkers. In accordance with Twitter regulations and in order to guarantee users’ privacy, users and texts will not be published.

6 Misinformation spread in Spanish tweets: an analysis of Covid-19 hoaxes

In this section, our goal is to analyse how misinformation has spread in Twitter during the COVID-19 pandemic. For this purpose, we use the NLI19-SP dataset presented in the previous section. Each tweet in the dataset receives a label (entailment, contradication or neutral) according to its relation with the most similar hoax. Additionally, tweets by Twitter accounts of fact-checkers have been also identified. All this information allows to infer relevant patterns and characteristics of misinformation and disinformation claims spread during the pandemic. To narrow the analysis, we focus on messages written in Spanish. Fig. 9 shows the distribution of tweets found according to the fact-checker nationality that was used to identify the hoax. Although there is an important number of tweets collected from hoaxes identified by Spanish fact-checkers, no big differences were found between Spanish speaking countries.
Figure 9: Map showing the number of tweets supporting a hoax according to the nationality of the fact-checker that has identified it. In the case of France, although it is not a Spanish speaking country, several hoaxes have been identified by Factual AFP fact-checker, a France agency.

Fig. 7 shows cumulative distribution plot for a general overview of the tweets collected that support the different hoaxes, represented with different colours. One of the most relevant conclusions that can be extracted from this analysis lies in the shared patterns among the different hoaxes, exhibiting a clear trend towards waves of misinformation. This behaviour reflects how misinformation inevitably feeds itself and how spreaders operate in a coordinated fashion, giving rise to waves of misinformation and disinformation. This phenomenon is also worth considering when taking steps to counter the propagation of misinformation. Besides, the large representation of specific hoaxes is also an important element to study. Thus, one of the most disseminated hoax (Hoax 31 in Table 7) is that “masks cause hipoxia”. The large number of tweets found supporting this false claim is the reason of the big wave centred on June 2020. Similarly, the peak located at April 2020 y mainly due to the hoax “Christine Lagarde said that the elderly live too long”.

In order to better visualise the distribution of tweets supporting hoaxes, in Fig. 8 the same plot is displayed without including the hoax 31, which concentrates large part of the tweets. Although the big wave disappears in this new plot, reflecting that it was caused by the hoax removed, one can see how the waves are still visible, evidencing the common behavioural patterns that describe how misinformation circulates.

For a deeper analysis of misinformation circulating during the Covid-19 pandemic, Fig. 10 shows the temporal distribution of tweets supporting a selection of hoaxes and tweets published by fact-checker Twitter accounts. In four cases, hoaxes 28, 37, 50 and 60, the campaign launched by fact-checking organisation resulted in a higher number of tweets countering the hoax that tweets actually supporting the hoax. For the rest of hoaxes analysed, fact-checkers started a very timid response. However, in case of the hoax 15, no presence of fact-checkers denying the hoax can be appreciated, a false claim stating that “The definition of pandemic was changed in 2009 by the WHO”. This reveals how complex this scenario and that further research is required in order to help fact-checkers to detect and to undertake activities to avoid the spreading of false claims. In any case, it must be taken into consideration that the response must be proportionate, avoiding an excessive response that could increase the dissemination of the hoax and amplify its effects.

7 Conclusion

In this article we have proposed FacTeR-Check to mitigate OSN misinformation. Our architecture provides two pipelines, one for semi-automated verification of claims; another for tracking known hoaxes on social media. The pipelines share three modules: a semantic similarity module, a NLI module and a information retrieval module. By using context-aware semantic similarity, we are able to find related fact-checks, while NLI allows to contrast the claim against reputable sources. This double process enables to perform semi-automated fact-checking. On the other hand, in order to track hoaxes, we retrieve tweets related to a hoax, filtering the most relevant tweets with semantic similarity and contrasting them with the original hoax, finding how this particular piece of misinformation has spread on a social media
Figure 10: Comparative for different hoaxes between the distribution of tweets supporting a specific hoax and tweet by fact-checkers rejecting it.
FacTeR-Check

platform. While our validation has been limited to COVID-19 and Twitter we want to emphasise that our architecture is adaptable to other knowledge domains as well as other social networks.

For the evaluation, we first assess each model individually. Then the modules are put together in both pipelines to test their joint performance. To begin with, the similarity module offers above average performance using multilingual models on the STS benchmark. The NLI module uses XLM-RoBERTa fine-tuned on XNLI and the SICK training dataset, which performs adequately on SICK test, offering similar results to state-of-the-art models, as well as offering multilingual capabilities. Finally, the information retrieval module is compared against KeyBERT and RAKE on a dataset of Spanish keywords from our gathered hoaxes. Using this architecture we built a dataset for misinformation detection using NLI in Spanish about COVID-19, as well as track a selection of hoaxes to analyse their spread. FacTeR-Check proves to extract insightful information about the spread of many hoaxes, showing aggregate frequency peaks matching COVID-19 waves in Spain. Identified hoaxes have their own particular activity peaks, some have more longevity than others, others are used much more; they are extremely diverse in lifetime and popularity.

In contrast to previous approaches, FacTer-Check relies on external databases to operate. If a rumour reaches the verification pipeline, and there is no related fact-check retrievable on the topic, only similar articles will be retrieved. This means that the verification pipeline is as robust as the fact-check database. Alternatives may include composing a massive database of hoax embeddings, as well as a dynamic information retrieval process to detect new hoaxes and calculate their embeddings. The architecture has been tested on OSNs, meaning that it is blind to outside information such as news sites or other valuable sources of information. If a piece of disinformation is published outside of the OSN, it will be out of the scope of the tracking algorithm. Finally, information is varied, coming in many shapes and forms, including text but also audio, video or images; the verification and tracking pipeline can only work on textual data, meaning that there is room for building systems that support other formats.

Acknowledgement

This work has been supported by the research project CIVIC: Intelligent characterisation of the veracity of the information related to COVID-19, granted by BBVA FOUNDATION GRANTS FOR SCIENTIFIC RESEARCH TEAMS SARS-CoV-2 and COVID-19, by the Spanish Ministry of Science and Innovation under FightDIS (PID2020-117263GB-100) and XAI-Disinfodemics (PLEC2021-007681) grants, by Comunidad Autónoma de Madrid under S2018/TCS-4566 grant, by European Comission under IBERIFIER - Iberian Digital Media Research and Fact-Checking Hub (2020-EU-IA-0252), and by “Convenio Plurianual with the Universidad Politécnica de Madrid in the actuation line of Programa de Excelencia para el Profesorado Universitario”.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in neural information processing systems, 2017, pp. 5998–6008.
[2] U. Naseem, I. Razzak, K. Musial, and M. Imran, “Transformer based deep intelligent contextual embedding for twitter sentiment analysis,” Future Generation Computer Systems, vol. 113, pp. 58–69, 2020.
[3] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “Bertscore: Evaluating text generation with bert,” arXiv preprint arXiv:1904.09675, 2019.
[4] W. Yang, Y. Xie, A. Lin, X. Li, L. Tan, K. Xiong, M. Li, and J. Lin, “End-to-end open-domain question answering with bertserini,” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), 2019, pp. 72–77.
[5] N. Zhang, “Learning adversarial transformer for symbolic music generation,” IEEE Transactions on Neural Networks and Learning Systems, 2020.
[6] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, “Image transformer,” in International Conference on Machine Learning. PMLR, 2018, pp. 4055–4064.
[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2019.
[8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” 2019.
[9] A. Conneau and G. Lample, “Cross-lingual language model pretraining,” Advances in Neural Information Processing Systems, vol. 32, pp. 7059–7069, 2019.
[10] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised cross-lingual representation learning at scale,” *arXiv preprint arXiv:1911.02116*, 2019.

[11] R. Mihalcea, C. Corley, C. Strapparava *et al.*, “Corpus-based and knowledge-based measures of text semantic similarity,” in *Aaai*, vol. 6, no. 2006, 2006, pp. 775–780.

[12] W. H. Gomaa, A. A. Fahmy *et al.*, “A survey of text similarity approaches,” *international journal of Computer Applications*, vol. 68, no. 13, pp. 13–18, 2013.

[13] E. Millar, D. Shen, J. Liu, and C. Nicholas, “Performance and scalability of a large-scale n-gram based information retrieval system,” *Journal of digital information*, vol. 1, no. 5, 2000.

[14] J. Singthongchai and S. Niwattanakul, “A method for measuring keywords similarity by applying jaccard's, n-gram and vector space,” *Lecture Notes on Information Theory*, vol. 1, no. 4, 2013.

[15] S. Dennis, T. Landauer, W. Kintsch, and J. Quesada, “Introduction to latent semantic analysis,” in *25th Annual Meeting of the Cognitive Science Society. Boston, Mass*, 2003, p. 25.

[16] P. Shrestha, “Corpus-based methods for short text similarity,” in *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles. REncontres jeunes Chercheurs en Informatique pour le Traitement Automatique des Langues (articles courts)*, 2011, pp. 1–6.

[17] M. Schuhmacher and S. P. Ponzetto, “Knowledge-based graph document modeling,” in *Proceedings of the 7th ACM international conference on Web search and data mining*, 2014, pp. 543–552.

[18] N. Peinelt, D. Nguyen, and M. Liakata, “tbert: Topic models and bert joining forces for semantic similarity detection,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7047–7055.

[19] P. Kasnesis, R. Heartfield, X. Liang, L. Toumaniédis, G. Sakellari, C. Patrikakis, and G. Loukas, “Transformer-based identification of stochastic information cascades in social networks using text and image similarity,” *Applied Soft Computing*, vol. 108, p. 107413, 2021.

[20] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” *arXiv preprint arXiv:1908.10084*, 2019.

[21] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, “Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation,” *arXiv preprint arXiv:1708.00055*, 2017.

[22] M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, and R. Zamparelli, “A sick cure for the evaluation of compositional distributional semantic models,” in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), May 2014, p. 216–223. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/363_Paper.pdf

[23] B. MacCartney, *Natural language inference*. Stanford University, 2009.

[24] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. R. Bowman, and N. A. Smith, “Annotation artifacts in natural language inference data,” *arXiv preprint arXiv:1803.02324*, 2018.

[25] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, “A large annotated corpus for learning natural language inference,” *arXiv preprint arXiv:1508.05326*, 2015.

[26] A. Williams, N. Nangia, and S. Bowman, “A broad-coverage challenge corpus for sentence understanding through inference,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 1112–1122.

[27] A. Conneau, R. Rinott, G. Lampe, A. Williams, S. R. Bowman, H. Schwenk, and V. Stoyanov, “Xnli: Evaluating cross-lingual sentence representations,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018.

[28] Q. Chen, X. Zhu, Z.-H. Ling, S. Wei, H. Jiang, and D. Inkpen, “Enhanced lstm for natural language inference,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 1657–1668.

[29] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, “Supervised learning of universal sentence representations from natural language inference data,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 670–680.

[30] J. Huertas-Tato, A. Martín, and D. Camacho, “Sml: a new semantic embedding alignment transformer for efficient cross-lingual natural language inference,” *arXiv preprint arXiv:2103.09635*, 2021.
[31] D. Graves, “Understanding the promise and limits of automated fact-checking,” 2018.
[32] J. Thorne and A. Vlachos, “Automated fact checking: Task formulations, methods and future directions,” in Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 3346–3359.
[33] M. Granik and V. Mesyura, “Fake news detection using naive bayes classifier,” in 2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON). IEEE, 2017, pp. 900–903.
[34] H. Karimi, P. Roy, S. Saba-Sadiya, and J. Tang, “Multi-source multi-class fake news detection,” in Proceedings of the 27th international conference on computational linguistics, 2018, pp. 1546–1557.
[35] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, “Fever: a large-scale dataset for fact extraction and verification,” arXiv preprint arXiv:1803.05355, 2018.
[36] A. Hanselowski, C. Stab, C. Schulz, Z. Li, and I. Gurevych, “A richly annotated corpus for different tasks in automated fact-checking,” in Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), 2019, pp. 493–503.
[37] A. Sathe, S. Ather, T. M. Le, N. Perry, and J. Park, “Automated fact-checking of claims from wikipedia,” in Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 6874–6882.
[38] Á. Huertas-García, J. Huertas-Tato, A. Martín, and D. Camacho, “Civic-upm at checkthat! 2021: integration of transformers in misinformation detection and topic classification,” 2021.
[39] N. Naderi and G. Hirst, “Automated fact-checking of claims in argumentative parliamentary debates,” in Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), 2018, pp. 60–65.
[40] N. Hassan, F. Arslan, C. Li, and M. Tremayne, “Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster,” in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 1803–1812.
[41] Á. Huertas-García, A. Martín, J. Huertas-Tato, and D. Camacho, “Profiling hate speech spreaders on twitter: Transformers and mixed pooling,” in CLEF (Working Notes) 2021, 2021.
[55] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, “SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation,” in Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 1–14. [Online]. Available: https://www.aclweb.org/anthology/S17-2001

[56] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised Cross-lingual Representation Learning at Scale,” arXiv, Nov 2019. [Online]. Available: https://arxiv.org/abs/1911.02116v2

[57] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, “Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers,” 2020.

[58] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, “Mplnet: Masked and permuted pre-training for language understanding,” 2020.

[59] G. Sidorov, A. Gelbukh, H. Gómez-Adorno, and D. Pinto, “Soft similarity and soft cosine measure: similarity of features in vector space model,” Computación y Sistemas, vol. 18, no. 3, 2014.

[60] A. Talman and S. Chatzikyriakidis, “Testing the generalization power of neural network models across nli benchmarks,” arXiv:1810.09774 [cs], May 2019, arXiv: 1810.09774. [Online]. Available: http://arxiv.org/abs/1810.09774

[61] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” arXiv, Dec 2014. [Online]. Available: https://arxiv.org/abs/1412.6980v9

[62] M. Grootendorst, “Keybert: Minimal keyword extraction with bert.” 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4461265

[63] W. Wagner, “Steven bird, ewan klein and edward loper: Natural language processing with python, analyzing text with the natural language toolkit: O’reilly media, beijing, 2009, isbn 978-0-596-51649-9,” Language Resources and Evaluation, vol. 44, no. 4, pp. 421–424, 2010.

[64] H. Schwenk, V. Chaudhary, S. Sun, H. Gong, and F. Guzmán, “Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia,” 2019.

[65] J. Tiedemann, “Parallel data, tools and interfaces in OPUS,” in Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12). Istanbul, Turkey: European Language Resources Association (ELRA), 2012, pp. 2214–2218.

[66] S. Subramanian, A. Trischler, Y. Bengio, and C. J. Pal, “Learning general purpose distributed sentence representations via large scale multi-task learning,” in International Conference on Learning Representations, 2018.

[67] S. Rose, D. Engel, N. Cramer, and W. Cowley, “Automatic keyword extraction from individual documents,” in Text Mining. Chichester, UK: John Wiley & Sons, Ltd, 2010, pp. 1–20.

[68] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, “spaCy: Industrial-strength Natural Language Processing in Python,” 2020. [Online]. Available: https://doi.org/10.5281/zenodo.1212303
| Id | Hoax (in Spanish) | Hoax (in English) | Fact-checkers |
|----|------------------|-------------------|--------------|
| 1  | La PCR no distingue entre coronavirus y gripe | PCR tests do not distinguish between coronavirus and the flu | Newtral.es |
| 2  | Las vacunas de ARN-m contra el coronavirus nos transforman en seres transgénicos | mRNA vaccines against coronavirus transform us into transgenic beings | Animal Político, Maldita.es, Newtral.es |
| 3  | La vacuna contra la COVID-19 se crea con células de fetos abortados | COVID-19 vaccines are made of cells from aborted fetuses | Agencia Ocote, Agência Lupa, Chequeado, ColombiaCheck, Maldita.es, Newtral.es |
| 4  | Merck asocia las vacunas contra la COVID-19 con un genocidio | Merck associates COVID-19 vaccines with a genocide | Ecuador Chequea, Newtral.es |
| 5  | Una imagen relaciona la prueba PCR con la destrucción de la glándula pineal en el Antiguo Egipto | An image links PCR tests to the destruction of the pineal gland | Maldita.es |
| 6  | La vacuna contra la COVID-19 produce parálisis facial | COVID-19 vaccines produce facial paralysis | Chequeado, Newtral.es |
| 7  | La primera ministra de Australia finge ponerse la vacuna contra la COVID-19 | Australia first minister pretends to get the COVID-19 vaccine | Agência Lupa, La Silla Vacía |
| 8  | La vacuna contra la COVID-19 produce convulsiones | COVID-19 vaccines produce seizures | Maldita.es, Newtral.es |
| 9  | Muere 53 personas en Gibraltar tras ponerse la vacuna contra la COVID-19 | 53 people dead after being vaccinated against COVID-19 in Gibraltar | Maldita.es, Newtral.es |
| 10 | Detienen en un Lidl de Gijón a 11 personas con COVID-19 | 11 people with COVID-19 arrested in Lidl supermarket in Gijón | Maldita.es, Newtral.es |
| 11 | Ya no existen las enfermedades respiratorias que no son COVID-19 | Respiratory diseases that are not COVID-19 do not exist anymore | Newtral.es |
| 12 | La ciudad de Rosario (Argentina) para la vacunación por los efectos adversos de la vacuna | The city of Rosario (Argentina) stops vaccination because of the adverse effects of the vaccine | Chequeado, Maldita.es |
| 13 | La OMS dice que llevar a los niños al colegio sirve como consentimiento para su vacunación | The WHO says that taking our children to school gives consent for their vaccination | Maldita.es |
| 14 | El Banco Mundial tenía planes para la COVID-19 desde 2017 | The World Bank had plans for COVID-19 since 2017 | Animal Político, Aos Fatos, Mala Espina Check |
| 15 | La vacuna contra la COVID-19 causa esterilidad | Pfizer vaccines cause sterility | Animal Político, Chequeado, ColombiaCheck, La Silla Vacía, Maldita.es, Newtral.es |
| 16 | Un estudio de 2008 financiado por la Comisión Europea ya incluía la COVID-19 | A study funded by the European Commission in 2008 already included COVID-19 | Newtral.es |
| 17 | La OMS dice que llevar a los niños al colegio sirve como consentimiento para su vacunación | The WHO says that taking our children to school gives consent for their vaccination | Maldita.es |
| 18 | En Países Bajos existe desde 2015 una patente de test de COVID-19 | A patent of COVID-19 test exists in the Netherlands since 2015 | Maldita.es, Newtral.es |
| 19 | Varios vacunados con la vacuna UQ-CSL contra la COVID-19 contraen el VIH | Several COVID-19 vaccinated people with UQ-CSL contracted HIV | Newtral.es |
| 20 | La primera ministra de Australia finge ponerse la vacuna contra la COVID-19 | Australia first minister pretends to get the COVID-19 vaccine | Agência Lupa, La Silla Vacía |
| 21 | La mascarilla produce enfermedades neurodegenerativas | Masks produce neurodegenerative diseases | Maldita.es, Newtral.es |
| 22 | En Países Bajos existe desde 2015 una patente de test de COVID-19 | A patent of COVID-19 test exists in the Netherlands since 2015 | Maldita.es, Newtral.es |
| 23 | La vacuna contra la COVID-19 es aún experimental porque está en fase 4 | Vaccines against COVID-19 are still experimental because they are in phase 4 | Animal Político, Maldita.es |
| 24 | El Banco Mundial tenía planes para la COVID-19 desde 2017 | The World Bank had plans for COVID-19 since 2017 | Animal Político, Aos Fatos, Mala Espina Check |
| 25 | La vacuna contra la COVID-19 destruye nuestro sistema inmunológico | Vaccines against COVID-19 destroy our immune system | Maldita.es, Newtral.es |
| 26 | Pirbright Institute patentó la COVID-19 en 2018 | Pirbright Institute patented COVID-19 in 2018 | Maldita.es |
| 27 | Las gargaras con agua y sal previenen o curan el coronavirus | Gargling with water and salt prevents or cures coronavirus | #NoComaCuento (La Nación), AFP, Chequeado, ColombiaCheck, Ecuador Chequea, Efecto Cocuyo, El Surtidor, La Silla Vacía, Maldita.es, Spondon Media, Verificador (La República) |
| 28 | La dieta alcalina previene o cura el coronavirus | Alcaline diets prevent or cure coronavirus | Agencia Lupa, Animal Político, Bolivia Verifica, Chequeado, ColombiaCheck, Cotejo.info, EFE Verifica, Ecuador Chequea, Efecto Cocuyo, #NoComaCuento (La Nación), La Silla Vacía, Mala Espina Check, Maldita.es, Newtral.es |
| 29 | El coronavirus fue fabricado en un laboratorio chino | Coronavirus was made in a Chinese lab | Chequeado, Ecuador Chequea, Estadío verifica |

Table 6: Relation of hoaxes 1 - 30.
| Id | Hoax (in Spanish) | Hoax (in English) | Fact-checkers |
|----|------------------|------------------|--------------|
| 31 | La mascarilla causa hipoxia | Masks cause hypoxia | Agencia Ocote, Agência Lupa, Animal Politico, Aos Fatos, Bolivia Verifica, Chequeado, ColombiaCheck, Corojito.info, EFE Verifica, Ecuador Chequea, Efecto Cocuyo, La Silla Vacía, Maldita.es, Newtral.es, Verificado, Verificador (La República) |
| 32 | El eucalipto previene o cura el coronavirus | Eucalyptus prevents or cures coronavirus | AFP |
| 33 | El matíco cura el coronavirus | Matíco cures coronavirus | Bolivia Verifica |
| 34 | El biomagnetismo mata el coronavirus | Biomagnetism kills coronavirus | Bolivia Verifica, Maldita.es |
| 35 | La hoja de guayaba previene o cura el coronavirus | Guava leaf prevents or cures coronavirus | Animal Politico, Bolivia Verifica, Maldita.es, Newtral.es |
| 36 | La NASA catalogó el dióxido de cloro como antídoto universal en 1988 | NASA catalogued chlorine dioxide as a universal antidote in 1988 | Animal Politico |
| 37 | El vino previene o cura el coronavirus | Wine prevents or cures coronavirus | Chequeado, EFE Verifica, Maldita.es, Newtral.es |
| 38 | La mascarilla causa la muerte por neumonía bacteriana | Masks cause death due to bacterial pneumonia | Maldita.es |
| 39 | La vitamina C previene o cura el coronavirus | Vitamin C prevents or cures coronavirus | AFP, Chequeado, EFE Verifica, Agência Lupa, Maldita.es, Verificado |
| 40 | La prueba de antígenos no sirve para el COVID-19 porque da positivo con Coca-Cola | Antigen tests are useless for COVID-19 because they test positive with Coca-cola | Maldita.es, Newtral.es |
| 41 | La homeopatía previene o cura el coronavirus | Homeopathy prevents or cures coronavirus | Chequeado, Mala Espina Check, Maldita.es, Periódismo de barrio / El Toque |
| 42 | La COVID-19, el MERS y el H1N1 coinciden con la instalación del 5G, 4G y 3G, respectivamente | COVID-19, MERS and H1N1 coincide with the installation of 3G, 4G and 5G, respectively | Poligrafo |
| 43 | Los indígenas protegen a los niños con hierbas y árboles frente a la COVID-19 | Indigenous groups protect their children from COVID-19 with herbs and trees | Ecuador Chequea |
| 44 | Los mosquitos transmiten el coronavirus de contagiados | Mosquitoes transfer coronavirus from infected people | Maldita.es |
| 45 | Beber agua o sorbos previene o cura el coronavirus | Drinking or sipping water prevents or cures coronavirus | #NoComaCuento (La Nación), AFP, Bolivia Verifica, ColombiaCheck, La Silla Vacía, Maldita.es, OjoPúblico |
| 46 | Mueren 55 personas en Estados Unidos tras vacunarse contra la COVID-19 | 55 people dead after being vaccinated against COVID-19 in the United States | EFE Verifica |
| 47 | Las mascarillas producen pleuresia y neumonía | Masks produce pneumonia and pleurisy | AFP |
| 48 | Las personas sanas llevan la mascarilla con la parte blanca hacia fuera para no contagarse de COVID-19 | Healthy people wear their masks with the white part on the outside not to get COVID-19 | Newtral.es |
| 49 | El SARS-COV-2 no cumple los postulados de Koch, Rivers e Inglis para considerarlo enfermedad y coronavirus | SARS-COV-2 does not fulfill Koch, Rivers and Inglis’ postulates in order to be considered as coronavirus and as a disease | EFE Verifica |
| 50 | Christine Lagarde dijo que los ancianos viven demasiado | Christine Lagarde said that the elderly live too long | Chequeado, ColombiaCheck, Ecuador Chequea, Maldita.es |
| 51 | La COVID-19 es una bacteria | COVID-19 is a bacteria | Animal Politico, Chequeado, ColombiaCheck, La Silla Vacía, Maldita.es, Verificador (La República) |
| 52 | Galicia aprueba una ley para aislar a los positivos COVID-19 en campos de concentración | Galicia approves a law to isolate COVID-19 positives in concentration camps | Maldita.es |
| 53 | Las ondas electromagnéticas del 5G propagan el coronavirus | 5G electromagnetic waves spread coronavirus | Chequeado, Ecuador Chequea |
| 54 | La OMS recomienda un test pulmonar para identificar el coronavirus | The WHO recommends a pulmonary test to detect the coronavirus | EFE Verifica |
| 55 | Las pandemias tienen lugar cada 100 años | Pandemics take place every 100 years | AFP, Animal Politico, ColombiaCheck, Verificador (La República) |
| 56 | El laboratorio de Wuhan tiene relación con Glaxo y Pfizer | Wuhan lab is related to Glaxo and Pfizer | Animal Politico, Chequeado, La Silla Vacía, Maldita.es, Newtral.es |
| 57 | El coronavirus desaparece a los 27 grados | Coronavirus disappears at 27 degrees | Bolivia Verifica, Convoca, Agência Lapa |
| 58 | Hubo 17000 y 26000 muertes más en 2019 y 2018 respectivamente que en 2020 | There were 17000 and 26000 more deaths in 2019 and 2018 respectively than in 2020 | Maldita.es, Newtral.es |
| 59 | El polisorbato 80 de la vacuna contra la gripe causa coronavirus | Polysorbate 80 in the flu vaccines cause coronavirus | EFE Verifica, Maldita.es |
| 60 | Detienen a Charles Lieber por crear y vender el coronavirus | Charles Lieber arrested for creating and selling coronavirus | #NoComaCuento (La Nación), AFP, Animal Politico, Efecto Cocuyo, Agência Lupa, Mala Espina Check, Maldita.es, Newtral.es |
| 61 | En Israel no mueren por coronavirus gracias a una receta de limón y bicarbonato | No deaths in Israel due to coronavirus thanks to a recipe with lemon and bicarbonate | Newtral.es, Verificado |

Table 7: Relation of hoaxes 31 - 61