Knowledge-Based Systems 251 (2022) 109265

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

FacTeR-Check: Semi-automated fact-checking through semantic similarity and natural language inference

Alejandro Martín *, Javier Huertas-Tato, Álvaro Huertas-García, Guillermo Villar-Rodríguez, David Camacho

Universidad Politécnica de Madrid, Spain

ARTICLE INFO

Article history:
Received 17 February 2022
Received in revised form 10 June 2022
Accepted 11 June 2022
Available online 20 June 2022

Keywords:
Misinformation
Transformers
COVID-19
Hoax
Natural language inference
Semantic similarity

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 evidence 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 and hoaxes propagation analysis that can be used to implement applications designed for both the general public and for fact-checking organisations. FacTeR-Check implements three different modules relying on the XLM-RoBERTa Transformer architecture to evaluate semantic similarity, to calculate natural language inference and to build search queries through automatic keywords extraction and Named-Entity Recognition. The three modules have been validated using state-of-the-art benchmark datasets, exhibiting good performance in all of them. Besides, FacTeR-Check is employed to collect and label a dataset, called NLI19-SP, composed of more than 40,000 tweets supporting or denying 60 hoaxes related to COVID-19, released publicly. Finally, an analysis of the data collected in this dataset is provided, which allows to obtain a deep insight of how disinformation operated during the COVID-19 pandemic in Spanish-speaking countries.

1. Introduction

Misinformation and disinformation are two terms that resound since a long time ago. These two forms of inaccurate information have been widely used for a variety of purposes for decades and centuries. However, the emergence of Internet, Online Social Networks and Instant Messaging Services have 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.[1]

The COVID-19 pandemic has definitely drawn attention to this problem, when misinformation and disinformation meet health and affect public safety. From the onset 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, death rates, or vaccine safety have been a constant threat to the control of this virus.

Fact-checking organisations are on the forefront of the fight against the propagation of false claims, where intensive work is done to validate new hoaxes and to retrieve evidences of information pieces 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, the large amount of new posts or tweets published on a daily basis causes this work to be 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.

In this research, we leverage the most recent advances in Natural Language Processing (such as the Transformer architecture [1]) 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 the general public to check the veracity of a claim (i.e. a tweet) through context-aware automated comparison against a databases of fact-checked claims. On the other hand, our proposal

* Corresponding author.
E-mail address: alejandro.martin@upm.es (A. Martín).

[1] https://dictionary.cambridge.org/es-LA/dictionary/english/disinformation.

https://doi.org/10.1016/j.knosys.2022.109265

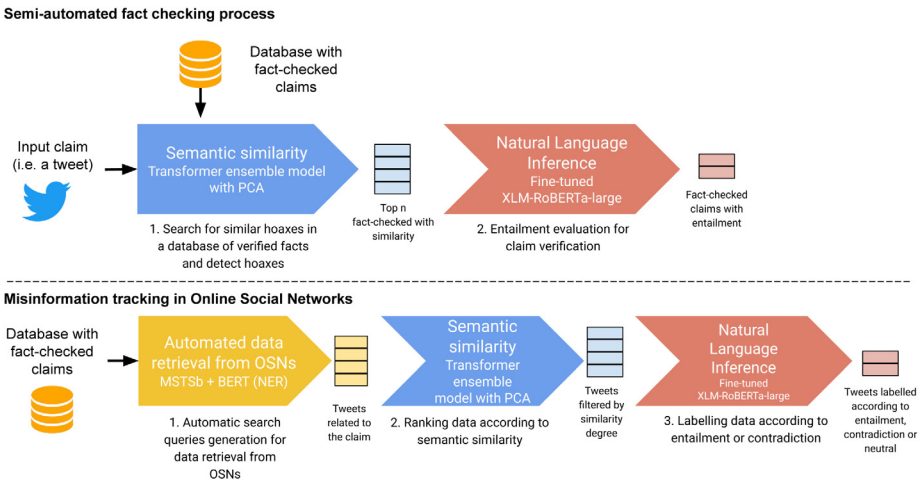

Fig. 1. Diagram showing the two possible usage flows of FacTeR-Check.

aims to provide useful tools for fact-checking organisations to detect and monitor hoaxes circulating in OSNs.

In contrast to previous approaches, our tool relies on a semi-automated fact-checking process, using fact-checkers' databases as source of verified information. 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 hoaxes. 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, regardless of the language of the candidate claim and the verified facts is. Finally, FacTeR-Check also allows to analyse the whole propagation cascade of the hoax, a very valuable tool to explore its whole story in a social network.

With FacTeR-Check, we aim to provide both the scientific community but also fact-checking organisations with a useful and powerful instrument to automatically retrieve tweets from Twitter related to a given false claim, to automatically filter by semantic similarity and label according to the alignment or contradiction against the false claim and to analyse the spreading of a hoax in this Social Network. Moreover, FacTeR-Check sets a significant improvement over existing approaches [2–5] through the use of fine-tuned language models that operate in a multilingual scenario. To the best of our knowledge, no existing research has proposed an architecture with all these functionalities and relying on advanced Transformer-based models. Finally, the architecture FacTeR-Check can be easily extended to integrate other information modalities, such as audio or image, given the use of embeddings to represent information.

We evaluate each FacTeR-Check module on state-of-the-art benchmark data sets, showing good performance in all the tasks tested. Furthermore, we use the proposed architecture to collect and analyse a large number of tweets supporting and denying 60 different hoaxes related to 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 labelling the degree of entailment of these tweets with a hoax, providing a useful insight into the propagation of hoaxes in Spanish on Twitter throughout one year.

In summary, this research presents the following contributions:

- A module to generate search queries for OSN's API composed of keywords, named entities, and logical operators.
- A module for context-aware multilingual semantic similarity evaluation, aimed at searching potentially related detected hoaxes or verified facts and filtering the candidate fact-claims for entailment.
- A module to perform semi-automated fact-checking through a natural language inference, that allows to check if there is an entailment, contradiction or neutral relation between two statements.
- A labelled dataset of Spanish tweets IDs with a degree of entailment against a list of 61 hoaxes.
- A deep insight into the misinformation and disinformation circulating on Twitter related to COVID-19 in Spanish-speaking countries for 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 on Twitter and publicly released in this research. Section 6 provides a detailed analysis of the propagation of hoaxes related to COVID-19 in Spanish on Twitter. Finally, Section 7 presents a summary of the work, the results obtained, and different lines of future work.

2. Background and related work

In this section, a selection of relevant background work is presented together with an overview of the state-of-the-art literature. The section presents some recent contributions and works focused on the transformer architecture (Section 2.1), semantic (textual-based) similarity methods (Section 2.2), natural language inference tasks (Section 2.3), automated fact-checking (Section 2.4), and misinformation tracking in OSNs (Section 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 able to address highly complex Natural Language Understanding (NLU) tasks efficiently. From then, an uncounted number of

applications, architectures, and models have been published to address tasks such as sentiment analysis [6], text generation [7] or question answering [8]. However, the attention concept was soon also exported to other domains, such as music generation [9] or image generation [10].

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 excellent results. 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, a key point to address complex NLU tasks.

As in the case of other deep architectures such as Recurrent Neural Networks (RNNs) or Convolutional Neural Networks (CNNs), the Transformer architecture involves a series of encoder and decoder layers that operate sequentially over the input. The goal 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 [11] 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 model, 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 [11], Roberta [12], XLM [13] or XLM-RoBERTa [14] 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 highest possible number 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 comparing the real semantic value [15]. These methods fail to detect similarity between synonyms, and they do not consider the existing relations between words of a sentence. Gomaa and Fahmy [16] proposed a taxonomy of similarity methods, where we can find string-based similarity methods operate with string and characters sequences or ngrams [17,18]. Corpus-based methods use large sets of words and texts and metrics such as latent semantic analysis [19] or building term vectors [20]. Knowledge-based methods allow to use the semantic content to provide more accurate comparisons, usually employing semantic networks [21]. The fourth category is made up of hybrid solutions that combine different methods [15].

The proposal of using an attention-based mechanism and its implementation into the Transformer architecture has meant a turning point. The embeddings generated with this type of architecture of a sentence or a text allow to build a rich multidimensional space where multiple characteristics are represented. Once the embedding vector of each document to be compared is obtained, a spatial distance such as cosine similarity can be used to measure the degree of similarity. Pre-trained models can be used for this purpose. However, if these models do not provide enough precision, they can be fine-tuned in a specific domain, thus allowing for more accurate similarity calculation. When training these models in a multilingual scenario, they generate a common features space for all languages represented in the training data, thus enabling comparing texts in different languages. This capability has revolutionised the research field of Natural Language Processing.

However, building precise models implies to narrow the application domain, specialising in a specific task but losing generalisation ability. As an example, transformers such as BERT have been combined with topic models to better deal with domain-specific language [22] or in combination with other information modalities such as image to improve similarity evaluation tasks [23]. To improve performance on tasks involving the comparison of different sentences and reduce computation resources required, sentence-oriented models such as Sentence-BERT [24] have been proposed, with the aim of providing better sentence embeddings by using siamese and triplet network architectures together with a pooling operation applied to the output of BERT or RoBERTa and the cosine similarity metric. Datasets such as STS benchmark [25] or SICK [26] are usually employed to train and evaluate these models.

2.3. Natural Language Inference

Natural Language Inference (NLI) is a task where the goal is to evaluate if a sentence called hypothesis (h) can be inferred given a sentence called premise (p) [27]. In other terms, given two sentences h and p, it is possible to infer if there is *entailment* between them, which means that h is definitely true based on the established knowledge of p. If the relation is a *contradiction*, h is definitely false based on p and if there is a *neutral* relation, although sentence h could be true, there is no actual basis on p [28]. In the three cases, the pair of sentences could involve high similarity, but detecting an entailment relation goes a step further, involving deeper natural language understanding models.

There are different datasets that have been designed to train and evaluate NLP models for NLI. The Stanford Natural Language Inference (SNLI) corpus [29] is composed of 570,000 pairs of sentences labelled with contradiction, neutral, or entailment labels by 5 human annotators. Multi-Genre Natural Language Inference (MultiNLI) [30] is another dataset designed to overcome several limitations of the SNLI dataset, where all sentences are extracted from image captions. Thus, MultiNLI is presented as a more complex corpus with a more varied language. Cross-lingual Natural Language Inference corpus (XNLI) [31] was built to serve as a cross-lingual corpus including sentence pairs from 15 different languages. Recurrent neural networks have proved to be able to achieve high performance in this domain, as it is the case of Long short-term memory networks (LSTMs) [32,33]. Various Transformer-based approaches have also been proposed, allowing interlingual sentences to be compared [34]. The importance of NLI is highlighted because of the importance of this task in the training of general-purpose language models, due to the high level of human language understanding to address this task.

In the case of automated fact-checking, NLI plays a very important role. Given a collection of false claims, the verification of a new information piece can be modelled as an NLI task, where our goal is to detect entailment with one of the false claims collected. Similarly, given a collection of true facts, we can model as a NLI task the process of determining if a new fact is true based on the existing facts in that collection.

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 evaluating at what level a claim can be considered true. These AFC methods typically integrate Machine Learning techniques, however, state-of-the-art research shows important limitations due to the training set used or the detection of paraphrasing [35]. Nevertheless, recent advances in this field, mainly because of the development of architectures using the attention-based mechanism, have led to important progress in the area.

Automated Fact-Checking is conducted using NLP models. There are different approaches to address this task according to the inputs [36]. One possibility is to derive the veracity of a claim without further knowledge or context [37], an approach highly unreliable. In this case, different datasets are usually employed to train machine learning-based tools for AFC to later classify news claims without considering recent knowledge [38]. For instance, FEVER is a dataset of claims extracted from Wikipedia and manually labelled as *Supported*, *Refuted* or *NotEnoughInfo* [39]. Hanselowski et al. [40] made public another dataset for automated fact-checking, with validated claims and documents annotated. WikiFactCheck-English [41] contains claims, context information, and evidence documents. Furthermore, Huertas et al. [42] presented a comparison of different transformer-based approaches to detect misinformation.

On the contrary, current research on automated fact-checking is moving towards the use of trusted information sources to retrieve evidence and make an accurate decision regarding the truth of the claim and provide fact-checkers with new technologies and instruments [43]. Zeng et al. [44] offer a broad description of the state-of-the-art following this research line. The authors consider two main categories to group the tasks around the fact-checking process: claim detection and claim validation. While the first task is a filtering process, aiming to detect those claims requiring verification, the second task involves deciding if the claim is true or not. In contrast to the two-step process proposed by Zeng et al. other authors [45] have considered that there should be a third intermediate step for evidence retrieval, a task that could be carried out computationally using a stance detection model [46].

Regarding the validation process, existing approaches can be categorised according to the truth source considered. While a few approaches rely on existing information sources such as Wikipedia, others rely on structured knowledge bases. However, both present limitations. While the former considers sources such as Wikipedia as trustworthy despite the possibilities of finding manipulated information, the latter expects that the required information is available, which is not always possible.

An example of a proposed approach relying on Wikipedia is WikiCheck [4]. The authors find that current approaches in the state-of-the-art entail overfitting issues, proposing a data filtering method to improve generalisation. The authors describe a new AFC mechanism based on a double process: Wikipedia search through the use of a NER algorithm and a natural language inference module, that implements a siamese network architecture to compare between a claim and multiple hypothesis sentences. The evaluation is mostly performed in the FEVER dataset. Also focused on Wikipedia, WhatTheWikiFact [47] was proposed in 2021 aiming to provide a confidence score for a claim and relevant articles from Wikipedia to provide evidence. It implements a Document Retrieval module to obtain relevant articles, a Sentence Retrieval module to target relevant sentences and a natural language inference module to infer if any of the sentences support or refute the claim.

Alonso-Reina et al. [5] also proposed a step-down process to retrieve documents and sentences, claiming that the sentence retrieval task is a complex part of the process and further research is required to improve the quality of this stage of the process. FAKTA [48] is an architecture for automatic fact-checking that implements evidence retrieval in addition to stance detection and linguistic analysis. Researchers have also explored the combination of multiple sources of information [49] or to leverage relevant knowledge to reach more reliable decisions. In contrast to these previous approaches, FacTer-Check is mostly focused on the validation process, aimed at providing a tool to verify any desired input claim on demand.

In general, most researchers have considered machine learning an essential ingredient to implement automated fact-checking tools and have proposed different implementations according to the features selected or the objectives. Naderi and Hirst [3] use linguistic features and a classifier in a statement multi-class classification task. Karadzhov et al. propose the use of LSTM networks to classify claims in combination with relevant fragments of text from external sources [50]. Kotonya et al. [2] provide a broad analysis of the state-of-the-art literature of automated fact checking approaches that focus on explainability. Other important implementation is ClaimBuster [51], 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. Stammbach and Ash [52] use GPT-3 to generate a summary of evidences for a fact check decision. The attention-based mechanism has also been used to identify worthy statements [53]. BERT has also been used to predict veracity and generate explanations in the public health domain [2].

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 presure* or *social influence* [54] hinder the process of analysing the propagation of a false claim. Monti et al. [55] 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 [56], 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 proposed the use of transformers to profile hate speech on Twitter [57].

The fight against misinformation in Online Social Networks has also been explored from the author's perspective, modelling user profiles and their characteristics according to the probability of trusting or distrusting false claims [58,59].

3. Fighting misinformation through semantic similarity and Natural Language Inference

FacTeR-Check aims at helping during 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 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

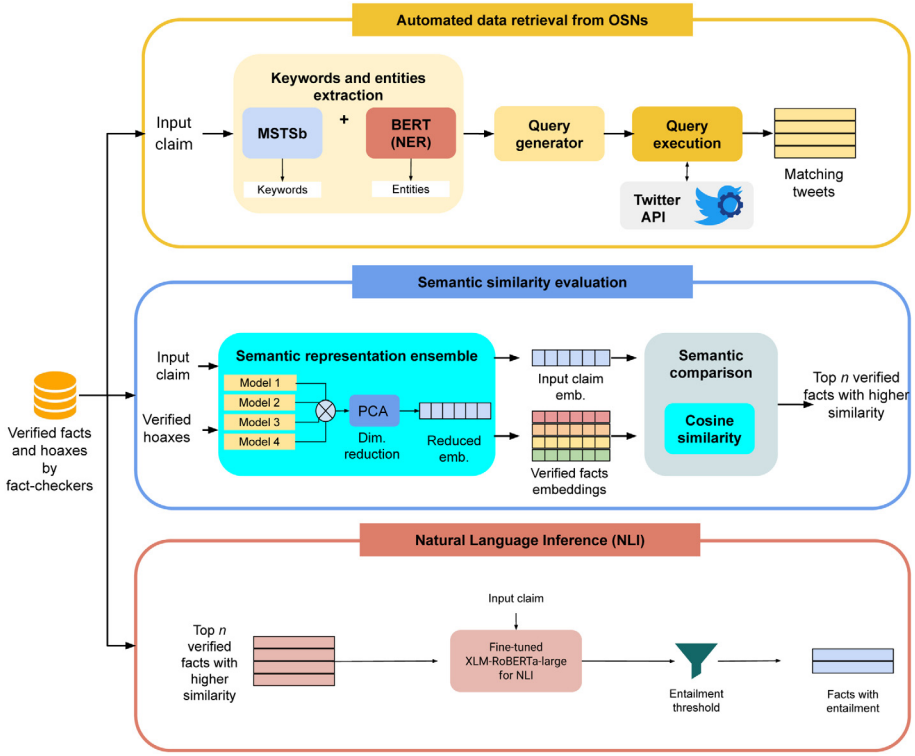

Fig. 2. Diagram showing the three components of the FacTeR-Check architecture.

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 entire 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 three different main functionalities (see Fig. 2):

1. **Multilingual Semantic similarity evaluation:** This component allows to measure the semantic distance between pairs of sentences. Thus, once received a new claim, 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. This module has two main goals: to identify semantically related hoaxes which can be used to inform the user of the tool and filter the fact-checked claims which are later evaluated using the NLI module.
2. **Multilingual Natural Language Inference:** In order to calculate if the input claim is aligned with a fact-checked claim, an NLI module calculates the probability of entailment with the input claim. If a coincidence is found (an entailment probability exceeds a certain threshold), the input claim is considered 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 search query composed of a series of relevant keywords and named entities 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.

The three functionalities described enable two different workflows, as already shown in Fig. 1. One is intended to provide a useful mechanism for semi-automated fact verification, checking claims against a database of fact-checked claims. 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 filtering step implemented through semantic similarity evaluation allows to offer fact-checked claims that could be relevant for the use in the case that no entailment is found with the NLI module. Besides, it also allows to improve the efficiency, reducing the time required by the NLI module to compare the input sentence with a series of candidate fact-checked claims (see Fig. 3).

The second workflow is designed to aid fact-checkers in the process of monitoring and tracking the life of a false claim in an Online Social Network. This involves retrieving a large number of tweets or posts that support a given false claim. For that purpose, it is required to build a search query composed of relevant keywords and named entities extracted from the hoax. The generated string is then sent to the API of the OSN in order to retrieve content related to the input claim. Once all the tweets or posts have been retrieved, the semantic similarity and NLI modules allow then to filter all the data to keep tweets or claims actually supporting the false claim. The next subsections describe in detail each functionality.

FacTeR-Check also implements multilingualism in all components. This is essential to be able to compare or verify information regardless of the language of the input sentence and the fact-checked claims.

3.1. The modules of FacTeR-check

This subsection describes in detail each module composing the FacTeR-Check architecture.

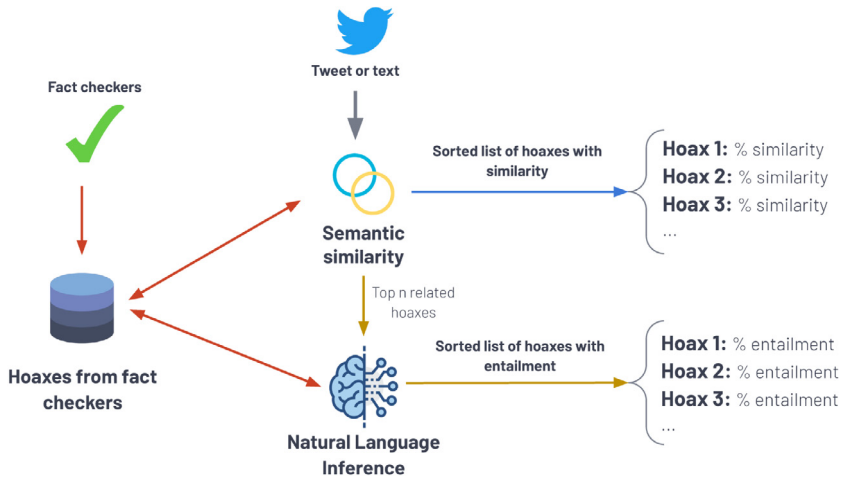

Fig. 3. Architecture for the evaluation of information pieces against hoaxes already identified by fact checkers. A first step allows to retrieve hoaxes that are semantically similar to the input text. In the second step, a natural language inference model measures the degree of entailment against each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

3.1.1. Semantic similarity module

Semantic is the level of language that deals with the meaning of a sentence by focusing on word-level interactions. 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 Transformer-based models, the goal is to evaluate the degree of similarity between a new input sentence and a database of fact-checked claims. The result will be a subset of fact-checked claims ensuring a certain minimum degree of similarity. This module is composed of an ensemble of models and a Principal Component Analysis to improve efficiency.

To measure the semantic similarity between texts, we use the cosine similarity function. This metric takes advantage 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 Eq. (1):

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

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 evaluated in the state-of-the-art literature [60–64], showing that a mixture of embeddings featuring different characteristics leads to more robust representations and better performance than single embedding-based methods. In addition, another advantage of ensemble methods is the expansion of vocabulary coverage.

To work in a multilingual scenario, the multilingual extended version of the Semantic Textual Similarity Benchmark (mSTSb)[2] [65] is used to fine-tune four well-known multilingual base models concatenated in the proposed ensemble (see Fig. 4). These single models are available on Sentence Transformers[3] [24].

mSTSb [65] departs from the original STS Benchmark [66], which consists of train-dev-test splits of 5749, 1500, and 1379 pairs of sentences, respectively, labelled with a similarity score between 0 and 5, from less to more similar. mSTSb expands these pairs to a multilingual level, translating from English to 15 languages[4] using the Google Translator python package,[5] following the same procedure that has been extensively used in the existing related literature [67,68]. Furthermore, the authors used this procedure because it allowed them to maintain the quality of the data, as the translated output received a confidence value. Thus, the mSTSb multilingual extension process dropped those sentence pairs with a confidence value below 0.7.

The multilingual SentenceTransformers models composing the semantic similarity ensemble are:

- **paraphrase-xlm-r-multilingual-v1**: Distilled version of RoBERTa [12] trained on large-scale paraphrase data using XLM-R [69] as the student model.
- **stsb-xlm-r-multilingual**: Distilled BERT [11] version trained in NLI [30] and STSb [66] using XLM-R as the student model.
- **paraphrase-multilingual-MiniLM-L12-v2**: Multilingual version of the MiniLM model from Microsoft [70] trained on large-scale paraphrase data.
- **paraphrase-multilingual-mpnet-base-v2**: Distilled version of the MPNet model from Microsoft [71] 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 [24]. To obtain the best results and avoid overfitting, we optimised the following hyperparameters using a 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 repository.[6]

As explained by Sidorov et al. [72], 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 problem, **Principal Component Analysis (PCA)** is computed and applied to

2 https://github.com/Huertas97/Multilingual-STSB.

3 https://www.sbert.net/.

4 ar, cs, de, en, es, fr, hi, it, ja, nl, pl, pt, ru, tr, zh-CN, zh-TW.

5 Google Translator python package: https://pypi.org/project/google-trans-new/.

6 Fine-tuned models available in https://huggingface.co/AIDA-UPM.

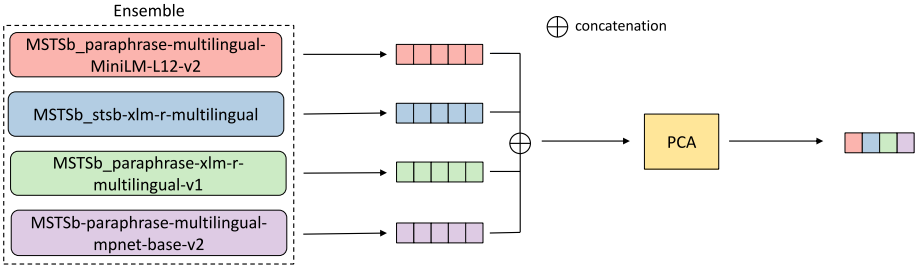

Fig. 4. Ensemble and dimensionality reduction approach proposed. Concatenation of embeddings from four multilingual sentence-transformers models applying PCA dimensionality reduction.

the entire architecture as shown in Fig. 4. This enables reducing 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 a new fact-checked claim is received, its embedding representation is obtained by applying the ensemble and PCA models to the concatenated outputs and saving them in the fact-checked claim database. This will allow to easily evaluate new claims, calculating the cosine distance to each fact-checked claim stored.

3.1.2. 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 cannot detect these finer nuances, an NLI model is able to detect a entailment or contradiction relationship given a pair of sentences. If we manage to detect if a 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 beyond this research.

To better describe the NLI task, let $\langle p, h\rangle$ be a sentence pair of a detected hoax and a statement to be verified. Using language inference we can infer *contradiction* and *neutral* probabilities between both, however, our main focus is on finding the degree of *entailment*. We find two possible outcomes when h is examined, either h is a hoax or h has undetermined factuality. These outcomes have a certain probability of happening that we want to approximate, which from now on we call events. The event of h actually being a hoax is called h_f, while the event for h having undetermined factuality is called h_u. Formally, we require to find a function f that is able to approximate Eq. (2).

$$f(p, h) \approx P(p|h_f) \tag{2}$$

where p is a hoax or fact-checked claim verified by fact-checkers and we have the certainty that it involves fake information, h is the statement to be verified 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 [69] is chosen. Transformer models for NLI have problems when transferring to unseen domains [73], so special consideration is given to the fine-tuning process. To train this network various datasets are used, Machine Translated MultiNLI (MNLI-MT) [30,31] and XNLI for validation and evaluation [31]. Auxiliary English data have also been used from other sources, including ANLI [74], SNLI [29] and FEVER [39] (adapted to NLI). We train two versions of the model, one with cross-lingual data (MNLI only) and other with all cited data. To ensure language understanding from any language to another, we fine-tuned our model with shuffled language pairings sampled from all possible combinations of 15 languages present in MNLI-MT.

Hyperparameters for fine tuning are a batch size of 1024 and a learning rate of $2e-5$ with Adam [75] optimiser. A learning rate schedule is used with warmup (6% of training steps) and linear decay. The final network is decided according to the validation set provided by XNLI.

3.1.3. Data retrieval from OSNs

The massive volume of information present on social media platforms makes it unmanageable to manually track and monitor hoaxes' evolution. 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 data downloaded 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 OSNs 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 obtain a more comprehensive picture.

As illustrated in Fig. 5, our proposal used for automatic keyword extraction, named **FactTeR-CheckKey**, combines KeyBERT [76] and multilingual Name Entity Recognition (NER) approaches with Spacy [77] and Flair [78] frameworks as keyword filtering steps.

KeyBERT is a Python package for keyword extraction that selects as keywords those words that are the most semantically similar to an input text. For this purpose, it leverages the semantic power of Transformer-based models to compute text embedding and word embeddings. Then it uses the cosine similarity distance metric to find the words most semantically similar to the text. It should be noted that the KeyBERT package provides the infrastructure to do this, but not the Transformer model employed, which the user must provide. Hence, to maintain consistency, FactTeR-CheckKey uses our previously mentioned multilingual `MSTSb-paraphrase-multilingual-mpnet-base-v2` model fine-tuned in mSTSb as the semantically aware model.

The multilingual NER module consists of a fine-tuned XLM-RoBERTa base model[7] available at Hugging Face [79], trained over the 40 languages proposed in the XTREME [80] dataset.

[7] https://huggingface.co/jplu/tf-xlm-r-ner-40-lang.

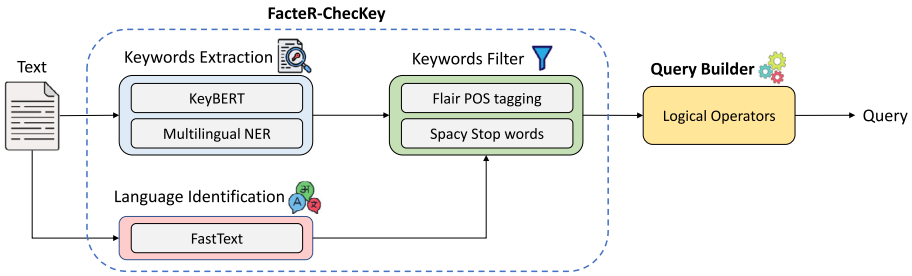

Fig. 5. Architecture of the FacteR-CheckKey module used for query building. Keywords are extracted using our own multilingual model in the KeyBERT package in combination with the multilingual NER model based on XLM-RoBERTa. Then, Flair multilingual tagger and Spacy models are used to delete verbs, conjunctions, adverbs, adpositions, and stop words from the extracted keywords. Finally, queries are built by concatenating the different keywords extracted with the AND logical operator.

- Example 1 -

Spanish Hoax → La prueba de antígenos no sirve para la COVID-19 porque da positivo con Coca-Cola
Keywords → prueba, antígeno, covid-19, positivo, coca-cola
Query → (prueba AND antígeno AND covid-19 AND positivo AND coca-cola)

English Hoax → Antigen tests are useless for COVID-19 because they test positive with CocaCola
Keywords → antigen, test, covid-19, cocacola
Query → (antigen AND test AND covid-19 AND cocacola)

- Example 2 -

Spanish Hoax → En Israel no mueren por coronavirus gracias a una receta de limón y bicarbonato
Keywords → coronavirus, receta, limón, bicarbonato
Query → (coronavirus AND receta AND limón AND bicarbonato)

English Hoax → No deaths in Israel due to coronavirus thanks to a recipe with lemon and bicarbonate
Keywords → israel, coronavirus, recipe, lemon, bicarbonate
Query → (israel AND coronavirus AND recipe AND lemon AND bicarbonate)

Fig. 6. Examples of queries generated with FactTeR-CheckKey from English and Spanish hoaxes for searching through the Twitter API.

To optimise multilingual keyword extraction for query building purposes, stop words are removed using Spacy *v3* [77]. Furthermore, multilingual part-of-speech tagging, POS tagging, is accomplished with Flair,[8] removing verbs, auxiliary verbs (AUX), coordinating conjunctions (CCONJ) and subordinating conjunctions (SCONJ), adverbs (ADV), and adpositions (ADP).

In contrast to the straightforward KeyBERT approach, FacteR-CheckKey automatically detects the language introduced to apply the appropriate stop-word list. To this end, the FastText `lid.176.bin` model[9] [81,82] is used as the language identification system. The criteria for selecting this language identification system were that compared to other systems (e.g., CLD2, langid, langdetect), it covers the most significant number of languages, totalising 179 languages trained both on long and short texts, it is fast and accurate.[10] It also returns a language identification probability value used in FacteR-CheckKey as a quality threshold. Thus, text inputs with a language identification probability below 0.7 are not considered and instead use English as the default language for stop words.

Fig. 6 provides an overview of the data from the hoaxes and the queries built to search the Twitter API. Queries are built by concatenating the different keywords extracted with the logical operator "AND".

3.2. Semi-automated (S-AFC) fact-checking through Natural Language Inference and semantic similarity

In this work, we propose a two-step process to perform semi-automated fact-checking (S-AFC). The semantic similarity and natural language inference modules described in the previous two sections Sections 3.1.1 and 3.1.2 are the pillars of this S-AFC process. The first step allows filtering 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 is obtained according to the degree of similarity, 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 be aligned with 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.

4. Evaluation of the FacTeR-check architecture

This section presents an evaluation of FacTeR-Check. Given it is not possible to evaluate the joint operation of all the modules due to there is not dataset available so far for this purpose, the Semantic Similarity, Natural Language Inference, and Keyword Extraction Modules (FactTeR-CheckKey) are evaluated individually using benchmark datasets from the state-of-the-art literature. The following subsections describe in detail the results obtained for each task.

4.1. Semantic similarity evaluation

The multilingual STS Benchmark [65], an extended version of the well-known STSb dataset, 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, and 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-`

[8] https://huggingface.co/flair/upos-multi.
[9] https://fasttext.cc/docs/en/language-identification.html.
[10] https://github.com/modelpredict/language-identification-survey.

Table 1
Spearman ρ and Pearson r correlation coefficient between the sentence representation from multilingual models and the gold labels for STS Benchmark test set.

Model	Dim	EN-EN		EN-ES		ES-ES		Avg	
		r	ρ	r	ρ	r	ρ	r	ρ
MSTSb_paraphrase-mltl-MiniLM-L12-v2	348	85.26	86.17	81.45	81.49	83.30	83.68	81.38	81.47
MSTSb_stsb-xlm-r-mltl	768	84.21	85.10	82.65	83.04	83.20	83.83	81.75	82.09
MSTSb_paraphrase-xlm-r-mltl-v1	768	84.80	85.59	82.90	83.19	83.41	83.71	82.39	82.60
MSTSb-paraphrase-mltl-mpnet-base-v2	768	86.80	**87.40**	84.42	84.45	85.19	**85.52**	83.48	83.59
Ensemble 2	1152	85.90	86.72	83.68	83.87	84.39	84.67	83.25	83.41
Ensemble 3	1920	86.34	87.13	84.18	84.34	84.86	85.14	83.67	**83.84**
Ensemble 4	2688	85.73	86.59	84.16	**84.53**	84.67	85.25	83.33	83.62

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.

Model + PCA	Dim	EN-EN		EN-ES		ES-ES		Avg	
		r	ρ	r	ρ	r	ρ	r	ρ
MSTSb_paraphrase-mltl-MiniLM-L12-v2	184	84.92	85.71	81.04	81.04	83.08	83.28	81.03	81.02
MSTSb_stsb-xlm-r-mltl	408	84.35	85.11	82.84	83.17	83.39	83.89	81.85	82.08
MSTSb_paraphrase-xlm-r-mltl-v1	286	84.79	85.50	82.73	82.97	83.38	83.58	82.23	82.39
MSTSb-paraphrase-mltl-mpnet-base-v2	306	86.69	87.27	84.21	84.28	84.93	85.19	83.20	83.28
Ensemble 2	347	85.91	86.72	83.49	83.69	84.42	84.68	83.12	83.28
Ensemble 3	367	86.64	87.55	84.50	84.80	85.24	85.72	83.85	84.21
Ensemble 4	429	86.77	**87.78**	85.00	**85.52**	85.56	**86.20**	84.24	**84.71**

Table 3
Spearman ρ correlation coefficient for the Ensemble 4 with 429 PCA reduced dimension for the STS Benchmark monolingual and cross-lingual tasks test sets.

Monolingual		Cross-lingual	
Language pair	ρ	Language pair	ρ
AR-AR	84.05	EN-AR	82.55
CS-CS	85.52	EN-CS	85.17
DE-DE	85.93	EN-DE	85.50
ES-ES	86.24	EN-ES	85.58
EN-EN	87.85	EN-FR	85.26
FR-FR	85.29	EN-HI	81.25
HI-HI	82–82	EN-IT	85.31
IT-IT	85.77	EN-JA	82.67
JA-JA	85.30	EN-NL	85.80
NL-NL	85.95	EN-PL	84.17
PL-PL	85.43	EN-PT	85.75
PT-PT	86.29	EN-RU	84.16
RU-RU	84.99	EN-TR	83.92
ZH-ZH	83.38	EN-ZH	83.08

`MiniLM-L12-v2`, Ensemble 3 adds the `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 (see Table 3).

As expected, the use of ensemble-based approaches dramatically increases the number of dimensions. To address 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. However, 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[11] extracted from three well-known resources (TED2020,[12] WikiMatrix [83] and OPUS-NewsCommentary [84]) are used to fit the PCA for each model. The relationship between the performance obtained and the reduction size is shown in Fig. 7. As can be seen, both in the case of single fine-tuned models and ensemble architectures, the performance converges with fewer than 200 principal components, which provides a substantial space reduction. The best PCA space is selected according to the average performance of the mSTSb development set across languages.

Table 2 shows the results after combining PCA and the ensemble approach, evidencing that this approach leads to better performance, dramatically reducing 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 also requires fewer dimensions than most of the single models. This demonstrates that ensemble approaches in combination with dimensionality reduction techniques allow one 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 the automated fact-checking process, that is to say, 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 XNLI. This dataset is the usual benchmark for this task, being commonly employed in similar works involving NLI tasks. Our motivation is to show that NLI results are aligned with the state-of-the-art, so our baselines are taken from popular and similar works. We showcase our results in Table 4. These results are obtained from the evaluation of paired language premises and

[11] The languages used in this scenario are: ar, cs, de, en, es, fr, hi, it, ja, nl, pl, pt, ru, tr, zh.

[12] https://www.ted.com/participate/translate.

Table 4
Evaluation for the XNLI test set, reporting accuracy percentages. Cited results reported from their original sources.

Model	Avg	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh
BiLSTM [85]	70.2	71.4	74.2	72.6	72.8	73.9	72.9	71.9	65.5	72.1	62.2	69.2	69.7	61.0	72.0	71.4
Unicoder [86]	78.5	78.2	81.4	80.9	79.5	85.6	82.3	81.1	73.4	79.7	73.8	77.1	76.8	69.6	77.9	80.5
XLM [13]	76.9	75.6	79.4	79.3	78.6	84.5	81.3	80.1	72.1	77.5	69.2	75.7	75.2	67.7	78.3	78.3
XLM-R + MNLI-MT	83.1	82.5	**86.3**	**85.1**	85.1	88.8	**86.2**	**84.7**	81.3	**83.1**	78.2	**81.4**	82.4	75.9	83.1	83.2
XLM-R + MNLI-MT + English	**83.2**	**83.0**	85.5	85.0	**85.6**	**88.9**	**86.2**	84.4	**81.4**	82.9	**78.3**	80.9	**82.6**	**76.1**	**83.4**	**83.4**

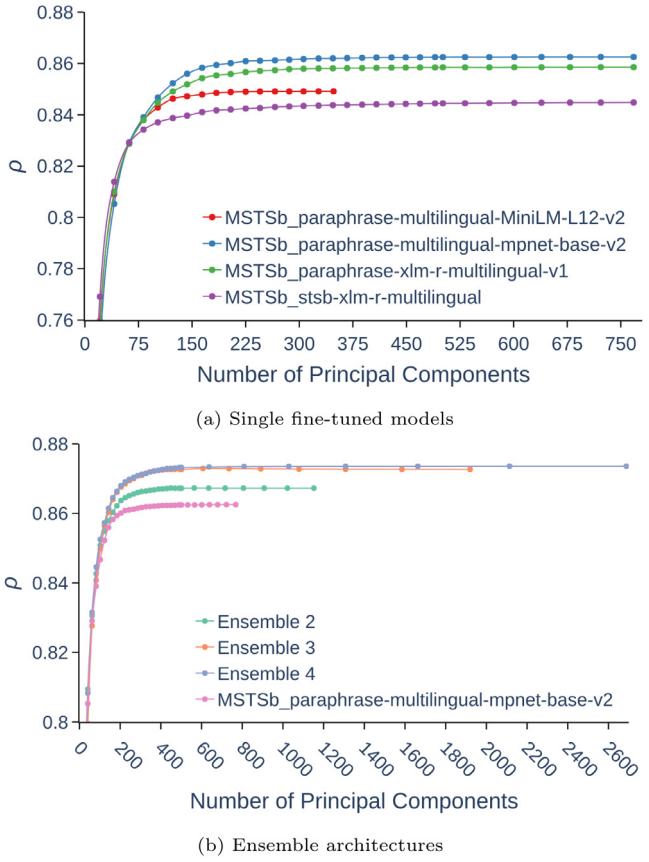

Fig. 7. Number of components selection in mSTSb development set. Average Spearman Correlation Coefficient of the single fine-tuned models (a), and ensemble architectures (b) using cosine similarity for the 15 languages as a function of the number of components from the extended STS-Benchmark development set. The average of correlation coefficients is computed by transforming each correlation coefficient into a Fisher z-value, averaging them and then backtransforming them to a correlation coefficient.

hypothesis. At first glance, XLM-R is superior to the baselines and is in line with the proposal in the original work [69] despite our differences in the fine-tuning procedure.

To better appreciate the advantages of adding more data to the fine-tuning, we extract results from all languages crossed with each other in Tables 5 and 6 for the MNLI-MT model and the model with additional English data. First, when evaluating any language to another, results are worse overall, dropping from 83% to 81% and 82%. While not a large difference, it means that language-crossing has some impact in the end results, which are not evaluated in the original XLM-RoBERTa article. We find the inclusion of additional data from challenging datasets such as ANLI or FEVER beneficial to the cross-lingual generalisation for two reasons: average accuracy is 1.3% higher and lower resource languages (swahili or urdu) have significantly higher accuracy when acting as either premise or hypothesis. It is important to highlight the high accuracy, usually larger than 80% for majority languages such as Chinese, Spanish or English, that is attained by the module when mixing languages, allowing for international detection of misinformation.

4.3. Performance of the keywords extraction module

For the evaluation of the FactTeR-CheckKey module, our approach is evaluated in 5 annotated datasets of Automatic Keyphrase Extraction from different languages:

- **110-PT-BN-KP** [87]: TV Broadcast News (BN) dataset that contains 110 transcription text documents from 8 broadcast news programmes from the European Portuguese ALERT BN database ranging from politics, sports, finance and other broadcast news. It uses an average of 23.73 gold keywords per document.
- **SemEval2017** [88]: 500 English paragraphs selected from ScienceDirect journal articles, among the domains of Computer Science, Material Sciences, and Physics. It uses an average of 18.19 gold keywords per document.
- **WikiNews** [89]: French version of WikiNews, which contains 100 news articles. It uses an average of 18.19 gold keywords per document. It uses an average of 11.77 gold keywords per document.
- **WICC** [90]: Composed of 1640 Spanish scientific articles published in the Workshop of Researchers in Computer Science (WICC). It uses an average of 4.57 gold keywords per document.
- **Pak2018** [91]: Dataset in Polish formed by 50 abstracts from journals on technical topics collected from Measurement Automation and Monitoring.[13] It uses an average of 4.64 gold keywords per document.

Additionally, we have included an evaluation on the 60 Spanish hoaxes presented in the following section. We compared the automatically extracted keywords with a series of keywords extracted manually by the authors, with an average of 4.5 gold keywords per tweet.

We have considered three baseline methods for comparison purposes: the multilingual statistical Rapid Automatic Keyword Extraction (RAKE) algorithm [92], Yet Another Keyword Extractor (YAKE) [91] and the straightforward KeyBERT using the `paraphrase-xlm-r-multilingual-v1` model. RAKE is a well-known unsupervised statistical method for keyword extraction based on the collocation and co-occurrence of words by eliminating stopwords and punctuation. Similarly, YAKE is also an unsupervised automatic keyword extraction method that includes more text features to statistically select the most relevant

[13] http://pak.info.pl/index.php?menu=menu&idMenu=924.

Table 5
Evaluation for the XNLI test set with unaligned premises and hypothesis, reporting accuracy percentages using the XLM-RoBERTa model with shuffled MNLI-MT data. Column is the language of the hypothesis and row is the language of the premise.

	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	Avg
ar	*82.6*	82.9	82.7	82.9	85.3	83.1	83.4	80.6	81.1	74.3	79.9	80.4	76.4	81.9	80.8	81.3
bg	82.1	*86.3*	84.9	84.8	87.0	85.3	84.8	82.0	83.7	75.1	82.1	81.8	77.9	83.2	82.9	
de	81.9	84.6	*85.1*	83.9	86.4	84.4	84.2	81.7	82.7	74.1	81.3	81.1	77.7	82.5	82.0	
el	81.6	84.0	84.2	*85.0*	86.2	84.5	84.4	81.6	82.2	74.1	80.9	81.1	77.4	82.2	82.0	
en	83.9	86.5	86.6	86.3	*88.8*	86.6	86.0	84.6	85.1	79.0	83.8	84.1	81.1	85.0	84.6	
es	81.9	84.8	83.5	84.1	87.0	*86.2*	84.6	81.7	83.0	73.2	81.4	81.6	77.9	82.6	82.3	
fr	81.3	84.5	83.7	83.7	86.1	84.3	*84.7*	81.7	82.7	73.6	81.0	80.9	77.7	82.4	82.5	
hi	79.6	82.2	81.7	81.5	84.3	82.4	82.1	*81.3*	80.7	72.8	79.7	78.8	77.3	80.1	80.5	
ru	81.2	83.5	83.8	82.8	85.3	83.7	83.5	80.6	*83.1*	74.8	80.3	81.1	77.9	81.4	81.7	
sw	77.9	80.3	79.9	80.0	83.3	80.3	80.3	77.1	78.6	*78.2*	78.2	76.7	73.7	78.5	78.9	
th	80.4	83.4	82.8	83.2	84.9	83.1	83.1	80.5	81.6	74.4	*81.4*	80.5	77.3	81.5	81.8	
tr	79.8	82.4	82.4	82.0	84.4	82.4	82.5	80.0	81.3	69.7	79.7	*82.4*	76.0	80.4	80.8	
ur	75.9	77.5	77.3	78.0	82.1	78.3	78.5	78.7	76.2	69.1	74.3	74.9	*75.8*	77.4	76.6	
vi	80.9	83.4	83.3	82.9	85.6	83.3	83.5	80.7	81.9	73.4	80.0	80.6	77.3	*83.1*	81.4	
zh	80.2	83.0	83.2	82.2	85.8	83.3	82.9	80.0	81.9	75.0	80.0	80.4	77.4	80.8	*83.2*	

Table 6
Evaluation for the XNLI test set with unaligned premises and hypothesis, reporting accuracy percentages using the XLM-RoBERTa model with shuffled MNLI-MT, ANLI, SNLI and FEVER. Column is the language of the hypothesis and row is the language of the premise.

	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	Avg
ar	*83.0*	84.4	84.5	84.1	85.3	84.5	84.1	82.2	82.5	78.4	81.1	82.6	78.5	83.2	83.0	82.6
bg	84.0	*85.5*	84.9	85.2	87.2	86.1	85.2	82.8	84.5	79.5	82.7	83.3	79.2	84.3	84.5	
de	83.1	85.3	*85.0*	84.7	86.6	85.3	85.0	82.7	83.3	79.3	82.1	82.8	78.7	83.3	83.8	
el	83.4	84.9	84.7	*85.6*	86.8	85.2	84.7	82.4	83.4	79.2	82.3	83.1	79.1	83.8	83.8	
en	85.0	87.4	86.8	86.7	*88.9*	87.8	86.8	85.1	85.5	80.7	84.4	85.1	81.4	85.8	85.9	
es	83.3	85.7	85.1	85.2	86.4	*86.2*	84.7	83.5	83.7	79.1	82.1	83.6	79.7	84.2	84.3	
fr	82.8	84.9	84.6	84.7	85.8	85.0	*84.4*	82.4	83.6	79.1	82.3	82.6	78.5	83.7	83.7	
hi	81.3	82.9	82.5	82.5	84.8	83.5	82.8	*81.4*	81.7	76.6	80.0	81.4	78.5	81.9	81.4	
ru	82.3	84.1	84.2	84.1	85.5	84.2	83.6	81.7	*82.9*	78.0	81.8	82.3	78.9	82.7	82.6	
sw	80.2	82.1	81.9	82.1	83.5	82.2	81.0	80.5	80.8	*78.3*	79.5	79.8	75.4	80.6	80.8	
th	81.2	83.7	83.0	83.6	84.9	84.0	83.1	81.6	82.1	77.5	*80.9*	81.4	77.7	82.6	82.6	
tr	82.2	83.8	83.2	83.4	85.4	83.9	83.8	81.6	82.0	77.1	80.5	*82.6*	77.9	82.3	82.4	
ur	78.6	80.6	79.9	80.6	82.5	80.9	79.8	79.2	78.9	75.2	77.6	78.7	*76.1*	79.5	79.9	
vi	83.2	84.3	83.8	83.8	85.4	84.7	83.8	81.8	82.7	78.4	81.6	82.1	78.4	*83.4*	83.3	
zh	81.9	83.8	83.9	83.9	85.5	84.0	83.9	81.9	82.3	77.6	81.3	81.8	78.7	82.7	*83.4*	

Table 7
Evaluation of Keyword-Extraction Approaches. P, R and F1 stand for Precision, Recall, and F1-score metrics, respectively.

	110-PT-BN-KP			SemEval 2017			WikiNews		
	P	R	F1	P	R	F1	P	R	F1
RAKE	0.3332	0.1266	0.1703	0.3509	0.1885	0.2370	0.2800	0.2609	0.2658
YAKE	0.4291	0.4108	0.3874	0.6398	0.3761	0.4565	0.1291	0.1526	0.1250
KeyBERT	0.4956	0.4615	0.4454	0.6154	0.3592	0.4378	0.3217	0.2958	0.3032
FacteR-ChecKey	0.6050	0.5787	**0.5670**	0.6662	0.3710	**0.4594**	0.3184	0.4795	**0.3724**
	WICC			Pak2018			Our Hoaxes		
	P	R	F1	P	R	F1	P	R	F1
RAKE	0.2198	0.1949	**0.1980**	0.0403	0.0183	0.0245	0.3608	0.4080	0.3693
YAKE	0.2224	0.1857	0.1921	0.0403	0.0183	0.0245	0.3603	0.4887	0.4058
KeyBERT	0.1225	0.1098	0.1102	0.1000	0.0645	0.0774	0.5383	0.7141	0.5990
FacteR-ChecKey	0.1060	0.2035	0.1280	0.1297	0.0984	**0.1087**	0.6489	0.7980	**0.7003**

keywords in a text. It does not need to be trained on a particular set of documents, nor does it depend on dictionaries, external corpora, text size, language, or domain. In contrast to KeyBERT and FacteR-ChecKey, the RAKE and YAKE algorithms do not take into account any semantic information for the extraction process.

Finally, precision, recall, and the F1 score are the metrics used to evaluate the ability to extract unigram keywords compared to gold keywords for each dataset considered.

As can be seen in Table 7, FacteR-ChecKey clearly has an advantage over RAKE, YAKE and KeyBERT approaches in 110-PT-BN-KP, WikiNews and PaK2018. It slightly improves performance in the SemEval 2017 dataset. Although FacteR-ChecKey did not show the best performance in the WICC dataset and did not overcome statistical methods, it is still improving the straightforward KeyBERT approach. These results indicate that FacteR-ChecKey is a powerful approach to extracting keywords in a multilingual scenario to build queries to search through the Twitter API.

Finally, one note about the query generation process using 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 based on the similarity score. For this reason, our method has many practical applications. From already checked hoaxes, it

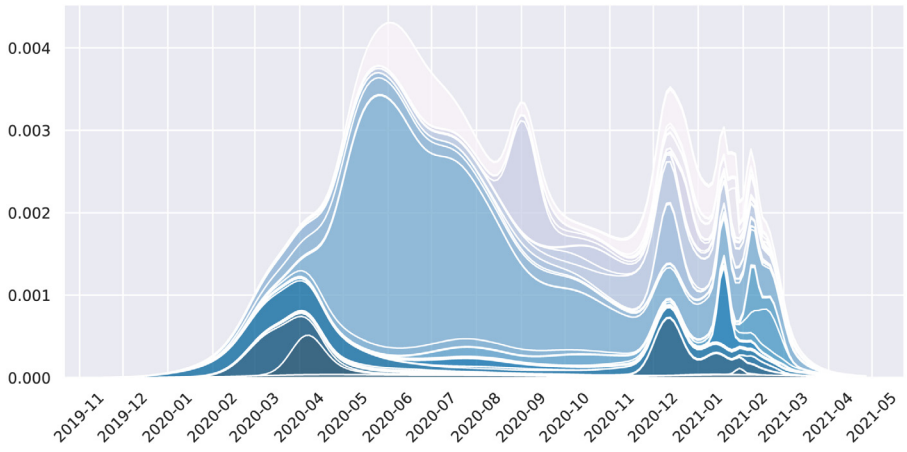

Fig. 8. Temporal distribution of tweets supporting the 61 hoaxes identified, evidencing common trends with multiple shared peaks.

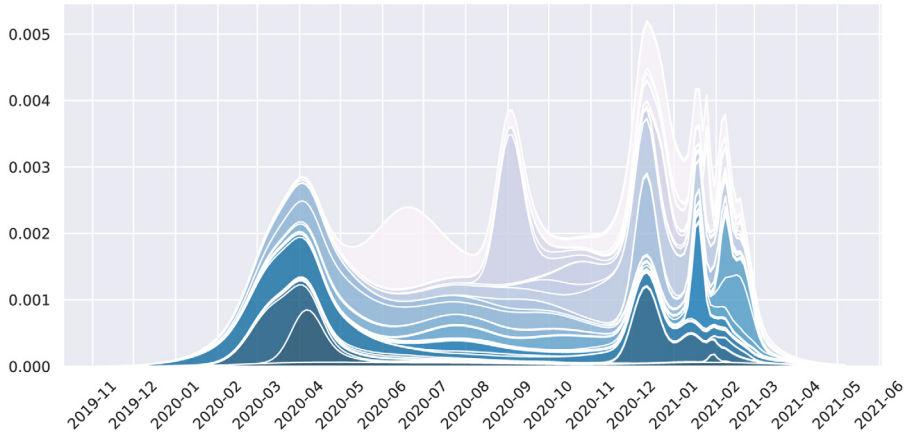

Fig. 9. Temporal distribution of tweets supporting the hoaxes identified without representing the hoax with id 31, related to the false claim "masks cause hypoxia".

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 is to show practica example of the use of FacTeR-Check, building a dataset of tweets spreading misinformation claims detected and verified by fact-checkers.[14] 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 related to the hoaxes from the Twitter API
3. **Tweets retrieving:** By using FacTeR-CheckKey, 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 applied 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 entail 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,[15] 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 registered false post. 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-CheckKey module. These search queries are then used through the Twitter API to find posts that share that type of disinformation. Each resulting search query is composed of representative keywords linked by search operators and the use of parentheses to improve the results. In some specific cases, we manually included more keywords to

[14] The dataset is available at: https://aida.etsisi.upm.es/datasets/.

[15] https://chequeado.com/latamcoronavirus/.

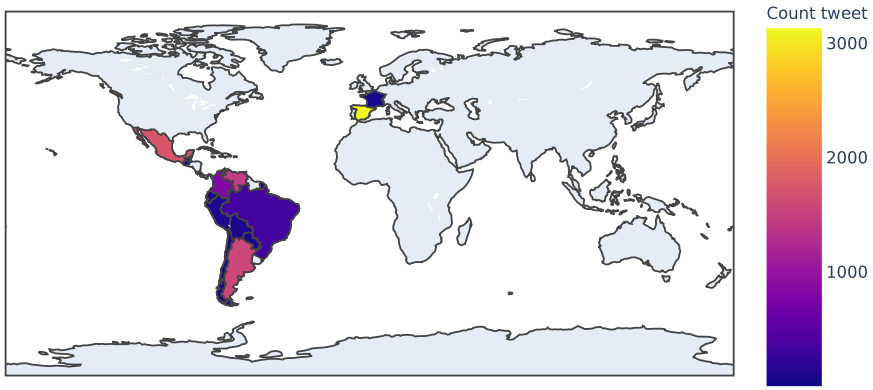

Fig. 10. Map showing the number of tweets supporting a hoax according to the nationality of the fact-checker that has identified it. In the case of France, although it is not a Spanish speaking country, several hoaxes have been identified by Factual AFP fact-checker, a France agency.

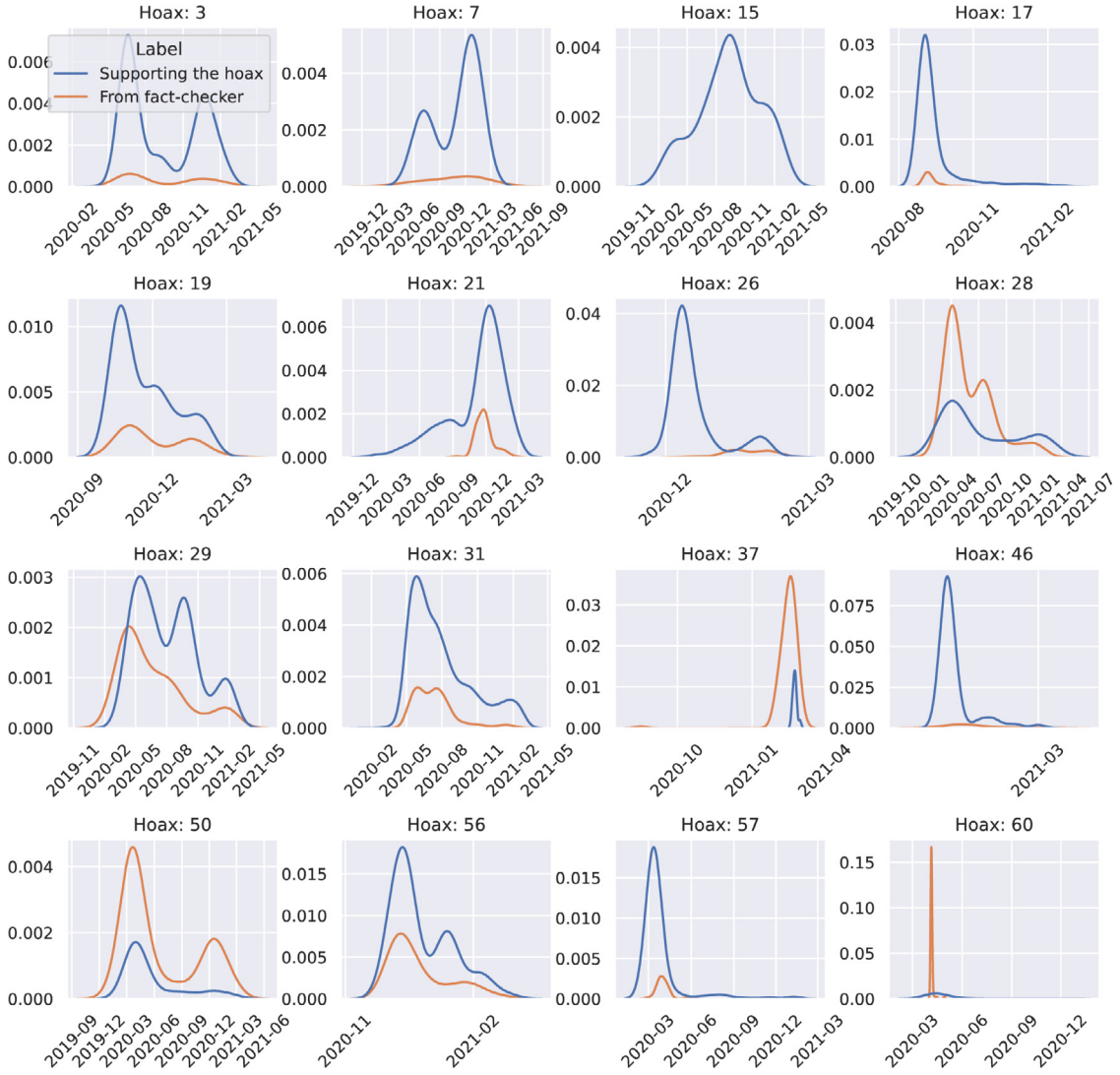

Fig. 11. Comparative for different hoaxes between the distribution of tweets supporting a specific hoax and tweet by fact-checkers rejecting it.

increase the size of the dataset, although the impact was minimal. Nevertheless, FacTeR-Check can operate automatically without human intervention.

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

Table A.8
Relation of hoaxes 1–61.

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	Mueren 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 PCR da positivo por nuestros genes endógenos, no por coronavirus	PCR tests positive due to our endogenous genes, not due to coronavirus	Newtral.es
13	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
14	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
15	La definición de pandemia cambió en 2009 por la OMS	The definition of pandemic was changed in 2009 by the WHO	Newtral.es
16	Muere una enfermera de Tennessee (Estados Unidos) tras vacunarse contra la COVID-19	A nurse from Tennessee (United States) died after being vaccinated against COVID-19	La Silla Vacía, Maldita.es, Newtral.es
17	Solo el 6% de las muertes por coronavirus en Estados Unidos fueron realmente por esta causa	Only 6% of coronavirus deaths in United States were actually due to this cause	AFP, Agência Lupa, Animal Político, Chequeado, La Silla Vacía
18	La PCR da positivo por los exosomas, no por coronavirus	PCR tests positive due to exosomes, not due to coronavirus	Newtral.es
19	La mascarilla produce enfermedades neurodegenerativas	Masks produce neurodegenerative diseases	Maldita.es, Newtral.es
20	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
21	La vacuna contra la COVID-19 causa esterilidad	Pfizer vaccines cause sterility	Animal Político, Chequeado, ColombiaCheck, La Silla Vacía, Maldita.es, Newtral.es
22	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
23	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
24	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
25	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

(*continued on next page*)

Table A.8 (*continued*).

26	La vacuna contra la COVID-19 destruye nuestro sistema inmunológico	Vaccines against COVID-19 destroy our immune system	Maldita.es, Newtral.es
27	Pirbright Institute patentó la COVID-19 en 2018	Pirbright Institute patented COVID-19 in 2018	Maldita.es
28	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, Spondeo Media, Verificador (La República)
29	La dieta alcalina previene o cura el coronavirus	Alcaline diets prevent or cure coronavirus	Agência 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
30	El coronavirus fue fabricado en un laboratorio chino	Coronavirus was made in a Chinese lab	Chequeado, Ecuador Chequea, Estadão verifica
31	La mascarilla causa hipoxia	Masks cause hypoxia	Agencia Ocote, Agência Lupa, Animal Político, Aos Fatos, Bolivia Verifica, Chequeado, ColombiaCheck, Cotejo.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 matico cura el coronavirus	Matico 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 Político, Bolivia Verifica, Maldita.es, Newtral.es
36	La NASA catalogueó el dióxido de cloro como antídoto universal en 1988	NASA catalogued chlorine dioxide as a universal antidote in 1988	Animal Político
37	El vino previene o cura el coronavirus	Wine prevents or cures coronavirus	Chequeado, EFE Verifica, Maldita.es, Newtral.es
38	La mascarila 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 la 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, Periodismo 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

(*continued on next page*)

Table A.8 (*continued*).

48	Las personas sanas llevan la mascarilla con la parte blanca hacia fuera para no contagiarse 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 fulfil 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 Político, 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 aisle 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 rcomienda un test pulmonar para identificar el coronavirus	The WHO recommends a pulmonary test to detect coronavirus	EFE Verifica
55	Las pandemias tienen lugar cada 100 años	Pandemics take place every 100 years	AFP, Animal Político, ColombiaCheck, Verificador (La República)
56	El laboratorio de Wuhan tiene relación con Glaxo y Pfizer	Wuhan lab is related to Glaxo and Pfizer	Animal Político, 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 Lupa
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 Político, 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

of March 2021. Moreover, tweets that reply to the query have not been excluded, since they can also misinform. 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 shows the hoaxes in Spanish, the English translation, and the list of fact-checkers that detected every hoax.

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. According to Twitter regulations and to guarantee user 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. The label assigned to each tweet (entailment, contradiction or neutral) according to its relation with the most similar hoax allows to represent how tweets supporting each hoax have been published and spread during the pandemic. 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. 10 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.

Fig. 8 shows the cumulative distribution plot for a general overview of the collected tweets that support the different hoaxes, represented in 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. Therefore, one of the most widely spread hoax (Hoax 31 in Table A.8) is that "masks cause hipoxia". The large number of tweets found supporting this false claim is the reason for the big wave centred on June 2020. Similarly, the peak located in April 2020 is 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. 9 the same plot is displayed without including the hoax 31, which concentrates a 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 more detailed analysis of the misinformation that circulated during the COVID-19 pandemic, Fig. 11 shows the temporal distribution of tweets that support 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 organisations resulted in a higher number of tweets countering and denying the hoax than the number of tweets actually supporting the hoax. For the rest of hoaxes analysed, fact-checkers started a very timid response. However, in case of hoax 15, a false claim stating that "The definition of pandemic was changed in 2009 by the WHO", no presence of fact-checkers denying the hoax can be appreciated. This reveals how complex is this scenario and that further research is required in order to help fact-checkers to detect and undertake further actions to avoid the spread of false claims. In any case, it must be taken into consideration that the response must be proportionate, avoiding an excessive reaction that could increase the dissemination of the hoax and amplify its effects.

7. Conclusion

In this article, we have presented FacTeR-Check, a multilingual machine learning-based architecture to mitigate OSN misinformation. Our architecture provides two pipelines, one for semi-automated verification of claims and another for tracking known hoaxes on social media. The pipelines share three modules: a semantic similarity module, an NLI module, and an 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 performing semi-automated fact-checking. On the other hand, 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 platform. While our case study 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 assess each model individually. 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 MNLI-MT with additional data, which performs adequately on XNLI test, offering similar results to state-of-the-art models, as well as offering multilingual capabilities. Finally, the information retrieval module is compared with the statistical algorithms RAKE and YAKE, and the straightforward KeyBERT on five multilingual keyword extraction benchmarks. 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. In addition, identified hoaxes have their own particular activity peaks. While some of them spread over a large time period, others have a greater impact; 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, which means 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 beyond the scope of the search algorithm. There is also room for improvement in the semantic similarity and NLI modules. For instance, further research is required on how to improve the NLI module and its efficiency or to further evaluate how it performs in exceptional cases such as sentences with low semantic similarity. 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 data modalities.

CRediT authorship contribution statement

Alejandro Martín: Conceptualization, Data curation, Writing, Funding acquisition, Supervision. **Javier Huertas-Tato:** Conceptualization, Software, Validation, Visualization, Methodology. **Álvaro Huertas-García:** Conceptualization, Software, Visualization, Writing, Methodology. **Guillermo Villar-Rodríguez:** Conceptualization, Software, Validation, Writing, Data curation. **David Camacho:** Conceptualization, Writing, Resources, Funding acquisition, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

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, Spain under S2018/TCS-4566 grant, by European Commission under IBERIFIER - Iberian Digital Media Research and Fact-Checking Hub (2020-EU-IA-0252), by "Convenio Plurianual with the Universidad Politécnica de Madrid in the actuation line of *Programa de Excelencia para el Profesorado Universitario*" and by the research project DisTrack: Tracking disinformation in Online Social Networks through Deep Natural Language Processing, granted by Barcelona Mobile World Capital Foundation.

Appendix

See Table A.8.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[2] N. Kotonya, F. Toni, Explainable automated fact-checking: a survey, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 5430–5443.

[3] N. Naderi, 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.

[4] M. Trokhymovych, D. Saez-Trumper, Wikicheck: An end-to-end open source automatic fact-checking api based on wikipedia, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 4155–4164.

[5] A. Alonso-Reina, R. Sepúlveda-Torres, E. Saquete, M. Palomar, Team GPLSI. Approach for automated fact checking, in: Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER), 2019, pp. 110–114.

[6] U. Naseem, I. Razzak, K. Musial, M. Imran, Transformer based deep intelligent contextual embedding for twitter sentiment analysis, Future Gener. Comput. Syst. 113 (2020) 58–69.

[7] T. Zhang, V. Kishore, F. Wu, K.Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, 2019, arXiv preprint arXiv:1904.09675.

[8] W. Yang, Y. Xie, A. Lin, X. Li, L. Tan, K. Xiong, M. Li, 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.

[9] N. Zhang, Learning adversarial transformer for symbolic music generation, IEEE Trans. Neural Netw. Learn. Syst. (2020).

[10] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, D. Tran, Image transformer, in: International Conference on Machine Learning, PMLR, 2018, pp. 4055–4064.

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2019, arXiv:1810.04805.

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, 2019, arXiv:1907.11692.

[13] A. Conneau, G. Lample, Cross-lingual language model pretraining, Adv. Neural Inf. Process. Syst. 32 (2019) 7059–7069.

[14] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, 2019, arXiv preprint arXiv:1911.02116.

[15] R. Mihalcea, C. Corley, C. Strapparava, et al., Corpus-based and knowledge-based measures of text semantic similarity, in: Aaai, Vol. 6, 2006, pp. 775–780.

[16] W.H. Gomaa, A.A. Fahmy, et al., A survey of text similarity approaches, Int. J. Comput. Appl. 68 (13) (2013) 13–18.

[17] E. Millar, D. Shen, J. Liu, C. Nicholas, Performance and scalability of a large-scale n-gram based information retrieval system, J. Digit. Inf. 1 (5) (2000).

[18] J. Singthongchai, S. Niwattanakul, A method for measuring keywords similarity by applying jaccard's, n-gram and vector space, Lecture Notes Inf. Theory 1 (4) (2013).

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

[20] 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.

[21] M. Schuhmacher, 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.

[22] N. Peinelt, D. Nguyen, 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.

[23] P. Kasnesis, R. Heartfield, X. Liang, L. Toumanidis, G. Sakellari, C. Patrikakis, G. Loukas, Transformer-based identification of stochastic information cascades in social networks using text and image similarity, Appl. Soft Comput. 108 (2021) 107413.

[24] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, 2019, arXiv preprint arXiv:1908.10084.

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

[26] M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, 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), 2014, pp. 216–223, URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/363_Paper.pdf.

[27] B. MacCartney, Natural Language Inference, Stanford University, 2009.

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

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

[30] A. Williams, N. Nangia, 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), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1112–1122, http://dx.doi.org/10.18653/v1/N18-1101.

[31] A. Conneau, R. Rinott, G. Lample, A. Williams, S.R. Bowman, H. Schwenk, 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.

[32] Q. Chen, X. Zhu, Z.-H. Ling, S. Wei, H. Jiang, 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.

[33] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, 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.

[34] J. Huertas-Tato, A. Martín, D. Camacho, SML: a new semantic embedding alignment transformer for efficient cross-lingual natural language inference, 2021, arXiv preprint arXiv:2103.09635.

[35] D. Graves, Understanding the promise and limits of automated fact-checking, 2018.

[36] J. Thorne, 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.

[37] M. Granik, 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.

[38] S. Miranda, D. Nogueira, A. Mendes, A. Vlachos, A. Secker, R. Garrett, J. Mitchel, Z. Marinho, Automated fact checking in the news room, in: The World Wide Web Conference, 2019, pp. 3579–3583.

[39] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, Fever: a large-scale dataset for fact extraction and verification, 2018, arXiv preprint arXiv:1803.05355.

[40] A. Hanselowski, C. Stab, C. Schulz, Z. Li, 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.

[41] A. Sathe, S. Ather, T.M. Le, N. Perry, J. Park, Automated fact-checking of claims from Wikipedia, in: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 6874–6882.

[42] A. Huertas-Garcıia, J. Huertas-Tato, A. Martín, D. Camacho, Civic-UPM at CheckThat! 2021: integration of transformers in misinformation detection and topic classification, 2021.

[43] P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barrón-Cedeño, P. Papotti, S. Shaar, G.D.S. Martino, Automated fact-checking for assisting human fact-checkers, 2021, arXiv preprint arXiv:2103.07769.

[44] X. Zeng, A.S. Abumansour, A. Zubiaga, Automated fact-checking: A survey, Lang. Linguist. Compass 15 (10) (2021) e12438.

[45] Z. Guo, M. Schlichtkrull, A. Vlachos, A survey on automated fact-checking, Trans. Assoc. Comput. Linguist. 10 (2022) 178–206.

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

[47] A. Chernyavskiy, D. Ilvovsky, P. Nakov, WhatTheWikiFact: Fact-checking claims against wikipedia, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 4690–4695.

[48] M. Nadeem, W. Fang, B. Xu, M. Mohtarami, J. Glass, FAKTA: An automatic end-to-end fact checking system, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), 2019, pp. 78–83.

[49] H. Karimi, P. Roy, S. Saba-Sadiya, J. Tang, Multi-source multi-class fake news detection, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 1546–1557.

[50] G. Karadzhov, P. Nakov, L. Màrquez, A. Barrón-Cedeño, I. Koychev, Fully automated fact checking using external sources, 2017, arXiv preprint arXiv:1710.00341.

[51] N. Hassan, F. Arslan, C. Li, 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.

[52] D. Stammbach, E. Ash, E-fever: Explanations and summaries for automated fact checking, in: Proceedings of the 2020 Truth and Trust Online Conference (TTO 2020), Hacks Hackers, 2020, p. 32.

[53] A. Pathak, M.A. Shaikh, R. Srihari, Self-supervised claim identification for automated fact checking, 2021, arXiv preprint arXiv:2102.02335.

[54] G. Ruffo, A. Semeraro, A. Giachanou, P. Rosso, Surveying the research on fake news in social media: a tale of networks and language, 2021, arXiv:2109.07909.

[55] F. Monti, F. Frasca, D. Eynard, D. Mannion, M.M. Bronstein, Fake news detection on social media using geometric deep learning, 2019, arXiv preprint arXiv:1902.06673.

[56] J. Zhang, B. Dong, S.Y. Philip, Fakedetector: Effective fake news detection with deep diffusive neural network, in: 2020 IEEE 36th International Conference on Data Engineering (ICDE), IEEE, 2020, pp. 1826–1829.
[57] A. Huertas-García, A. Martín, J. Huertas-Tato, D. Camacho, Profiling hate speech spreaders on Twitter: Transformers and mixed pooling, in: CLEF (Working Notes) 2021, 2021.
[58] K. Shu, S. Wang, H. Liu, Understanding user profiles on social media for fake news detection, in: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), IEEE, 2018, pp. 430–435.
[59] K. Shu, X. Zhou, S. Wang, R. Zafarani, H. Liu, The role of user profiles for fake news detection, in: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2019, pp. 436–439.
[60] R. Speer, J. Chin, An ensemble method to produce high-quality word embeddings (2016), 2019, arXiv:1604.01692.
[61] W. Yin, H. Schütze, Learning meta-embeddings by using ensembles of embedding sets, 2015, arXiv:1508.04257.
[62] N. Alami, M. Meknassi, N. En-nahnahi, Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning, Expert Syst. Appl. 123 (2019) 195–211, http://dx.doi.org/10.1016/j.eswa.2019.01.037, URL https://www.sciencedirect.com/science/article/pii/S0957417419300375.
[63] S.A. Devi, S. Sivakumar, A hybrid ensemble word embedding based classification model for multi-document summarization process on large multi-domain document sets, Int. J. Adv. Comput. Sci. Appl. 12 (9), (2021).
[64] B. Subba, S. Kumari, A heterogeneous stacking ensemble based sentiment analysis framework using multiple word embeddings, Comput. Intell. (2021).
[65] A. Huertas-García, J. Huertas-Tato, A. Martín, D. Camacho, Countering misinformation through semantic-aware multilingual models, in: H. Yin, D. Camacho, P. Tino, R. Allmendinger, A.J. Tallón-Ballesteros, K. Tang, S.-B. Cho, P. Novais, S. Nascimento (Eds.), Intelligent Data Engineering and Automated Learning – IDEAL 2021, Springer International Publishing, Cham, 2021, pp. 312–323.
[66] D. Cer, M. Diab, E. Agirre, I.n. Lopez-Gazpio, 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), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1–14, http://dx.doi.org/10.18653/v1/S17-2001, URL https://www.aclweb.org/anthology/S17-2001.
[67] J. Ham, Y.J. Choe, K. Park, I. Choi, H. Soh, Kornli and korsts: New benchmark datasets for Korean natural language understanding, 2020, arXiv:2004.03289.
[68] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, 2020, arXiv:2004.09813.
[69] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzman, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, 2019, arXiv arXiv:1911.02116, URL https://arxiv.org/abs/1911.02116v2.
[70] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020, arXiv:2002.10957.
[71] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, Mpnet: Masked and permuted pre-training for language understanding, 2020, arXiv:2004.09297.
[72] G. Sidorov, A. Gelbukh, H. Gómez-Adorno, D. Pinto, Soft similarity and soft cosine measure: similarity of features in vector space model, Comput. Y Sistemas 18 (3) (2014) http://dx.doi.org/10.13053/cys-18-3-2043.
[73] A. Talman, S. Chatzikyriakidis, Testing the generalization power of neural network models across nli benchmarks, 2019, arXiv:1810.09774, [Cs], URL http://arxiv.org/abs/1810.09774.
[74] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, D. Kiela, Adversarial NLI: A new benchmark for natural language understanding, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020.
[75] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv arXiv:1412.6980, URL https://arxiv.org/abs/1412.6980v9.
[76] M. Grootendorst, Keybert: Minimal keyword extraction with bert, 2020, http://dx.doi.org/10.5281/zenodo.4461265.
[77] I. Montani, M. Honnibal, S. Van Landeghem, A. Boyd, "Spacy: Industrial-strength natural language processing in python", 2020, http://dx.doi.org/10.5281/zenodo.1212303.
[78] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, R. Vollgraf, FLAIR: An easy-to-use framework for state-of-the-art NLP, in: NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), 2019, pp. 54–59.
[79] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T.L. Scao, S. Gugger, M. Drame, Q. Lhoest, A.M. Rush, Huggingface's transformers: State-of-the-art natural language processing, 2019, http://dx.doi.org/10.48550/ARXIV.1910.03771, arXiv URL https://arxiv.org/abs/1910.03771.
[80] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, M. Johnson, Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization, 2020, CoRR arXiv:2003.11080.
[81] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, 2016, arXiv preprint arXiv:1607.01759.
[82] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, T. Mikolov, Fasttext.zip: Compressing text classification models, 2016, arXiv preprint arXiv:1612.03651.
[83] H. Schwenk, V. Chaudhary, S. Sun, H. Gong, F. Guzmán, Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia, 2019, arXiv:1907.05791.
[84] J. Tiedemann, Parallel data, tools and interfaces in OPUS, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 2214–2218.
[85] M. Artetxe, H. Schwenk, Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond, Trans. Assoc. Comput. Linguist. 7 (2019) 597–610, http://dx.doi.org/10.1162/tacl_a_00288, http://arxiv.org/abs/https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00288/1923278/tacl_a_00288.pdf.
[86] H. Huang, Y. Liang, N. Duan, M. Gong, L. Shou, D. Jiang, M. Zhou, Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 2485–2494, http://dx.doi.org/10.18653/v1/D19-1252, URL https://aclanthology.org/D19-1252.
[87] L. Marujo, A. Gershman, J. Carbonell, R. Frederking, J.P. Neto, Supervised topical key phrase extraction of news stories using crowdsourcing, light filtering and co-reference normalization, 2013, http://dx.doi.org/10.48550/ARXIV.1306.4886, arXiv URL https://arxiv.org/abs/1306.4886.
[88] I. Augenstein, M. Das, S. Riedel, L. Vikraman, A. McCallum, Semeval 2017 task 10: Scienceie - extracting keyphrases and relations from scientific publications, 2017, http://dx.doi.org/10.48550/ARXIV.1704.02853, arXiv URL https://arxiv.org/abs/1704.02853.
[89] A. Bougouin, F. Boudin, B. Daille, Topicrank: Graph-based topic ranking for keyphrase extraction, in: Proceedings of the Sixth International Joint Conference on Natural Language Processing, Asian Federation of Natural Language Processing, Nagoya, Japan, 2013, pp. 543–551.
[90] G.O. Aquino, L.C. Lanzarini, Keyword identification in spanish documents using neural networks, J. Comput. Sci. Technol. 15 (2015).
[91] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, A. Jatowt, Yake! keyword extraction from single documents using multiple local features, Inform. Sci. 509 (2020) 257–289, http://dx.doi.org/10.1016/j.ins.2019.09.013.
[92] S. Rose, D. Engel, N. Cramer, W. Cowley, Automatic keyword extraction from individual documents, in: Text Mining, John Wiley & Sons, Ltd, Chichester, UK, 2010, pp. 1–20.