Semantic similarity models for automated fact-checking: *ClaimCheck* as a claim matching tool

Irene Larraz; Rubén Míguez; Francesca Sallicati

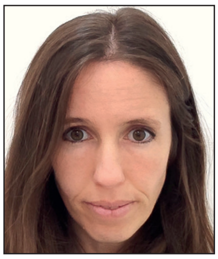
Nota: Este artículo se puede leer en español en:
https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/87284

Recommended citation:

Larraz, Irene; **Míguez, Rubén**; **Sallicati, Francesca** (2023). "Semantic similarity models for automated fact-checking: *ClaimCheck* as a claim matching tool". *Profesional de la información*, v. 32, n. 3, e320321.

https://doi.org/10.3145/epi.2023.may.21

Manuscript received on 15th February 2023
Accepted on 18th May 2023



Irene Larraz ✉
https://orcid.org/0000-0002-9164-6610

Universidad de Navarra
Campus Universitario
31009 Pamplona, Spain
ilarraz@alumni.unav.es

Rubén Míguez
https://orcid.org/0000-0001-9395-1849

Newtral
Vandergoten, 1
28014 Madrid, Spain
ruben.miguez@newtral.es

Francesca Sallicati
https://orcid.org/0000-0001-9981-8360

Newtral
Vandergoten, 1
28014 Madrid, Spain
francesca.sallicati@gmail.com

Abstract

This article presents the experimental design of *ClaimCheck*, an artificial intelligence tool for detecting repeated falsehoods in political discourse using a semantic similarity model developed by the fact-checking organization *Newtral* in collaboration with *ABC Australia*. The study reviews the state of the art in algorithmic fact-checking and proposes a definition of claim matching. Additionally, it outlines the scheme for annotating similar sentences and presents the results of experiments conducted with the tool.

Keywords

Verification; Automated fact-checking; Claim matching; Semantic similarity; Paraphrase models; Disinformation; Artificial intelligence; AI; Algorithms; Software.

Funding

This publication is part of the R&D&I project "Digital native media in Spain: typologies, audiences, building trust and keys to journalistic sustainability" - *Diginativemedia*. Reference PID2021-122534OB-C22, funded by the *Ministry of Science and Innovation/State Research Agency*/10.13039/501100011033/ and the *European Regional Development Fund* "A way of doing Europe".

1. Introduction

In 2020, *The Washington Post* identified more than 50 lies that former US President Donald Trump had repeated at least 20 times during his term in office. Some of them had been uttered more than 200 times (**Kessler**; **Fox**, 2021). Two years earlier, the media had to add a new category to rate such lies repeated more than 20 times. The head of fact-checking at the media outlet, Glenn Kessler, explained that this figure is sufficient to demonstrate that, far from being a mistake, there is an intention to deceive, which ultimately turns into propaganda (*The Washington Post*, 2018).

Fact-checkers dedicate considerable time and effort to combat misinformation, providing verified information to the public discourse and increasing the political cost of lying. Therefore, identifying repeated lies has become a priority task in political discourse fact-checking.

In recent years, advancements in the field of artificial intelligence and language models through deep learning have propelled the development of automated solutions to detect these repetitions. The aim is to assist journalists in enhancing their ability to monitor politicians' statements, increase the reach of fact-checking, and provide faster responses to audiences. However, the use of natural language processing (NLP) techniques has encountered various challenges in identifying similar phrases, particularly when pronounced with different words or expressions. Consequently, most models designed for this purpose are not yet operational in newsrooms.

Fact-checking organizations have also begun developing their own tools to automate claim matching, which refers to the task of identifying phrases with the same meaning. In the context of fact-checking, this concept entails the process of comparing a claim with previously verified ones to determine if there is any match or similarity.

The UK-based fact-checking organization *Full Fact* is working on a tool to review whether a politician's claim resembles another that they have already verified, comparing it against their archive of publications and alerting journalists when a repetition is detected (**Corney**, 2021a; **Floodpage**, 2021). Similarly, the *Duke Reporters' Lab* at *Duke University* is experimenting with *Squash*, a tool that allows for finding similar claims among fact-checks from various US media outlets to provide live alerts during electoral debates (**Adair**, 2021).

The purpose of this study is to address the challenge of detecting repeated false information by designing an experiment to validate a potential solution: the reuse of previous fact-checks and the improvement of semantic similarity systems to search for linguistic repetitions in political discourse. The central hypothesis is that the use of semantic similarity systems can significantly enhance the automation of fact-checking by identifying false claims that utilize paraphrasing to express the same idea. This article provides a detailed description of the experimental methodology and the obtained results, as well as their implications for future research on detecting false information in the political domain.

Intending to improve the precision of detecting similar phrases, this study conducts a comparative analysis of three architectures by evaluating their performance in creating *ClaimCheck*. This tool was developed jointly by the fact-checking organization *Newtral* and *ABC Australia*, as part of the *JournalismAI* program coordinated by the *London School of Economics*. The system was designed to meet the specific needs of journalists and is the result of collaboration between engineers and fact-checkers. Furthermore, the model has been tested in a newsroom environment, where it has been used to continuously improve its results.

Additionally, the study reviews the state of the art in the application of such algorithms for fact-checking (Section 2), proposes a definition of claim matching (Section 3), details the methodology of the experimental design, including an analysis of the challenges faced by these models (Section 4), and conducts an experiment to compare three potential architectures (Section 5). Finally, it presents an analysis of the tool's implementation in the newsroom and suggests future research directions to enhance claim matching and contribute to the automation of fact-checking through artificial intelligence (Section 6).

2. State of the art

The academic literature has extensively documented the need to address repeated lies and the problems generated by the repetition of misinformation, as well as the negative effects of amplifying false messages (**Phillips**, 2018; **Wardle**, 2018).

Babakar and **Moy** (2016) emphasize that public discourse heavily relies on repetition, making it a crucial aspect of the roadmap for automated fact-checking and a challenge that researchers still need to address. **Graves** (2018) also highlights the detection of repeated falsehoods as a fundamental element of automated fact-checking and suggests that the most effective method for these tools is to compare statements with a library of previously verified claims by one or multiple fact-checking organizations to enhance their reach and responsiveness to false assertions.

Thorne and **Vlachos** (2018) also refer to claim matching as a popular model employed by fact-checking organizations, wherein a claim is compared with previously verified ones. Similarly, **Nakov** *et al.* (2021) provide an overview of the technological advancements available for automated fact-checking and the detection of previously verified claims as one of the primary approaches. They also emphasize the role of context in this task, including the use of neighboring phrases, coreference resolution, and reasoning about the target text.

> We understand claim matching as the task of identifying statements that share a common meaning, even if they are expressed in different ways

The effects of repeated lies have also been examined in other fields, such as Psychology, which has investigated what is known as the illusory truth effect. This effect occurs when the repetition of a deception creates familiarity in the listener, reducing doubts about its truthfulness and leading to the belief that what is being said is true (**Hassan**; **Barber**, 2021; **Murray** *et al.*, 2020; **Agadjanian** *et al.*, 2019). Some of these studies have also attempted to measure the likelihood of politicians repeating a claim after it has been fact-checked. For instance, one study found that the chances of a claim being repeated in the five days following a fact-check publication decreased by 9.5 percentage points (**Lim**, 2018).

Scientific evidence has also analyzed the various political responses that arise from fact-checking. In some cases, it has been found that politicians continue to cling to falsehoods even after being debunked with compelling evidence against their integrity. This can be attributed to reasons ranging from disagreement to demagogic strategies (**Sippitt**, 2020; **Porter**; **Wood**, 2021).

On the other hand, in the field of computer science, claim matching has emerged as one of the primary benchmarks for NLP (**Nakov** *et al.*, 2022). Although paraphrasing and semantic similarity models still pose unresolved challenges, advancements in language models, specifically large language models (LLMs), have brought new perspectives to enhance the accuracy of these methods. Specifically, the domain of Semantic Textual Similarity (STS) focuses on the task of measuring the similarity in meaning between sentences. In this regard, models assess the extent to which two claims are semantically, textually, lexically, and referentially similar (**Hövelmeyer**; **Boland**; **Dietze**, 2022; **Martín** *et al*., 2021; **Sheng** *et al*., 2021).

To train these models, researchers work with datasets and pre-trained models such as *SemEval-PIT2015* (en), the *STS Benchmark*, or the *Microsoft Research Paraphrase Corpus*, among others (**Dolan**; **Brockett**, 2005; **Lan** *et al.*, 2019).

Researchers have attempted to address some of the challenges faced by the technology, such as the ambiguity of claims and the need for more contextual information (**Shaar** *et al*., 2021a; 2021b; **Reimers**; **Gurevych**, 2019; **Nguyen**; **Karimi**; **Xing**, 2021). Others have taken different approaches, for example, framing it as a classification problem over a collection of previously verified claims (**Mansour**; **Elsayed**; **Al-Ali**, 2022). **Kazemi** *et al.* (2022) also propose a model based on a binary classifier using *XLM-RoBERTa* (*XLM-R*), a popular multilingual language model, and a semantic similarity search system utilizing sentence embeddings from *LaBSE*, *SBERT*, and pairwise cosine similarity.

To achieve this, various paths have been explored. For instance, the *University of Texas at Arlington* explores structured and semantic models that can capture multiple aspects of a factual claim, such as the topic, the type of expressed data, the involved entities and their relationships, quantities, times, intervals, comparisons, and aggregated structures (**Arslan**, 2021). Another example is the *Berkeley FrameNet* system (**Baker**; **Fillmore**; **Lowe**, 1998), which relies on frame semantics, a branch of meaning theory, with 13 categories of factual claims. Currently, the tools are designed as hybrid systems in which the judgment of the journalist is necessary, and the technology is considered a support to enhance search capabilities and improve response speed when a falsehood resurfaces.

Within the broader framework of studies at the intersection of Journalism and artificial intelligence, fact-checking has been the subject of extensive analysis due to its potential for automation (*Stanford Institute for Human-Centered Artificial Intelligence*, 2023). Despite persistent challenges in achieving full automation of fact-checking, such as the need to improve the accuracy of natural language processing tools and a more comprehensive database of verified facts, there is a growing number of initiatives that leverage AI to perform this task.

In the field of automated fact-checking, claim matching is intertwined with other tasks, such as claim detection, which involves identifying factual claims within political discourse; check-worthiness, which assesses the relevance of a claim to prioritize verifications; and claim validation, which involves corroborating the data to confirm the veracity of a claim by seeking evidence from open data sources (**Hassan** *et al.*, 2015; 2017; **Graves**, 2018; **Zeng**; **Abumansour**; **Zubiaga**, 2021). These tasks have driven the development of artificial intelligence applied to fact-checking, an area that has attracted significant attention due to its distinctive features, such as the structure of fact-checks or the methodology employed (**Nakov** *et al.*, 2021).

3. Towards a definition of claim matching

Revised definitions identify claim matching as a task aimed at detecting pairs of statements whose meaning coincides, even if the words or grammatical structures used to convey that meaning differ. However, although there is a general consensus on the nature of this task, there is no universal definition that precisely delineates the concept.

Full Fact describes claim matching for fact-checking purposes as a task to identify statements with a shared truth condition (**Corney**, 2021b). In other words, similar phrases are considered if both contain a variable that makes them true. This means that there cannot be one certain statement and another that is not. For example, they cite the statement “It is raining in London,” which is true under the condition that it is raining in that city. A similar phrase could be “It is damp outside” or “It is pouring.”

On the other hand, **Kazemi** *et al.* (2021) define claim matching as the task of identifying pairs of textual messages containing statements that can be addressed with the same fact-check. From a more technical perspective, **Shaar** *et al.* (2020) reflect a similar idea, framing the task as a classification problem to find verifications that can help debunk the initial claim. This implies that the fact-check contains the new statement. The work of **Adair** *et al.* (2018) follows the same line of thought.

Jiang *et al.* (2021) provide a broader definition that includes

> "messages containing falsehoods, inaccuracies, rumors, decontextualized truths, or deceptive logical leaps that convey similar information/themes as the claim but, for example, with a different name of the person or event (...)."

For these authors, it is not only about identifying phrases with a common meaning but also grouping statements that share a theme or information, even if they are presented differently or use different names or events. This expansion of the definition highlights the importance of considering the broader context in which they are presented, such as when a politician repeats the same data but changes the reference to the city in which it is being uttered.

In summary, our proposal is based on the definitions provided by the aforementioned authors and aims to encompass the common factors among them. Thus, we understand claim matching as the task of identifying statements that share a common meaning, even if they are expressed in different ways.

4. Experimental design methodology

The experimental design of the claim matching model was based on an iterative process that involved the evaluation and comparison of various architectures to determine their effectiveness. To carry out this task, a comprehensive analysis of previously published scientific literature in related studies was conducted, and consultations were made with experts in the field of data verification. Based on this knowledge, empirical evaluation of the models was performed through testing using a training dataset that contained phrases with varying degrees of similarity.

In doing so, the work of **Dolan** and **Brockett** (2005) on the *Microsoft Research Paraphrase Corpus* and the study by **Kazemi** *et al.* (2021) served as references. From this corpus, an annotation guideline was developed, which defined the criteria for labeling pairs of statements.

Table 1. Examples of annotated statements and reasoning for the decision

Claim 1	Claim 2	Tag	Reason
More than 4,100 arrested for gender-based violence since the start of the lockdown	Nearly 9,000 arrested for gender-based violence during the state of emergency	Similar	Both messages contain a figure regarding the number of arrests during the state of emergency.
The government inflates the data of tests conducted to boast that we are in the 'top ten' worldwide.	The government falsifies the data it sends to the *OECD* to be in the top 10 of tests conducted.	Similar	Both statements point to manipulated data by the government regarding the number of tests conducted to appear "in the top 10".
20% of the infections in Spain are among healthcare personnel.	For the first time, more than 20 percent of the infected population consists of healthcare workers.	Similar	There are three shared elements: the percentage, and references to infections and healthcare personnel.
"Gender-based violence is the most common crime during the state of alarm."	The pandemic of sexism: Aggressions against minors or people close to women increase as a form of sexist violence, according to the *Prosecutor's Office*.	Dissimilar	Both refer to gender-based or sexist violence during the lockdown, however, one focuses on the frequency of the crime while the other addresses where it occurs.
282,891 more unemployed, and we must add the million self-employed who have ceased their activity and the 3.3 million Spaniards in ERTE*.	To boast about management with over 70,000 deaths, nearly four million unemployed, 750,000 in ERTE*, and a million self-employed on the brink can only be done by someone living in La Moncloa, oblivious to the problems of the Spanish people.	Similar	Although both statements provide multiple data points, there is one that coincides in both regarding the number of self-employed individuals who have lost their activity.

* Employment Regulation File

The training data was obtained from a database of 200,000 factual statements provided by *Newtral*. For each sentence, the three most similar candidates were selected based on cosine similarity scores, a measure used to assess the similarity between two entities in a vector space. From these candidates, sentence pairs were extracted for annotation based on threshold values predefined in a similar dataset (**Dolan**; **Brockett**, 2005). According to these thresholds, the following were selected:

- 50% of pairs with cosine similarity >= 0.8
- 25% of pairs with cosine similarity < 0.8 and >= 0.7
- 25% of pairs with cosine similarity < 0.7 and >= 0.4

The objective was to obtain a training dataset consisting of 10,000 annotated pairs of statements, including both manually annotated pairs and pairs with weak labels obtained through heuristics. The heuristics involved applying rules to identify common features in factual statements, such as the presence of numbers, percentages, geographic names, and other indicators of factuality. While labels obtained through heuristics are less precise than manually obtained labels, they allow for a significant expansion of the sample size and, consequently, improve the effectiveness of the training process.

Once the dataset was obtained, labels or similarity parameters were assigned to the sentences, and annotation criteria were defined. Different approaches were considered in this process, including assigning numerical values for measuring similarity or classifying similarity into distinct categories.

> The experimental design of the claim matching model was based on an iterative process that involved the evaluation and comparison of various architectures to determine their effectiveness

Firstly, an evaluation was conducted using a preliminary definition of criteria on 100 pairs of statements, with three different annotators labeling the pairs of sentences into three categories: similar, dissimilar, or related. We also included a "N/A" mark in case the proposed sentence was not factual. The results revealed a high discrepancy, with a 50% difference among the annotators. Out of those sentences, half showed a complete disagreement among the annotators, as each one applied a different label. Subsequently, the three annotators conducted a joint review to align the annotation criteria. A second trial was prepared, and this time an improvement in the alignment of the annotations was observed. Although a 30% discrepancy persisted, only one annotator marked a different label on this occasion. Finally, the annotation criteria were reviewed, and it was agreed to eliminate one of the labels ('related') to avoid noise. With this new approach, the data were reannotated.

4.1. Confining degrees of similarity: a proposed framework for data annotation

Confining the definition of claim matching entails identifying what is considered a similar sentence. Several factors come into play in this step. The first one arises from language usage issues, such as the use of pronouns like 'he' or temporal and spatial references like 'yesterday' or 'here'. An extreme example of this is how one can refer to the same person using their name, surname, initials, position, or former position, among many others, which makes it challenging for automated systems to identify them. This problem falls within a broader issue, that of the ambiguity of some statements.

In addition to the mentioned factors, another essential component for deciphering the meaning of a statement and determining its similarity to others is context. This element is crucial in understanding short statements where the context is difficult to extrapolate (**Shaar** *et al.*, 2021b). Particularly in oral language statements, necessary details are often omitted to understand what is being referred to. The conversation's context or the speaker's context may contain relevant information that algorithms find difficult to deduce otherwise. For instance, if it is uttered by a regional politician, it likely refers to that specific region, whereas if the same statement is made by a national politician, it extends to a national scope.

The data itself poses another obstacle to retrieving previously verified sentences, as the same data can be presented in different ways. In some cases, politicians may use rounding or switch from absolute numbers to relative values, which can lead to interpretation errors. Fact-checkers have pointed out that in many cases, the idea politicians seek to support is more relevant than the data itself. An example of this is the statement made by some members of the *Popular Party* in Spain regarding the number of unemployed individuals in the country, where a variety of data points were used, exceeding the actual figure. In total, different party members had repeated at least 19 times that there were more millions of unemployed individuals than those actually recorded by the *National Institute of Statistics* (**Real**, 2021). However, the figures mentioned by the political representatives ranged from four to six million unemployed, making it difficult for a system that only identified the repetition of figures to detect this discrepancy.

Taking the above into consideration, for the development of *ClaimCheck* , two sentences were considered similar if they met the following criteria:

- They refer to the same thing, even if the data varies or is contradictory.
- One of the statements includes specific details that do not invalidate the other.
- One of the statements refers to the same reality from a different perspective than the other; for example, one provides the total quantity while the other refers to the percentage variation, but both convey the same message.

According to the established criteria, the data annotation stage used to train a claim matching model is a critical aspect for the success of its implementation, as the selected parameters and labels must accurately reflect the similarity between the statements. In the case of *ClaimCheck*, a binary classification with two categories was chosen in order to simplify the classification task and obtain more precise results. It is important to emphasize that the selection of categorization should be carefully evaluated to avoid classification errors and ensure the quality of the annotations made.

4.2. Issues in fact-checks retrieval

Beyond what is considered similar in the annotation phase, there are problems arising from its specific application for fact-checking in order to retrieve previously published fact-checks. In these cases, other factors also come into play, such as temporality. A statement that was false at a certain moment may become true over time, or vice versa. This problem arises when retrieving fact-checks whose data may have been modified over time due to updated statistics and fluctuating values, as well as variations in the context in which the data is used. In other words, Statement A and Statement B are similar, and the similarity between them persists over time, but it may not be possible to use Statement A to retrieve a fact-check with Statement B if that verification has "expired," even though it does not affect the underlying concept or definition of similarity between them.

> Finding a balance between system precision and its ability to retrieve all similar phrases is a challenging task on claim matching

Nuances are another crucial element in identifying two similar statements in the field of fact-checking, as the introduction of a single word can completely change the meaning of the statement, altering the rating or classification given to it before, shifting from true to false or other categories. For example, the statement "50% of new contracts are indefinite" was true at a certain moment if it referred to those contracts generated from a specific date, in this case, following labor reform. However, if the word "new" is removed and the totality of contracts is analyzed, it would be false to claim that "50% of contracts are indefinite." Again, in this case, both statements are similar according to the criteria established in the experimental design, but it would create problems to retrieve a fact-check to assess the other.

The same occurs with the context in the specific case of fact-checking. Referring to the previous example, it is not the same for the statement "50% of new contracts are indefinite" to be made by a minister, who is referring to the national level, as another statement indicating that "50% of the contracts we have made are indefinite" said by a regional president, who generally refers to the regional scope of their community. Both statements are semantically similar, but the context implies that they are referring to different data related to different locations. These types of problems present challenges when interpreting the results generated by the algorithm.

5. Method and experiments on *ClaimCheck*

ClaimCheck is built on the foundation of *ClaimHunter*, a claim detection tool that identifies factual statements in *Twitter* messages (**Beltrán**; **Míguez**; **Larraz**, 2019). These are different algorithms applied in different phases of the process. While the latter focuses on the detection of statements to be verified, the former works once that filtering is done and seeks similar cases that have been previously verified. Thus, to process each statement, political discourse first goes through the *ClaimHunter* algorithm to confirm its factual nature and then through the *ClaimCheck* algorithm, which is responsible for retrieving candidate phrases to be labeled as similar.

The first step is to identify what is used as the archive, whether it includes only the dataset of previous fact-checks or also other statements from politicians on social media or in the media. *ClaimCheck* utilizes both sources to achieve two objectives: on one hand, it allows the reuse of previously conducted verifications for faster action, and on the other hand, it helps identify disinformation campaigns when multiple statements articulated in different ways convey the same false message. To do this, *ClaimCheck* collects fact-checks from the *Fact Check Explorer*, a *Google* tool that stores fact-checkers' publications using the *ClaimReview* markup. The *ClaimReview* markup is a structured data system that facilitates the extraction of categories within a verification, such as the claim being verified, the author of the statement, and the rating indicating the degree of falsehood. Additionally, *ClaimCheck* also utilizes verifications from other fact-checkers not included in the *Google* tool. As a result, it has a database of 300,000 fact-checks in over 20 different languages, which in the future could expand its scope beyond political discourse fact-checking to include the debunking of disinformation circulating on social media.

Next, we detail each of the phases: the construction of the training and test datasets, the design of the architectures, and the analysis of the experimental results.

5.1. Construction of the training and test datasets

To train the model, *ClaimCheck* required identifying valid candidates, i.e., detecting phrases in the fact-checking database that had the potential to convey the same meaning as the input phrase. In this process, the aforementioned problems of semantic similarity arise (see Section 4.2), as the system searches for similar phrases. Additionally, the *ClaimCheck* system needed to evaluate whether the two phrases refer exactly to the same thing, which poses a problem of semantic meaning.

The *ClaimCheck* system, therefore, consists of two components: a first component that performs a search for similar phrases, and a classifier that evaluates the similarity between the original claim and the proposed candidates.

To construct the training dataset, annotations from the *Fact Check Explorer* were used. On the other hand, pairs were selected from the *ClaimHunter* dataset to construct the test dataset (test benchmark), which was used to evaluate the classification model.

This led to the adoption of two strategies. On one hand, phrases were extracted from tweets that could contain more than one sentence, as well as complete tweets that included additional information. This extraction of phrases was performed by applying tokenization functions to obtain the different phrases that compose each tweet, and they were then passed through *ClaimHunter* to discard non-factual phrases. Specifically, three types of pairs were generated from the *ClaimHunter* database:

- Tweet-to-tweet pairs
- Phrase-to-tweet pairs
- Phrase-to-phrase pairs

All these possible combinations of phrase pairs underwent cosine similarity filters using the sentence-transformer library and the *sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2* model. Specifically, pairs with a cosine similarity below 0.7 were discarded, and a subset was selected for annotation. In total, the training dataset consisted of 7,762 pairs, with 66% being similar claims, while the test (*ClaimHunter*) benchmark comprised 2,246 pairs with a 45% similarity. Additionally, the *Prodigy* tool was used for annotation purposes.

Table 2. Datasets used for the development of *ClaimCheck*

Dataset	Source	Annotation	Similar pairs	Manually annotated pairs	Total pairs
Training	*Fact Check Explorer* and others	Similar	66%	46% *	7,760
Test	*ClaimHunter Benchmark*	Similar	45%	100%	2,246
	Prodigy Benchmark	Similar	31%	100%	7,443

*The other 54% are weak labels.

5.2. Experimental design

The experimentation with the design of the architectures took place between July and October 2022, carried out by a team of three researchers who implemented the prototyping proposals. For this purpose, a protocol was developed for data collection, and three potential experimental scenarios with their respective prototypes were designed. Each architecture presented below represents a different experiment, approaching the steps of experimentation in its own manner.

5.2.1. Classifier

The first approach involved retrieving information using standard *ElasticSearch* methods, a system that utilizes keywords and rules to search for the top K suitable candidates, and then filtering them using a classifier. Specifically, a *BERT*-like classifier was employed to detect pairs of similar sentences.

5.2.2. Semantic search + threshold

The second strategy also relied on information retrieval, but this time using a semantic search approach with *K-Nearest Neighbors* (*KNN*) implemented with *OpenSearch*. *KNN* is a type of supervised learning algorithm used for both regression and classification. Regular search with *ElasticSearch* to find matches is useful for general queries, but *KNN*-based search is more suitable for specific searches. This is because in *KNN*-based search, language features known as embeddings or 'vector representations' are extracted. The term 'embedding' is used in natural language processing to refer to the technique of representing words or phrases as numerical vectors, enabling measurement of proximity between them. Similarity is established based on how close these values are in the vector space.

Mukherjee, **Sela** and **Al-Saadoon** (2020) provide the following example: when searching for a wedding dress using a *KNN*-based search application, it produces similar results if you type "wedding dress" or "bridal gown." Thus, "summer dress" and "floral summer dress" are considered similar due to the proximity between their embeddings, unlike "summer dress" and "wedding dress."

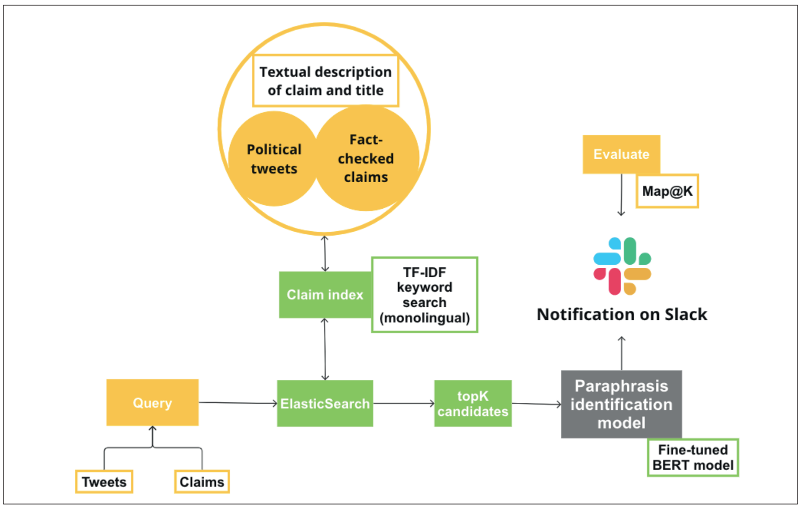


Figure 1. The architecture of the model with a classifier

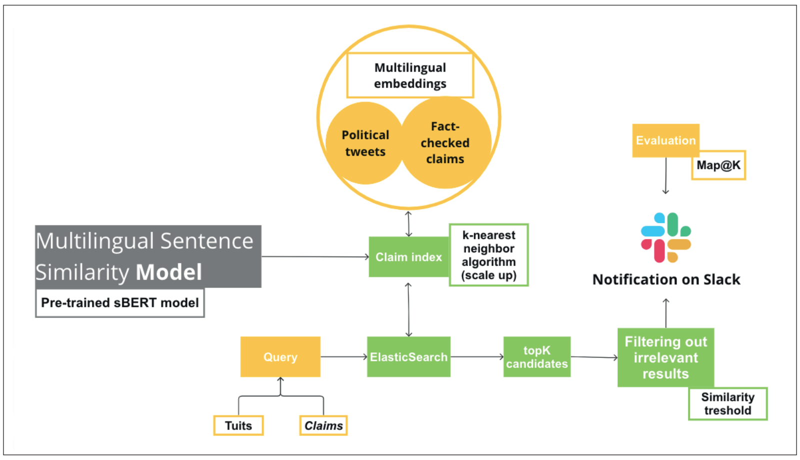


Figure 2. The architecture of the model with a semantic search system and a threshold

To retrieve the most similar or vectorially closest candidate phrases, a cosine similarity threshold is set to filter out irrelevant candidates. This concept is used to establish a critical value or cutoff point from which a decision or classification is made. The challenge lies in defining an appropriate similarity threshold for our solution. In this case, instead of using an AI-based classifier like in point 1, we simply used the similarity threshold as the classifier. With this approach, we consider the retrieval model to be good enough so that anything exceeding the threshold would be classified as a similar phrase.

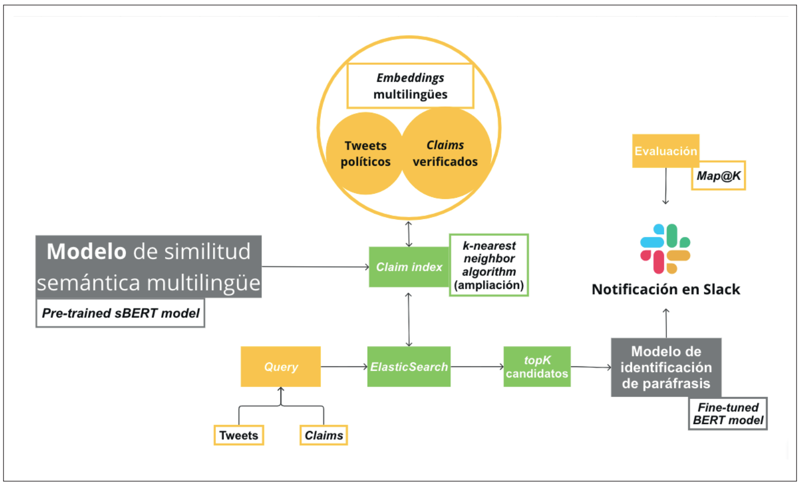


Figure 3. The architecture of the model with a semantic search system and classifier

5.2.3. Semantic search + classifier

The third proposal involved combining two artificial intelligence models:

- one for generating vector representations in the pre-trained semantic search process to retrieve candidates, and
- another model that acts as a binary classifier to identify paraphrases among the retrieved candidates, i.e., to determine whether the two phrases actually convey the same meaning or not.

Similar to the previous experiment, the first step was to train a pre-trained semantic search model, and in the second step, train a binary paraphrase classifier model using the labeled training dataset specifically created for this purpose. Both models were integrated into a prototype that allows for semantic search first and, in a subsequent phase, identifies paraphrases using the classification system.

As shown in Figure 3, the prototype is based on an embedding generator using politicians' statements on *Twitter* and the constructed fact-checking archive from previously published fact-checks (*Claim index*). From new tweets and statements (Query), a search is performed (*ElasticSearch*) on the embeddings, and the most accurate results (topK candidates) are retrieved. To filter the topK candidates, the paraphrase model is utilized, which selects the results that yield better matching (Paraphrase identification model), and sends an alert to the *Slack* messaging program.

The feedback from journalists on *Slack*, where they label whether the candidates selected by the tool are similar or not, enables an evaluation of the model in the real world using the mean average precision at different values of K (*MAP@K*).

To determine the best option, we conducted an evaluation using *MAP@K* for each of the three architectures. Specifically, this involves quantitatively assessing the percentage of topK candidates retrieved, which represents the system's precision. On an individual level, in the first and third cases, the evaluation focuses on the classifier's precision (percentage of correctly classified records), recall (percentage of similar records that are retrieved), and F1 score (a metric that combines the two previous metrics).

5.3. Results of the experiments

The third strategy (semantic search + classifier) proved to be the most successful for our use case. Regarding the similarity classification model, we conducted several tests to choose one of the pre-trained versions of *BERT*-like models offered by *Huggingface*, which is one of the most popular and widely used libraries in the field of natural language processing. Among these tests, the models with the best performance were *microsoft/Multilingual-MiniLM-L12-v2* and *xlm-roberta-base*.

Both models performed well in terms of precision and recall, but we chose the *microsoft/Multilingual-MiniLM-L12-v2* model because it is lighter, making it more suitable for use in the editorial workflow.

The *sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2* model was used in the experiments to test candidate retrieval. The effectiveness of the model was evaluated by adding an additional filter using the *microsoft/Multilingual-MiniLM-L12-H384 model*. Three commonly used metrics in information retrieval systems were used to measure performance: *MAP@K*, *Recall@K*, and *Mean Reciprocal Ranking*. The goal was to evaluate the ability of each model to retrieve the most relevant results according to the user's search criteria.

> The combination of semantic search systems and classifiers for retrieving previous fact-checks can enhance the efficiency and effectiveness of claim matching, enabling to detect previously verified claims more quickly and accurately

Table 3. Preliminary results of the trial with different pre-trained models

Models - 2checks Training	Train set	Thres-hold	Set	Similarity %	Precision	Recall	F1-score	Accuracy
microsoft/Multilingual-MiniLM-L12-H384	2checks (66% similar)	0.5	Test MRPC	66%	67.32%	98.86%	80.10%	67.35%
			Test Benchmark - ClaimHunter	45%	79.45%	71.32%	75.16%	78.05%
			Test PAWSX-es	44%	44.79%	99.89%	61.84%	44.82%
xlm-roberta-base (bce - preprocess tweets)	2checks (66% similar)	0.5	Test MRPC	66%	67.36%	99.21%	80.24%	67.52%
			Test Benchmark - ClaimHunter	45%	80.88%	69.98%	75.04%	78.32%
			Test Benchmark - ClaimHunter - only Sentences (Set6)	60%	92.14%	50.96%	65.62%	67.58%
			Test PAWSX-es	44%	44.76%	99.89%	61.82%	44.77%
xlm-roberta-base (bce - preproc-assign-labels -current ml commons version)	2checks (66% similar)	0.5	Test MRPC	66%	67.46%	98.77%	80.17%	96.18%
			Test Benchmark - ClaimHunter	45%	79.59%	67.11%	72.82%	76.67%
			Test PAWSX-es	44%	44.79%	100.00%	61.87%	44.82%
xlm-roberta-base (bce - preproc-assign-labels -local)	2checks (66% similar)	0.5	Test MRPC	66%	66.90%	98.77%	79.77%	66.71%
			Test Benchmark - ClaimHunter	45%	78.38%	66.54%	71.98%	75.87%
			Test PAWSX-es	44%	44.79%	100.00%	61.87%	44.82%
xlm-roberta-base (bce - preprocess tweets)	2checks (66% similar)	ROC optimal = 0.331424	Test MRPC	66%	67.35%	98.25%	79.91%	67.17%
			Test Benchmark - ClaimHunter	45%	75.52%	75.81%	75.67%	77.29%
			Test PAWSX-es	44%	44.82%	100.00%	61.89%	44.87%
microsoft/mdeberta-v3-base (bce - preprocess tweets)	2checks (66% similar)	0.5	Test MRPC	66%	66.98%	99.82%	80.17%	67.17%
			Test Benchmark - ClaimHunter	45%	66.44%	92.73%	77.41%	74.80%
			Test PAWSX-es	44%	44.77%	100.00%	61.85%	44.77%
microsoft/mdeberta-v3-base (bce - preprocess-tweets) adjustment of deberta specific parameters (warmup steps, epsilon, weight decay)	2checks (66% similar)	0.5	Test MRPC	66%	67.24%	99.74%	99.69%	67.52%
			Test Benchmark - ClaimHunter	45%	66.17%	92.73%	77.23%	74.53%
			Test PAWSX-es	44%	44.77%	100.00%	61.85%	44.77%
microsoft/mdeberta-v3-base (bce - preprocess-tweets) adjustment of deberta specific parameters (warmup steps, epsilon, weight decay)	2checks (66% similar)	ROC optimal = 0.998581	Test MRPC	66%	67.80%	99.21%	80.55%	68.18%
			Test Benchmark - ClaimHunter	45%	78.55%	80.88%	79.70%	80.81%
			Test PAWSX-es	44%	44.79%	100.00%	61.87%	44.82%
bert-base-multilingual-cased (bce - preprocess-tweets)	2checks (66% similar)	0.5	Test MRPC	66%	66.71%	99.82%	79.97%	66.76%
			Test Benchmark - ClaimHunter	45%	71.76%	76.29%	73.96%	74.98%
			Test PAWSX-es	44%	44.79%	100.00%	61.87%	44.82%
sentence-transformers/ paraphrase-multilingual-mpnet-base-v2 (bce - preprocess-tweets)	2checks (66% similar)	0.5	Test MRPC	66%	66.39%	99.30%	79.58%	66.12%
			Test Benchmark - ClaimHunter	45%	56.22%	82.89%	67.03%	61.98%
			Test PAWSX-es	44%	44.79%	100.00%	61.87%	44.82%
sentence-transformers/ stsb-xlm-r-multilingual (bce - preprocess-tweets)	2checks (66% similar)	0.5	Test MRPC	66%	67.53%	93.43%	78.40%	65.77%
			Test Benchmark - ClaimHunter	45%	66.97%	69.79%	68.35%	69.90%
			Test PAWSX-es	44%	44.79%	99.89%	61.84%	44.82%
sentence-transformers/ paraphrase-xlm-r-multilingual-v1 (bce - preprocess-tweets)	2checks (66% similar)	0.5	Tesr MRPC	66%	66.82%	90.46%	76.86%	63.80%
			Test Benchmark - ClaimHunter	45%	62.49%	69.12%	65.64%	66.30%
			Test PAWSX-es	44%	44.81%	99.89%	61.87%	44.87%

Tabla 4. Resultados de los ensayos del modelo de clasificación

Semantic Similarity - Classifier Model													
Modelo	**Training dataset**	**Training distribution**	**Size**	**Test dataset**	**Test size**	**Test distribution**	**Precision**		**Recall**		**F1**		**Accuracy**
							class 0	**class 1**	**class 0**	**class 1**	**class 0**	**class 1**	
xlm-roberta-base	Fact Check Explorer y otros	66% similar	7,762	Test Benchmark	2,246	45% similar	77.2%	80%	84.5%	71.3%	80.7%	75.4%	78.4%
				Prodigy Benchmark	7,443	31% similar	85.4%	67%	84.4%	68.8%	84.9%	67.9%	79.5%
microsoft/ Multilingual-MiniLM-L12-H384	Fact Check Explorer y otros	66% similar	7,762	Test Benchmark	2,246	45% similar	77%	79.4%	83.9%	71.3%	80.3%	75.2%	78%
				Prodigy Benchmark	7,443	31% similar	86.3%	64.7%	82%	71.8%	84.1%	68.1%	78.8%

Table 5. Average results of the three experiments

Evaluation of candidate retrieval (average of the three architectures)				
	K = 1	K = 3	K = 5	K = 10
MAP@K	0.7471	0.6971	0.6842	0.6864
Recall@K	0.3383	0.5615	0.6646	0.7836
MRR	0.8528			

Evaluation of candidate retrieval **(+ filtering model)** (average of the three architectures)				
	K = 1	K = 3	K = 5	K = 10
MAP@K	0.8237	0.7122	0.6837	0.6646
Recall@K	0.3638	0.5625	0.6392	0.7148
MRR	0.9169			

By comparing the results of the experiment, it was confirmed that the filtering of the model with the third architecture contributes to retrieving more suitable candidates in the top positions (Table 5). This translates into higher precision of the first retrieved candidates, as can be observed in the higher values of the *MAP@K* metric for smaller values of K.

6. Discussion and future research directions

The first finding of this research indicates that finding a balance between system precision and its ability to retrieve all similar phrases is a challenging task. As the retrieval rate increases, the precision in the recommendations provided to fact-checkers decreases, which negatively affects their confidence in the algorithm and the effort required to review the information. Given that, in general, the volume of unrelated phrases is significantly higher than that of similar phrases, the goal is to optimize the system's precision in order to avoid a high incidence of false positives.

Another issue in the process is related to the decrease in search speed as the fact-checks database grows. This occurs if the models used do not scale linearly. For this reason, the choice has been made to select faster models at the expense of heavier models, as long as the performance improvement provided by the latter is marginal, i.e., not exceeding 3%. In real-world situations, where the number of pairs of phrases to compare can reach millions, prioritizing scalability over precision is crucial.

In the experiments, it is also observed that a significant challenge is faced regarding the lack of temporal and spatial context. Therefore, in the future, it is necessary to design a strategy that allows the generation and storage of relevant metadata associated with each tweet and phrase in the system. The candidate retrieval process should consider both the semantic relevance of the phrases and the consistency of their spatial and temporal metadata. To address the temporal issue in the current system, the use of temporal windows is suggested to effectively retrieve the topK valid candidates. Although a simple solution, this strategy proves to be effective in retrieving relevant information within the proposed architecture.

The precise identification of entities (entity linking) represents another challenging task for claim matching systems in the political context. Since the model only has access to the information within the given phrase, it is unable to interpret that terms and expressions such as "Feijóo," "the president of the populares," or "the leader of the opposition" refer to the same person, as discussed in Section 4.2. While in some specific cases, the use of dictionary-based approximations may mitigate this problem, the development of more general solutions is required to address this issue.

7. Conclusions

The results of the study indicate that semantic similarity systems are a valuable tool for detecting paraphrases in political discourse and improving the effectiveness of claim matching models that contribute to automated fact-checking. Specifically, the combination of semantic search systems and classifiers for retrieving previous fact-checks can enhance the efficiency and effectiveness of claim matching, enabling fact-checking organizations to detect previously verified claims more quickly and accurately.

The general conclusion that can be drawn from the experiment results is that filtering the model using this type of architecture can be highly beneficial in improving the precision of the retrieved top candidates. The evidence provided by the experiments shows that by using this technique, more appropriate candidates from similar phrases can be retrieved and placed in the top positions, gaining agility without sacrificing precision, making it a particularly effective option for fact-checking. This result may have important practical implications in various domains.

> Semantic similarity systems are a valuable tool for detecting paraphrases in political discourse and improving the effectiveness of claim matching models that contribute to automated fact-checking

The *ClaimCheck* tool employs an approach based on the *sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2* model, which combines semantic search with a classifier to filter the results obtained through the *microsoft/Multilingual-MiniLM-L12-H384* model. The results obtained by *ClaimCheck* are superior to those of other evaluated models, as evidenced by the metrics of *MAP@K*, *Recall@K*, and *Mean Reciprocal Ranking*.

The experience of *ClaimCheck* demonstrates the most successful experimental paths with positive outcomes for the *Newtral* newsroom, while also highlighting some issues that require resolution to improve the effectiveness of the model. These include the lack of context, precise entity recognition, and the need to balance precision with the retrieval of all similar phrases.

The experimental design presented in this article can serve as a starting point for future research in the field of automated fact-checking and the use of semantic similarity systems in identifying false claims in political discourse. This extends beyond extracting previous fact-checks and enabling their reuse more quickly without duplicating the effort but also extends to real-time fact-checking or creating alert systems against misinformation campaigns. Furthermore, this study suggests potential directions for future research to address the challenges present in the proposed model and improve its adoption by specialized fact-checking organizations.

8. References

Adair, Bill (2021). "The lessons of Squash, Duke's automated fact-checking platform". *Poynter*, 16 June.
https://www.poynter.org/fact-checking/2021/the-lessons-of-squash-the-first-automated-fact-checking-platform

Adair, Bill; **Li, Chengkai**; **Yang, Jun**; **Yu, Cong** (2018). *Automated pop-up fact-checking: challenges & progress*.
https://ranger.uta.edu/~cli/pubs/2019/popupfactcheck-cj19-adair.pdf

Agadjanian, Alexander; **Bakhru, Nikita**; **Chi, Victoria**; **Greenberg, Devyn**; **Hollander, Byrne**; **Hurt, Alexander**; **Kind, Joseph**; **Lu, Ray**; **Ma, Annie**; **Nyhan, Brendan**; **Pham, Daniel**; **Qian, Michael**; **Tan, Mackinley**; **Wang, Clara**; **Wasdahl, Alexander**; **Woodruff, Alexandra** (2019). "Counting the Pinocchios: the effect of summary fact-checking data on perceived accuracy and favorability of politicians". *Research & politics*, v. 6, n. 3.
https://doi.org/10.1177/2053168019870351

Arslan, Fatma (2021). *Modeling factual claims with semantic frames: definitions, datasets, tools, and fact-checking applications*. Doctoral dissertation. The University of Texas at Arlington.
https://rc.library.uta.edu/uta-ir/bitstream/handle/10106/30765/ARSLAN-DISSERTATION-2021.pdf

Babakar, Mevan; **Moy, Will** (2016). *The state of automated factchecking. How to make factchecking dramatically more effective with technology we have now*. Full Fact.
https://fullfact.org/media/uploads/full_fact-the_state_of_automated_factchecking_aug_2016.pdf

Baker, Collin F.; **Fillmore, Charles J.**; **Lowe, John B.** (1998). "The Berkeley FrameNet project". In: *Proceedings of the joint conference of the international conference on computational linguistics and the Association for Computational Linguistics (Coling-ACL)*, pp. 86-90.
https://aclanthology.org/C98-1013.pdf

Beltrán, Javier; **Míguez, Rubén**; **Larraz, Irene** (2019). "ClaimHunter: an unattended tool for automated claim detection on *Twitter*". *KnOD@WWW. CEUR workshop proceedings*, v. 2877, n. 3.
https://ceur-ws.org/Vol-2877/paper3.pdf

Corney, David (2021a). "How does automated fact checking work?". *Full Fact*, 5 July.
https://fullfact.org/blog/2021/jul/how-does-automated-fact-checking-work

Corney, David (2021b). "Towards a common definition of claim matching". *Full Fact*, 5 October.
https://fullfact.org/blog/2021/oct/towards-common-definition-claim-matching

Dolan, William B.; **Brockett, Chris** (2005). "Automatically constructing a corpus of sentential paraphrases". In: *Proceedings of the third international workshop on paraphrasing* (*IWP2005*), pp. 9-16.
https://aclanthology.org/I05-5002.pdf

Floodpage, Sebastien (2021). "How fact checkers and *Google.org* are fighting misinformation". *Google*, 31 March.
https://blog.google/outreach-initiatives/google-org/fullfact-and-google-fight-misinformation

Graves, Lucas (2018). *Understanding the promise and limits of automated fact-checking*. Reuters Institute for the Study of Journalism. Factsheets.
https://ora.ox.ac.uk/objects/uuid:f321ff43-05f0-4430-b978-f5f517b73b9b

Hassan, Aumyo; **Barber, Sarah J.** (2021). "The effects of repetition frequency on the illusory truth effect". *Cognitive research: principles and implications*, v. 6, n. 38.
https://doi.org/10.1186/s41235-021-00301-5

Hassan, Naeemul; **Adair, Bill**; **Hamilton, James T.**; **Li, Chengkai**; **Tremayne, Mark**; **Yang, Jun**; **Yu, Cong** (2015). "The quest to automate fact-checking". In: *Proceedings of the 2015 computation + journalism symposium*. Columbia University.
http://cj2015.brown.columbia.edu/papers/automate-fact-checking.pdf

Hassan, Naeemul; **Arslan, Fatma**; **Li, Chengkai**; **Tremayne, Mark** (2017). "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* (*KDD '17*). New York: Association for Computing Machinery, pp. 1803-1812.
https://doi.org/10.1145/3097983.3098131

Hövelmeyer, Alica; **Boland, Katarina**; **Dietze, Stefan** (2022). "SimBa at CheckThat! 2022: lexical and semantic similarity based detection of verified claims in an unsupervised and supervised way". In: *CLEF 2022: Conference and labs of the evaluation forum*, 5-8 September, Bolonia, Italia.
https://ceur-ws.org/Vol-3180/paper-40.pdf

Jiang, Ye; **Song, Xingyi**; **Scarton, Carolina**; **Aker, Ahmet**; **Bontcheva, Kalina** (2021). "Categorising fine-to-coarse grained misinformation: an empirical study of Covid-19 Infodemic". *Arxiv*.
https://doi.org/10.48550/arXiv.2106.11702

Kazemi, Ashkan; **Garimella, Kiran**; **Gaffney, Devin**; **Hale, Scott A.** (2021). "Claim matching beyond English to scale global fact-checking". In: *Proceedings of the 59th Annual meeting of the Association for Computational Linguistics and the 11th International joint conference on natural language processing*. Association for Computational Linguistics, pp. 4504-4517.
https://doi.org/10.18653/v1/2021.acl-long.347

Kazemi, Ashkan; **Li, Zehua**; **Pérez-Rosas, Verónica**; **Hale, Scott A.**; **Mihalcea, Rada** (2022). "Matching tweets with applicable fact-checks across languages". *Arxiv*.
https://doi.org/10.48550/arXiv.2202.07094

Kessler, Glenn; **Fox, Joe** (2021). "The false claims that Trump keeps repeating". *The Washington Post*, 20 January.
https://www.washingtonpost.com/graphics/politics/fact-checker-most-repeated-disinformation

Lan, Zhenzhong; **Chen, Mingda**; **Goodman, Sebastian**; **Gimpel, Kevin**; **Sharma, Piyush**; **Soricut, Radu** (2020). "ALBERT: a lite Bert for self-supervised learning of language representations". In: *Conference paper at International conference on learning representations (ICLR)*. *Arxiv*.
https://doi.org/10.48550/arXiv.1909.11942

Lim, Chloe (2018). "Checking how fact-checkers check". *Research & politics*, v. 5, n. 3.
https://doi.org/10.1177/2053168018786848

Mansour, Watheq; **Elsayed, Tamer**; **Al-Ali, Abdulaziz** (2022). "Did I see it before? Detecting previously-checked claims over *Twitter*". *Lecture notes in computer science*, pp. 367-381.
https://doi.org/10.1007/978-3-030-99736-6_25

Martín, Alejandro; **Huertas-Tato, Javier**; **Huertas-García, Álvaro**; **Villar-Rodríguez, Guillermo**; **Camacho, David** (2021). "FacTeR-check: semi-automated fact-checking through semantic similarity and natural language inference". *Arxiv*.
https://doi.org/10.48550/arXiv.2110.14532

Mukherjee, Amit; **Sela, Eitan**; **Al-Saadoon, Laith** (2020). "Building an NLU-powered search application with *Amazon SageMaker* and the *Amazon opensearch* service KNN feature". *Amazon SageMaker, artificial intelligence*, 26 October.
https://aws.amazon.com/es/blogs/machine-learning/building-an-nlu-powered-search-application-with-amazon-sagemaker-and-the-amazon-es-knn-feature

Murray, Samuel; **Stanley, Matthew**; **McPhetres, Jon**; **Pennycook, Gordon**; **Seli, Paul** (2020). "'I've said it before and I will say it again…': repeating statements made by Donald Trump increases perceived truthfulness for individuals across the political spectrum". *PsyArXiv preprints*, 15 January.
https://doi.org/10.31234/osf.io/9evzc

Nakov, Preslav; **Corney, David**; **Hasanain, Maram**; **Alam, Firoj**; **Elsayed, Tamer**; **Barrón-Cedeño, Alberto**; **Papotti, Paolo**; **Shaar, Shaden**; **Da-San-Martino, Giovanni** (2021). "Automated fact-checking for assisting human fact-checkers". *International joint conference on artificial intelligence*. *Arxiv*.
https://doi.org/10.48550/arXiv.2103.07769

Nakov, Preslav; **Da-San-Martino, Giovanni**; **Alam, Firoj**; **Shaar, Shaden**; **Mubarak, Hamdy**; **Babulkov, Nikolay** (2022). “Overview of the CLEF-2022 CheckThat! Lab task 2 on detecting previously fact-checked claims”. In: *CLEF 2022: conference and labs of the evaluation forum*, 5-8 septiembre, Bolonia, Italia.
https://ceur-ws.org/Vol-3180/paper-29.pdf

Nguyen, Vincent; **Karimi, Sarvnaz**; **Xing, Zhenchang** (2021). “Combining shallow and deep representations for text-pair classification”. In: *Proceedings of the 19th Annual workshop of the Australasian Language Technology Association*, pp. 68-78.
https://aclanthology.org/2021.alta-1.7.pdf

Phillips, Whitney (2018). *The oxygen of amplification. Better pratices for reporting on extremists, antagonists, and manipulators online*. Data & Society Research Institute.
https://datasociety.net/wp-content/uploads/2018/05/FULLREPORT_Oxygen_of_Amplification_DS.pdf

Porter, Ethan; **Wood, Thomas J.** (2021). “The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom”. *Proceedings of the National Academy of Sciences of the United States of America*, v. 118, n. 37.
https://doi.org/10.1073/pnas.2104235118

Real, Andrea (2021). “Casado mezcla diferentes estadísticas de empleo para asegurar que hay 4 millones de parados, pero es falso”. *Newtral*, 6 octubre.
https://www.newtral.es/parados-espana-casado-pp-factcheck/20211007

Reimers, Nils; **Gurevych, Iryna** (2019). “Sentence-bert: sentence embeddings using siamese bert-networks”. 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*). Hong Kong, November, pp. 3982-3992.
https://doi.org/10.18653/v1/D19-1410

Shaar, Shaden; **Alam, Firoj**; **Da-San-Martino, Giovanni**; **Nakov, Preslav** (2021a). “The role of context in detecting previously fact-checked claims”. *Arxiv*.
https://doi.org/10.48550/arXiv.2104.07423

Shaar, Shaden; **Babulkov, Nikolay**; **Da-San-Martino, Giovanni**; **Nakov, Preslav** (2020). “That is a known lie: detecting previously fact-checked claims”. In: *Proceedings of the 58th Annual meeting of the Association for Computational Linguistics*, pp. 3607-3618.
https://doi.org/10.18653/v1/2020.acl-main.332

Shaar, Shaden; **Haouari, Fatima**; **Mansour, Watheq**; **Hasanain, Maram**; **Babulkov, Nikolay**; **Alam, Firoj**; **Da-San-Martino, Giovanni**; **Elsayed, Tamer**; **Nakov, Preslav** (2021b). “Overview of the CLEF-2021 CheckThat! Lab task 2 on detecting previously fact-checked claims in tweets and political debates”. In: *CLEF 2021: Conference and labs of the evaluation forum*, 21-24 September, Bucharest, Romania.
https://ceur-ws.org/Vol-2936/paper-29.pdf

Sheng, Qiang; **Cao, Juan**; **Zhang, Xueyao**; **Li, Xirong**; **Zhong, Lei** (2021). “Article reranking by memory-enhanced key sentence matching for detecting previously fact-checked claims”. In: *Proceedings of the 59th Annual meeting of the Association for Computational Linguistics and the 11th International joint conference on natural language processing* (volume 1, Long papers).
https://doi.org/10.18653/v1/2021.acl-long.425

Sippitt, Amy (2020). *What is the impact of fact checkers’ work on public figures, institutions and the media?*. Africa Check, Chequeado and Full Fact.
https://fullfact.org/media/uploads/impact-fact-checkers-public-figures-media.pdf

Stanford Institute for Human-Centered Artificial Intelligence (2023). *Artificial intelligence index*. Stanford University.
https://aiindex.stanford.edu/report

The Washington Post (2018). “Meet the bottomless Pinocchio | Fact Checker”. [Video]. *YouTube*, 10 December.
https://www.youtube.com/watch?v=zoS1sVZRfUU

Thorne, James; **Vlachos, Andreas** (2018). “Automated fact checking: task formulations, methods and future directions”. *Arxiv*.
https://doi.org/10.48550/arXiv.1806.07687

Wardle, Claire (2018). “Lessons for reporting in an age of disinformation”. *Medium*, 28 December.
https://medium.com/1st-draft/5-lessons-for-reporting-in-an-age-of-disinformation-9d98f0441722

Zeng, Xia; **Abumansour, Amani S.**; **Zubiaga, Arkaitz** (2021). “Automated fact-checking: a survey”. *Language and linguistics compass*, v. 15, n. 10.
https://doi.org/10.1111/lnc3.12438