Natural Language Processing Journal 7 (2024) 100066

Contents lists available at ScienceDirect

Natural Language Processing Journal

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

Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research

Rrubaa Panchendrarajan *, Arkaitz Zubiaga

School of Electronic Engineering and Computer Science, Queen Mary University of London, 327 Mile End Rd, Bethnal Green, London, E1 4NS, United Kingdom

ARTICLE INFO

Keywords:
Multilingual claim detection
Claim prioritization
Claim matching
Claim clustering
Fact-checking

ABSTRACT

Automated fact-checking has drawn considerable attention over the past few decades due to the increase in the diffusion of misinformation on online platforms. This is often carried out as a sequence of tasks comprising (i) the detection of sentences circulating in online platforms which constitute claims needing verification, followed by (ii) the verification process of those claims. This survey focuses on the former, by discussing existing efforts towards detecting claims needing fact-checking, with a particular focus on multilingual data and methods. This is a challenging and fertile direction where existing methods are yet far from matching human performance due to the profoundly challenging nature of the issue. Especially, the dissemination of information across multiple social platforms, articulated in multiple languages and modalities demands more generalized solutions for combating misinformation. Focusing on multilingual misinformation, we present a comprehensive survey of existing multilingual claim detection research. We present state-of-the-art multilingual claim detection research categorized into three key factors of the problem, verifiability, priority, and similarity. Further, we present a detailed overview of the existing multilingual datasets along with the challenges and suggest possible future advancements.

Contents

* Corresponding author.
E-mail addresses: r.panchendrarajan@qmul.ac.uk (R. Panchendrarajan), a.zubiaga@qmul.ac.uk (A. Zubiaga).
URL: https://www.zubiaga.org (A. Zubiaga).

https://doi.org/10.1016/j.nlp.2024.100066
Received 19 January 2024; Received in revised form 8 March 2024; Accepted 17 March 2024

1. Introduction

Misinformation poses a significant threat to society, a threat that has escalated with the advent and widespread use of social media platforms. This is demanding an additional layer of verifying online information to ensure the integrity and validity of the information that people read online. However, verifying the content circulating on online platforms is a time-consuming task that if done manually can only encompass a small portion of the available information, which demands the development of methods to enable automated fact-checking, a process that starts off by identifying information needing verification and ends up by verifying whether a claim is supported or refuted by a reputable piece of evidence, or occasions with a verdict that there is no sufficient evidence to determine its accuracy. This is often carried out as a series of steps involving (i) the identification of claims to be checked, (ii) prioritization of important claims to deal with, (iii) gathering evidence associated with those claims, and (iv) concluding with the final verdict by checking the claim against the associated evidence. While there are several dedicated organizations such as PoliFact,[1] Full Fact[2] and Newtral[3] established in recent years, research in this direction is experiencing a substantial increase in both fact-checking organizations and academic research due to the growing pressure of dealing with online misinformation.

There are some recent surveys presenting overviews of existing research on the fact-checking pipeline and its underlying components (Zeng et al., 2021; Guo et al., 2022; Das et al., 2023). However, their holistic focus on the entire fact-checking process impedes them from providing a detailed study of each component of the pipeline. In addition to these surveys, some studies have focused on reviewing each a particular aspect of the fact-checking problem. For example, Hardalov et al. (2022) highlights the role of *stance detection* in automated misinformation detection, and Kotonya and Toni (2020a,b) focus on the explainability aspect of automated fact-checking. Different from these survey papers, we present a comprehensive study on the claim detection component of the automated fact-checking pipeline, with a specific focus on multilingual research.

This survey presents a comprehensive review of the state-of-the-art techniques used for a wide range of claim detection tasks. Fig. 1 depicts the claim detection tasks discussed in this paper. Given that the claim detection task can have different objectives and hence be formulated in different ways, we discuss the different claim detection subtasks by grouping them into the following three categories:

- **Verifiability**: Identifying claims that are verifiable. We further discuss the definition of verifiable clams, and the tasks associated with it in Section 2.
- **Priority**: Not all the verifiable claims are worthy of fact-checking, and prioritization of claims plays a vital role in effective fact-checking. We further discuss the factors determining the priority of claims, and the tasks associated with it in Section 2.
- **Similarity**: A massive amount of unverified online content often comprises repeated information. Hence, identifying similar claims is important for avoiding the repetition of fact-checking similar claims. We introduce the similarity identification tasks in Section 2.

The rest of the survey is structured as follows. Section 2 introduces the fact-checking pipeline, different definitions used in the literature to define a claim, and the multilingual view of the claim detection problem. Sections 3 and 4 present existing research on identifying *verifiability* and *priority* of claims. Similarity identification of the claims is discussed in Sections 5 and 6. We outline the challenges associated with claim detection in Section 7, followed by the conclusions in Section 8.

2. Background: Automated fact-checking pipeline

An automated fact-checking pipeline is typically composed of five major components (Das et al., 2023) as depicted in Fig. 2. The process begins with detecting, out of a collection of input sentences, verifiable factual statements referred to as *claims*. This component is commonly applied to social platforms and online resources such as news articles to identify statements that require verification, hence getting rid of the remainder of the sentences which do not need to go through the rest of the fact-checking pipeline for not needing verification. Identifying the verifiability of claims is often carried out as a binary decision indicating whether a statement is verifiable or not. However, determining the claim type indicating the type of factual information presented in the claim can also be performed as a fine-grained analysis, as for example (Konstantinovskiy et al., 2021) introduced a taxonomy of types of claims.

Once the claims are extracted, they go through a prioritization process to estimate the worthiness of the claim to be verified. The criteria used to estimate the worthiness may vary according to the topic or domain of the claim and the user groups (i.e. audience) who are interested in the veracity of the claim. Some of the popular criteria used in the literature are the virality of the claim, the interest of the public in the veracity of the claim, the impact that the claim could make on society, and its timeliness (Das et al., 2023; Micallef et al., 2022). While the determination of priority is often modeled as the estimation of the *check-worthiness* of a claim, similar prioritization tasks including estimation of *attention-worthiness* and *harmfulness* are also carried out in the literature (Nakov et al., 2022) for claim prioritization.

Following the prioritization, relevant evidence is retrieved which could ideally help support or refute the prioritized claims, and finally, the verdict of the claim indicating whether the fact discussed in the

[1] https://www.politifact.com/.
[2] https://fullfact.org/.
[3] https://www.newtral.es/.

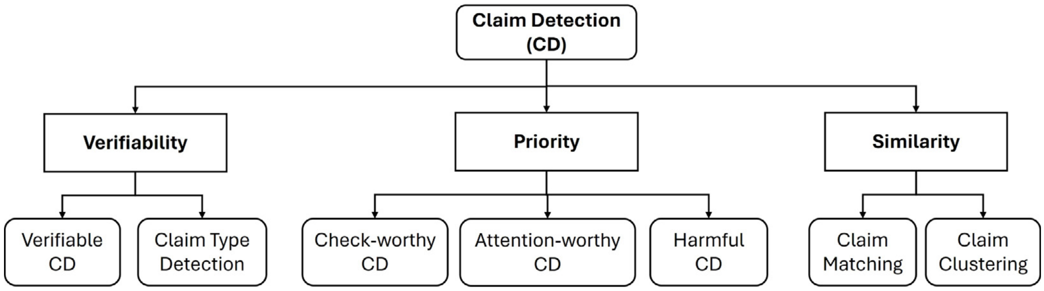

Fig. 1. Claim detection tasks.

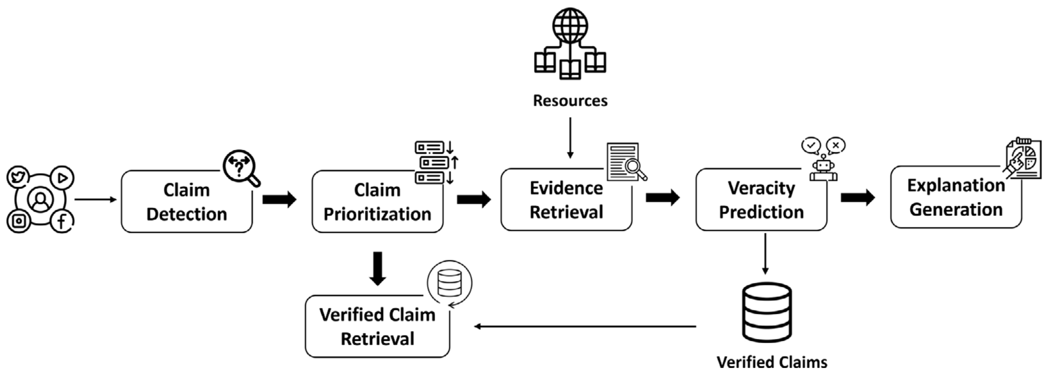

Fig. 2. Fact-checking Pipeline.

statement is supported or refuted[4] is predicted with respect to the evidence retrieved (Guo et al., 2022), while also on alternative occasions predicting that there is 'not enough information' to determine the support or refutal. Often the evidence retrieval and veracity prediction tasks are tackled jointly in the literature and referred to as the *fact verification* process (Guo et al., 2022). The latest addition to the automated fact-checking pipeline is the explanation generation (Kotonya and Toni, 2020a), where researchers aim to automatically generate a reason for the verdict predicted to boost the interpretability and explainability of the system.

Apart from these five major components, as a parallel component, researchers have focused on retrieving, from the database of previously fact-checked claims, existing claims that resemble or relate to newly retrieved claims. This task is referred to as *verified claim retrieval* or *claim matching* in the literature. Matching new claims with previously fact-checked claims can then help avoid a repeated attempt of claim verification as the verdict was already done in the past. While this can avoid the substantial time involved in processing an unverified claim through the remaining components of the pipeline, the impact and spread of the claim can also be minimized with timely verdicts.

2.1. Definition of claim

There are different ways of defining a claim depending on the objective of different components of the automated fact-checking pipeline. This includes the following three key stages of a claim as depicted in Fig. 3:

- *Claim* - A claim is defined as a statement containing a purported fact about the real world (Das et al., 2023). For example, the statement *"team X is the best team in FIFA 2023"* is an opinion, whereas, the statement *"United States is one of the host countries of the next FIFA World Cup"* is a claim that can be checked against an objective piece of evidence. Zeng et al. (2021) give a slightly different definition for a claim as a "factual statement under investigation".
- *Verifiable Claim* - A verifiable claim is defined as a "factual statement that can be checked" by Micallef et al. (2022), and a similar definition of "assertion about the world that is checkable" is given by Konstantinovskiy et al. (2021). Claims about personal experience are neither an assertion about the world nor can be verified. Therefore they are not considered as *verifiable claims*. While the definition of a *verifiable claim* enforces the possibility of determining the veracity of a claim, this aspect predominantly relies on the availability of evidence. However, the availability of evidence cannot be determined until the execution of the evidence retrieval task. Therefore, verifiable claim detection tasks often ignore the availability of evidence and aim at identifying claims about the real world.
- *Check-worthy Claim* - Verifying all the assertions about the real world is impractical, and hence it demands a prioritization process. This objective is generally handled as an estimation of check-worthiness of a verifiable claim in the literature. Due to the subjective nature of this task, it is hard to define the check-worthiness of a verifiable claim and it may vary according to the topic discussed in the claim and the user group who are interested in the claim. Further, the worthiness may vary over time (Guo et al., 2022), as the interest of a claim may fade or increase as stories develop, which makes it more challenging. Some of the common factors used to determine the check-worthiness of verifiable claims include the popularity of the claim, amount of public interest in the verdict of the claim, impact of the verdict, and timeliness of the verdict (Das et al., 2023; Micallef et al., 2022). In addition to the check-worthiness, the following criteria are used in the literature for prioritizing claims:
  - *Attention-worthy Claims*: A verifiable claim that should get the attention of the policymakers and government entities (Nakov et al., 2022; Shaar et al., 2021a).

[4] Occasionally some researchers may choose to output 'true' or 'false' predictions, although 'supported' and 'refuted' are the more widely used alternatives indicating how the retrieved evidence aligns with the claim, rather than concluding the veracity of the claim.

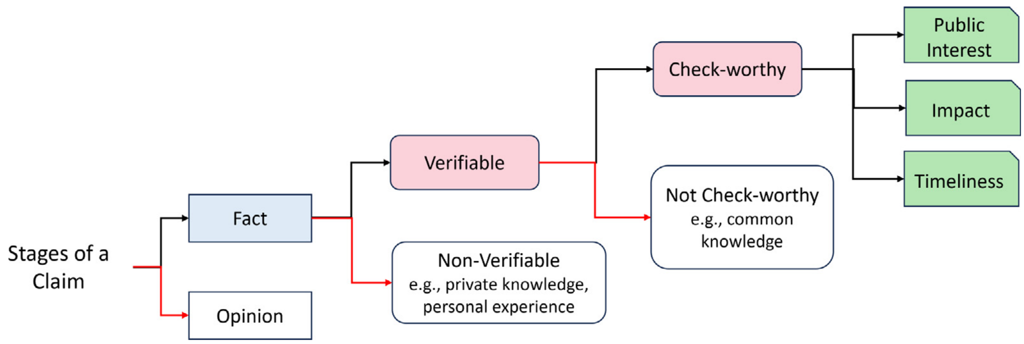

Fig. 3. Stages of a claim. Black arrows in the figure indicate claims and red arrows indicate non-claims.

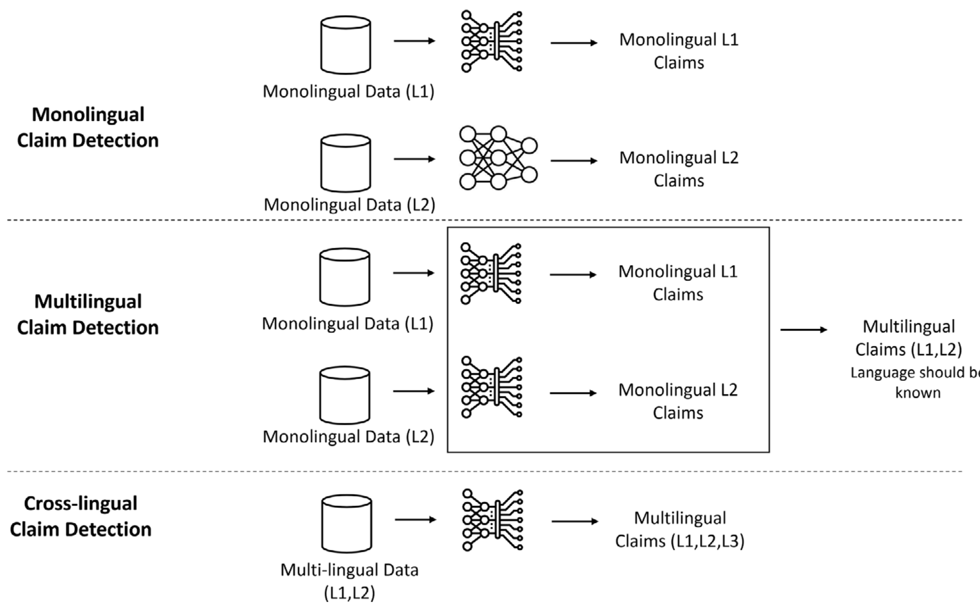

Fig. 4. Monolingual vs. Multilingual vs. Cross-lingual claim detection.

– *Harmful Claims*: A verifiable claim that is harmful to society (Nakov et al., 2022; Shaar et al., 2021a).
– *Interesting to the general public*: A verifiable claim that may have an impact on society or attract interest from the general public (Shaar et al., 2021a).

Apart from these three key stages of a claim, *rumor* detection is also typically handled as a similar problem in the literature, although not strictly formulated as a fact-checking task, where a statement can be categorized as a rumor or non-rumor. A rumor is defined in the literature (Zubiaga et al., 2018) as "an unverified claim circulating on social platforms" (Guo et al., 2022) and as "unverified and instrumentally relevant information statements in circulation" (DiFonzo and Bordia, 2007).

2.2. Cross-lingual claim detection

We define the following terms used in the literature to refer to the multilingual aspect of the claim detection problem. Fig. 4 summarizes these definitions.

- *Monolingual* Claim Detection - Achieved via developing a language-specific model for identifying the claims of a specific language *L1*, and training the model on the data from the same language. Repeating the process for language *L2* remains as *monolingual* claim detection as the introduction of new language *L2* requires developing a language-specific model.
- *Multilingual* Claim Detection - Achieved via developing a language -independent claim detection model, and training it on the data from language *L1*. Applying the model for other languages requires retraining the model for other languages on language-specific training data. Therefore, this approach demands training data from each language the fact-checking pipeline deals with and the language of the input statement to be known to determine the corresponding *monolingual* model. Further, the *monolingual* models do not generalize their knowledge beyond the language it has seen in the training data.
- *Cross-lingual* Claim Detection - Achieved via developing a single model for identifying the claims of any language or a broader set of languages than the one(s) seen during training. The *cross-lingual* model is trained using training data composed of one or more languages. Further, it generalizes its knowledge beyond the languages it has seen in the training data, to identify claims written in new languages.

2.3. Transformer models

The transformer architecture (Vaswani et al., 2017) proposed in 2017 was a breakthrough in the evolution of neural architectures. The model was composed of layers of encoders and decoders with multi-headed attention, enabling them to be very effective and powerful in sequence transduction tasks. Later, the architecture was explored by researchers to introduce several influential models, resulting in a family of transformer models. Some of the notable models include BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), GPT (Generative Pre-trained Transformer) (Radford and Narasimhan, 2018), BART (Bidirectional Auto-Regressive Transformers) (Lewis et al., 2020), and their variations.

Due to the powerful nature of these models in understanding the language and training data, fine-tuning them with limited training data showed a significant performance increase in several Natural Language Processing (NLP) tasks (Pfeiffer et al., 2020). This generalization capability of the transformer models was further extended to multi-lingual settings, by training them on massive amounts of multi-lingual data. mBERT, mGPT, mBART, and XLM-R (Kalyan et al., 2021) are some of the popular multi-lingual pre-trained transformer models, which offer state-of-the-art performance in cross-lingual settings in various NLP tasks.

3. Verifiable claim detection

This section summarizes the existing verifiable claim detection research in the literature. This task is often treated as a binary classification problem to identify whether a statement is a verifiable claim or not. However, the problem is treated as a multi-class classification too by either adding an uncertainty label (Kazemi et al., 2021a) or identifying fine-grained verifiable claim types (Konstantinovskiy et al., 2021).

3.1. Datasets

A key challenge in cross-lingual setting is the availability of multilingual training data. Some notable datasets released for multilingual verifiable claim detection were the NLP4IF 2021 shared task data (Shaar et al., 2021a) and CheckThat 2022 data (Nakov et al., 2022). Both datasets contain tweets related to the COVID-19 pandemic. NLP4IF 2021 includes tweets written in English, Arabic, and Bulgarian languages labeled as verifiable claims or not. The annotation also includes six other misinformation labels including the label indicating whether the tweet requires fact-checking or not. In addition to these three languages, the CheckThat 2022 dataset includes Dutch and Turkish tweets labeled for the verifiable claim detection and claim prioritization tasks. This dataset was expanded with more languages and data via several stages (Alam et al., 2021; Shaar et al., 2021b; Nakov et al., 2021), and the final version was released in 2022 (Nakov et al., 2022). Following these, several other multilingual verifiable claim detection datasets were released focusing on topics including COVID-19 (Kazemi et al., 2021a) and politics (Kazemi et al., 2021a; Dutta et al., 2022). Table 1 summarizes the existing multilingual claim detection datasets. While there are more datasets of monolingual nature, only the four datasets listed in the table include multilingual data, with dataset sizes varying from 1.3K to 6K and including a diverse set of languages.

3.2. Cross-lingual claim detection

One of the pioneering works in cross-lingual claim detection was carried out by the authors of the NLP4IF dataset (Alam et al., 2020). They experimented with both cross-lingual and mono-lingual settings using multi-lingual models including mBERT (multilingual BERT) (Devlin et al., 2019) and XLM-r (Kalyan et al., 2021) and language-specific BERT models. The authors observed similar or improved performance in cross-lingual settings when the models were trained using the combined dataset (dataset including all three languages under study). Further, the authors experimented with the impact of tweet-specific features on the performance. This includes various features of a tweet and its author details such as verified status, number of friends, and followers. Additionally, the botness of the tweet indicating the chances of the author being a bot is also included as a feature for the model. Compared to other features, the authors observed an increase in performance when the botness feature was injected into the classification model. A similar attempt was made by Uyangodage et al. (2021) in experimenting with both cross-lingual and mono-lingual settings using BERT variations. The authors fine-tuned publicly available language-specific BERT models and mBERT (Devlin et al., 2019) using the language-specific and combined training data respectively. They observed similar or improved performance in cross-lingual settings on the NLP4IF dataset.

Panda and Levitan (2021) performed a similar analysis of utilizing mBERT (Devlin et al., 2019) for the classification task in the NLP4IF 2021 dataset, and the authors reported that mBERT can achieve an impressive score in identifying misinformation labels even without fine-tuning on language-specific training data. A recent study by Agrestia et al. (2022) showed, that fine-tuning GPT-3 model (Radford and Narasimhan, 2018) using English data only gives competitive performance to the BERT models trained on language-specific training data for both verifiable claim detection and claim prioritization tasks.

3.3. Multilingual claim detection

A straightforward approach to implementing the multi-lingual setting for verifiable claim detection tasks is fine-tuning the multi-lingual pre-trained models on language-specific training data. Following this approach (Hüsünbeyi et al., 2022) used XLM-R (Kalyan et al., 2021) for training language-specific models in CheckThat 2022 English and Turkish data for both the verifiable claim detection and claim prioritization tasks. The authors compared the performance with monolingual pre-trained models and observed that multilingual models achieve competitive performance when trained on language-specific training data. Savchev (2022) performed a similar analysis using the XLM-R multilingual model. The author further used back translation, translating a tweet to a target language, and then translating back from the target language to the original language as the data augmentation technique. They observed an increase in overall performance with the incorporation of the data augmentation technique. The latest study by Eyuboglu et al. (2023) applied a wide range of pre-trained models and observed that, generally, BERT variants are very powerful in classifying both the verifiability and priority level of claims.

3.4. Monolingual claim detection

One of the earliest works in the direction of monolingual verifiable claim detection was carried out by Prabhakar et al. (2020). The authors used the verifiable claims from the dataset FEVER (Thorne et al., 2018) and collected the non-claims from Wikipedia articles with the assumption that sentences without any citation are non-claims according to Wikipedia's verifiability policy.[5] This resulted in a massive amount of claim and non-claim samples in English. The authors fine-tuned BERT (Devlin et al., 2019) and DistilBERT (Sanh et al., 2019) models to identify the claims and observed that the fine-tuned BERT model achieved an F1-score of 98%. Similarly, Alam et al. (2021) released the initial version of the CheckThat dataset and the authors experimented with different sets of BERT variations for each language. Among the variations of BERT models, they observed that XLM-r (Kalyan et al., 2021) outperformed other models in several languages.

Similar to the cross-lingual setting, Suri and Dudeja (2022) used the BERT model with a data augmentation technique to detect verifiability and priority of claims in the English language in CheckThat 2022 data. The authors translated the training data from other languages to English to increase the training data size. Further, they injected both tweet-specific and author-specific features as additional input to the model for improving the classification performance. Apart from these studies, language-specific pre-trained BERT models were often used as

[5] https://en.wikipedia.org/wiki/Wikipedia:Verifiability.

Table 1
Multilingual verifiable claim detection datasets.

Dataset	Objective	Label	Language	Topic	Source	Size	Evaluation
NLP4IF 2021 (Shaar et al., 2021a)	Verifiable Claims	Yes No	English Arabic Bulgarian	Covid-19	Twitter	1.3K–4K	Precision Recall F1 Score
Kazemi et al. (2021a)	Claim-like Statements	Yes No Probably	English Hindi Bengali Malayalam Tamil	Covid-19 Politics	WhatsApp	5K	Accuracy Precision Recall F1 Score
Dutta et al. (2022)	Verifiable Claims	Yes No	English Hindi Bengali Code-mixed	Politics	Twitter	600–1.4K	Precision Recall F1 Score
CheckThat 2022 (Nakov et al., 2022)	Verifiable Claims	Yes No	English Arabic Bulgarian Dutch Turkish	Covid-19	Twitter	4K–6K	Accuracy

the effective monolingual solution (Henia et al., 2021; Hussein et al., 2021).

Different from these approaches, Konstantinovskiy et al. (2021) performed a fine-grained analysis of verifiable claims by classifying a sentence into non-claim or six sub-categories of a claim. The authors annotated around 6300 sentences from subtitles of TV political debates and trained various traditional machine learning models with a wide range of features. Notable textual features include TF-IDF, Part-of-speech (POS) tags, Named entity recognition (NER), and word embeddings. The authors observed that the logistic regression classifier (LaValley, 2008) obtained the highest F1 score in classifying the sentences as a claim or non-claim, and injecting POS and NER information did not improve the performance of the optimal classifier. The proposed solution was tested in a live feed of transcripts from TV shows. Similar claim type identification research has been carried out in the literature using rule-based approaches (Rony et al., 2020).

The most recent attention on monolingual claim detection has been given to developing domain-specific solutions. Woloszyn et al. (2021) focused on identifying *green claim*, a claim discussing an issue related to the environment. The authors compared three pre-trained models, RoBERTa (Liu et al., 2019), BERTweet (Nguyen et al., 2020), and Flair (Akbik et al., 2018), and observed that generally, RoBERTa outperformed the other two models in the green claim detection task. Smeros et al. (2021) extracted scientific claims by introducing three variants of BERT, SciBERT, NewsBERT, and SciNewsBERT fine-tuned using scientific articles and news headlines. Recent research has also shown that the difficulty of identifying check-worthy claims varies across domains. Abumansour and Zubiaga (2023) showed that a model trained on check-worthy and non-check-worthy claims pertaining to a set of topics can struggle to perform the claim detection task on a new, unseen topic, a limitation that can be mitigated through the use of data augmentation strategies.

3.5. Evaluation

As previously mentioned, identifying a verifiable claim is consistently approached as a classification task, and typically as a binary classification task. The following evaluation metrics are used to evaluate the classification models in the literature.

- Accuracy (Nakov et al., 2022) - Proportion of correctly classified data instances among the total number of data instances
- Precision (Shaar et al., 2021a) - Indicates the proportion of correctly classified positive data instances over the total number of data instances classified as positive samples, and computed as follows:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{1}$$

- Recall (Shaar et al., 2021a) - Indicates the proportion of correctly classified positive data instances over the total number of positive data instances, and computed as follows:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{2}$$

- F1 Score (Alam et al., 2021; Shaar et al., 2021a) - F1 score is the harmonic mean of precision and recall. The score is computed as follows:

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

Table 1 shows the evaluation metrics used in the existing verifiable claim detection datasets. It can be observed that precision, recall, and F1 score are widely used for the evaluation with the exception of the CheckThat 2022 dataset (Nakov et al., 2022). This is due to the fairly balanced nature of the CheckThat 2022 dataset, where the authors recommended reporting only the accuracy as the evaluation metric.

3.6. Discussion

Table 3 summarizes the existing work on verifiable claim detection. Among the three language settings, minimal effort has been made in experimenting with cross-lingual solutions for verifiable claim detection tasks. Moreover, the verifiable claim detection task is often treated with the claim prioritization task using the same solutions. However, these two tasks are different in nature, and require special attention to differentiate between verifiability and priority of a claim. For example, Fig. 5 shows the statistics of the verifiable and check-worthy claim detection datasets from CheckThat 2020 (Nakov et al., 2022). Compared to the amount of verifiable claims, relatively very few check-worthy claims are presented in the dataset. This shows the importance of giving attention to the objective of the task to distinguish between verifiability and priority of a claim.

Further, it can be observed that most of the studies on verifiable claim detection rely on pre-trained transformer models, especially BERT and its variations to obtain state-of-the-art performance with limited fine-tuning. Moreover, limited research has utilized tweet-specific features for the classification of Twitter data, and translation is used as a data augmentation technique to handle the limited training data issue. Further, it can be noticed that performing fine-grained verifiable claim-type identification has been carried out in the literature only in monolingual settings. This could be due to the unavailability of the training data in a multilingual environment and the challenges associated with developing multilingual data annotated with fine-grained claim types. Moreover, very little effort has been made to date in the literature towards utilizing large language models (LLMs) for claim detection. We consider however that employing LLMs for claim detection represents an interesting avenue for future research.

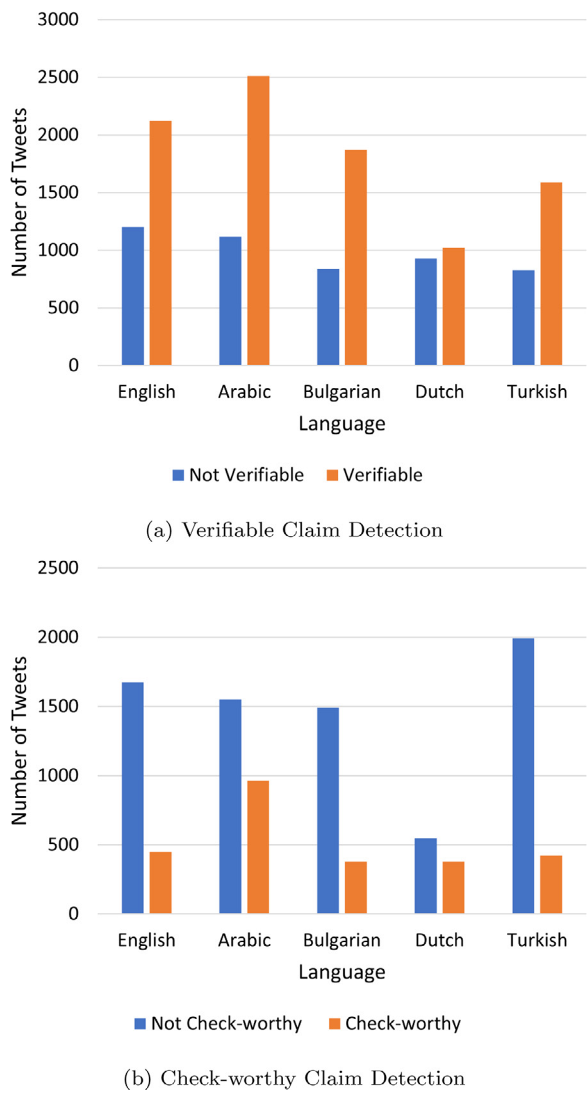

Fig. 5. Statistics of CheckThat 2022 verifiable and check-worthy datasets.

4. Claim prioritization

Prioritizing verifiable claims is a key task in the fact-checking pipeline as not all the verifiable claims can be fact-checked due to a limited availability of resources. However, defining the priority of a verifiable claim is a subjective decision and it depends on multiple factors as discussed in Section 2. Prioritizing verifiable claims is often treated as a check-worthy claim detection task, which aims at classifying a claim as check-worthy or not. However, the list of claims can be ranked according to the check-worthiness for prioritization. This enables the problem to be solved as either a classification or regression or ranking task. Apart from the check-worthiness, criteria such as attention-worthiness (Nakov et al., 2022), harmfulness (Nakov et al., 2022; Shaar et al., 2021a), and interest to the general public (Shaar et al., 2021a) are used to prioritize the claims. Similar to verifiable claim detection, the prioritization task does not ensure the availability of evidence, and prioritized claims may not have supporting or refusing evidence impeding the determination of a verdict.

4.1. Datasets

A series of CheckThat datasets released since 2018, serve as a rich source of training data for the claim prioritization task in multilingual settings. The initial version of the CheckThat 2018 dataset (Nakov et al., 2018) was released with political debates annotated for ranking check-worthy sentences only in English and Arabic languages. This dataset was collected in the English language, and the Arabic version of it was obtained through manual translation. Later, the problem was tackled as a classification task, and Tweets annotated with various prioritization criteria including check-worthiness, harmfulness, and attention-worthiness were released in multiple languages (Nakov et al., 2022). In addition to these criteria, the NLP4IF 2021 dataset includes Tweets annotated with binary labels indicating the *interest to the general public* (Shaar et al., 2021a). One of the largest annotated datasets for check-worthy claim detection was released by the authors of ClaimHunters (Beltrán et al., 2021) focusing on languages specific to a region. Table 2 presents the multilingual claim prioritization datasets.

4.2. Cross-lingual claim prioritization

ClaimRank (Jaradat et al., 2018) is one of the earliest attempts at detecting the check-worthiness of claims in a cross-lingual setting. The authors trained a neural network classifier and input cross-lingual embedding representation of the text along with a wide range of contextual features. This includes TF-IDF scores, part-of-speech (POS) tags, named entities, and topic vectors learned using the topic modeling technique, Latent Dirichlet Allocation (LDA) (Jelodar et al., 2019). Finally, the likelihood of the classifier was used to rank the claims.

Similar to verifiable claim detection, multilingual BERT (mBERT) (Devlin et al., 2019) has been widely used as a solution for claim prioritization in cross-lingual settings. Uyangodage et al. (2021) experimented with both cross-lingual and mono-lingual settings using BERT variations. The authors fine-tuned publicly available language-specific BERT models and mBERT using the language-specific training data and merged training data respectively. They observed similar or improved performance in cross-lingual settings in the CheckThat 2021 dataset. Zengin et al. (2021) analyzed a similar approach with language-specific BERT models and mBERT. The authors additionally explored data augmentation techniques such as machine translation and under-sampling (decreasing the size of majority classes) to overcome the imbalanced nature of the dataset. However, none of these techniques increased the performance of check-worthy claim identification.

Hasanain and Elsayed (2022) and Kartal and Kutlu (2022) analyzed the zero-shot learning by training the mBERT model only in training data of one language and testing its generalization capability in other languages. The authors observed that mBERT performs as good as the monolingual models trained on target languages. Similar studies have been conducted in the literature using other multilingual models such as XLM-R (Beltrán et al., 2021). Schlicht et al. (2023) proposed to modify the architecture of transformer models by introducing a world language adaptor, a lightweight and modular neural network on top of the multilingual transformers. During the experiments, the authors observed that adaptors trained for world languages were capable of transferring knowledge across languages.

Multi-tasking was experimented as an alternative solution to improve the performance of claim prioritization in cross-lingual settings. Schlicht et al. (2021) performed multi-task learning by jointly detecting the language and check-worthy claims. The authors used Sentence BERT (Reimers and Gurevych, 2019) trained on a multilingual dataset with dedicated fully connected layers for each task. Du et al. (2022) extended this work by performing a wide range of auxiliary tasks to enhance the performance, and observed an increase in performance for check-worthy claim detection tasks. Notable auxiliary tasks jointly learned include translation to English, verifiable claim detection, harmful tweet detection, and attention-worthy tweet detection.

Table 2
Multilingual claim prioritization datasets.

Dataset	Criteria	Task	Label	Language	Topic	Source	Size	Evaluation
CheckThat 2018 (Nakov et al., 2018)	Check-worthy	Ranking	0–1	English Arabic	Politics	Political Debate	7K–9K	MAP MRR MAP@K
NLP4IF 2021 (Shaar et al., 2021a)	Interesting Harmful Requires attention	Classification	Yes No	English Arabic Bulgarian	Covid-19	Twitter	1.3K–4K	Precision Recall F1 Score
ClaimHunter (Beltrán et al., 2021)	Check-worthy	Classification	Yes No	Spanish Catalan Galician Basque	Politics	Twitter	30K	Precision Recall F1 Score
CheckThat 2022 (Nakov et al., 2022)	Check-worthy Harmful Attention-worthy	Classification	Yes No	English Arabic Bulgarian Spanish Turkish Dutch	COVID-19 Politics	Twitter	4K–6K	F1 Score

4.3. Multilingual claim prioritization

Similar to the verifiable claim detection in a multilingual setting, a straightforward solution to identify claim priority in the multilingual environment can be achieved by training multilingual pre-trained models such as mBERT (Hasanain and Elsayed, 2020; Tarannum et al., 2022; Sadouk et al., 2023), XLM-r (Tarannum et al., 2022; Sadouk et al., 2023; Aziz et al., 2023), and GPT-3 (Sadouk et al., 2023) in language-specific training data. Different from these approaches, Nakov et al. (2021) translated the text written in other languages to English first using Google Translation API, and then trained a Support Vector Machine (SVM) classifier (Suthaharan and Suthaharan, 2016) for classifying check-worthy tweets in English. However, the effectiveness of this method was highly dependent on the precision of machine translation.

4.4. Monolingual claim prioritization

Developing a monolingual classification of claim priority was often solved by fine-tuning language-specific pre-trained models (Williams et al., 2021; Zhou et al., 2021) or combining other classification approaches such as neural networks (Martinez-Rico et al., 2020; Rony et al., 2020; Dutta et al., 2022), and SVM (Cheema et al., 2020) with monolingual word embedding representation. However, the problem can also be approached as a regression task and Logistic Regression (LaValley, 2008) has been widely used with the word embedding representation for ranking claim priority (Kartal, 2020; Kartal et al., 2020).

The latest attention in this direction has been given to identifying domain-specific claims priority. Pathak and Srihari (2021) and Pathak et al. (2020) developed solutions specific to news articles based on the assumption that sentences that could well represent the headlines are more check-worthy, and experimented with unsupervised techniques to prioritize check-worthy sentences. Gollapalli et al. (2023) attempted to extract medical claims and claim types discussing prevention, diagnoses, cures, treatments, and risks. The authors finetuned the Text-to-Text Transfer Transformer (T5) model (Raffel et al., 2020) for identifying claim priority, and the BART (Lewis et al., 2020) model was used to detect the claim types in a zero-shot setting.

4.5. Evaluation

Evaluation of claim prioritization depends on the nature of the task. The following measures are used in the literature to evaluate the ranking tasks.

- MAP@K (Mean Average Precision @ K) (Nakov et al., 2022) - Computed as the mean of the average precision value of all the data instances. Here, the average precision is computed as the average precision score for the range of K value (average of precision @ 1 to precision @ K).
- MRR (Mean Reciprocal Ranking) (Shaar et al., 2020; Kazemi et al., 2021b) - Given the actual ranking of top K elements, MRR is calculated using their corresponding retrieved as follows,

$$MRR = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{rank_K} \tag{4}$$

Metrics such as accuracy, precision, recall, and F1-score (refer Section 3.5) are widely used as an evaluation measure when the prioritization is carried out as a classification task. Table 2 shows the evaluation metrics used in the existing claim prioritization datasets. It can be observed that a wide selection of metrics is used in the literature depending on whether the task is modeled as a classification or ranking problem. It is worth noting that the CheckThat 2022 dataset (Nakov et al., 2022) recommended reporting the F1 score to account for class imbalance.

4.6. Discussion

Prioritizing a claim is generally modeled as an estimation of the check-worthiness of a claim, and the problem is solved as either a classification or ranking task in the literature. Table 3 summarizes the existing work on claim prioritization. It can be observed that transformer-based models are widely used in all three settings, and various data augmentation techniques including up-sampling, down-sampling, and machine translation were used to overcome the imbalance in the training dataset. Further, multi-tasking is also experimented as a solution to transfer the language and task knowledge in cross-lingual settings. Similar to the trend of treating both verifiable and check-worthy claims with the same solutions, different prioritization tasks are treated the same without giving much attention to the actual objective (e.g. check-worthiness, attention-worthiness, and harmfulness). Possibly this could be an interesting future direction for developing prioritization solutions incorporating the actual priority criteria.

Table 4 summarizes the models used in the literature for claim detection and claim prioritization. It can be noticed that transformer-based models are widely used compared to machine learning approaches and other deep learning learning approaches. Especially the BERT model and its architectural variations such as RoBERTa, XLM-r, ALBERT, BERTweet, Sentence BERT, DistilBERT, and ConvBERT, its multilingual variations such as mBERT, and the language-specific variations such as AraBERT, Spanish BERT, and BERTurk are commonly applied. Further, it can be noticed that large language models are yet to be explored in this line of research.

5. Claim matching

Claim matching is the task of identifying a pair of claims that can be addressed with the same fact-check (Kazemi et al., 2021b). This

Table 3
Summary of existing claim detection and prioritization research.

Research	Objective		Language			Model	Data augmentation		Input	Learning	Dataset		
	Verifiability	Priority	Cross-lingual	Multi-lingual	Mono-lingual	Transformer	Machine translation	Sampling	Twitter features	Multi-tasking	CheckThat	NLP4IF	Other
Panda and Levitan (2021)	✓	✓	✓	✓	–	✓	–	–	–	–	–	✓	–
Uyangodage et al. (2021)	✓	✓	✓		✓	✓	–	–	–	–	✓	✓	–
Alam et al. (2020)	✓	✓	✓		✓	✓	–	–	✓	–	–	✓	–
Agrestia et al. (2022)	✓	✓	✓	–	–	✓	–	–	–	–	✓	–	–
Hüsünbeyi et al. (2022)	✓	✓	–	✓	✓	✓	–	–	–	–	✓	–	–
Savchev (2022) and Eyuboglu et al. (2023)	✓	✓	–	✓	✓	✓	✓	–	–	–	✓	–	–
Suri and Dudeja (2022)	✓	✓	–	–	✓	✓	✓	–	✓	–	✓	–	–
Henia et al. (2021) and Hussein et al. (2021)	✓	✓	–	–	✓	✓	–	–	–	–	–	✓	–
Alam et al. (2021) and Smeros et al. (2021)	✓	✓	–	–	✓	✓	–	–	–	–	–	–	✓
Woloszyn et al. (2021) and Prabhakar et al. (2020)	✓	–	–	–	✓	✓	–	–	–	–	–	–	✓
Konstantinovskiy et al. (2021)	✓	–	–	–	✓	–	–	–	–	–	–	–	✓
Zengin et al. (2021)	–	✓	✓		✓	✓	✓	✓	–	–	✓	–	–
Hasanain and Elsayed (2022) and Kartal and Kutlu (2022)	–	✓	✓		✓	✓	–	–	–	–	✓	–	–
Schlicht et al. (2021)	–	✓	✓	–	–	✓	–	–	–	✓	✓	–	–
Du et al. (2022)	–	✓	✓	–	–	✓	–	✓	✓	✓	✓	–	–
Beltrán et al. (2021)	–	✓	✓	–	–	✓	–	–	–	–	–	–	✓
Schlicht et al. (2023)	–	✓	✓	–	–	✓	–	–	–	–	✓	–	–
Jaradat et al. (2018)	–	✓	✓	–	–	–	–	–	–	–	✓	–	–
Sadouk et al. (2023)	–	✓	–	✓	✓	✓	–	✓	–	–	✓	–	–
Aziz et al. (2023)	–	✓	–	✓	✓	✓	–	–	–	–	✓	–	–
Hasanain and Elsayed (2020)	–	✓	–	✓	–	✓	–	–	✓	–	✓	–	–
Tarannum et al. (2022)	–	✓	–	✓	–	✓	–	✓	–	–	✓	–	–
Nakov et al. (2021)	–	✓	–	✓	–	–	✓	–	–	–	–	–	✓
Cheema et al. (2020), Williams et al. (2021), Kartal (2020) and Kartal et al. (2020)	–	✓	–	–	✓	✓	–	–	–	–	✓	–	–
Pathak and Srihari (2021) and Gollapalli et al. (2023)	–	✓	–	–	✓	✓	–	–	–	–	–	–	✓
Martinez-Rico et al. (2020) and Zhou et al. (2021)	–	✓	–	–	✓	–	–	✓	–	–	✓	–	–
Rony et al. (2020)	–	✓	–	–	✓	–	–	–	–	–	–	–	✓

can be handled as either a classification task to classify whether the two claims match or not, or a regression or semantic similarity task to generate a score indicating the strength of the match. When modeling as a classification problem, the likelihood of the classifier can also be used as a score indicating the probability of the two statements discussing the same claim. The task can be further extended as a search problem from a database of verified claims, by producing a ranked list of verified claims matching the input claim using the scores obtained via classification, or regression, or semantic similarity function. This extended task is referred to as *fact-checked claim retrieval* or *verified claim retrieval*. Interestingly, the claim matching task is highly related to the next component of the fact-checking pipeline; evidence retrieval. Here, the underlying idea of finding the relationship between two claims or a claim-evidence pair remains the same, and two or more claims matched to the same evidence can be treated as similar claims that can be fact-checked together.

5.1. Datasets

As mentioned previously, claim-matching tasks are treated with various objectives in the literature, and a wide range of datasets serving these objectives are available in multilingual environments. This includes matching two tweets (Kazemi et al., 2021a), matching tweets with a report (Kazemi et al., 2022), and also matching a verified claim with tweets or social media posts (Shaar et al., 2021b; Nielsen and McConville, 2022; Pikuliak et al., 2023). Table 5 summarizes the existing multilingual claim-matching datasets.

5.2. Cross-lingual claim matching

Very little effort has been made in the direction of experimenting with claim matching in cross-lingual settings. The authors of the *MultiClaim* dataset (Pikuliak et al., 2023) used various multilingual and monolingual embedding representations of posts and claims, and the distance between the vector representation was used as a similarity function to retrieve fact-checked claims. Further, they compared the embedding-based retrieval using the BM25 ranking algorithm (Robertson et al., 2009). The experiment results showed the BM25 algorithm is ineffective in handling multilingual environments and multilingual embedding representations such as LaBSE (Feng et al., 2022) retrieve similar claims effectively. Instead of utilizing the embeddings directly, the authors of *MMTweets* dataset (Singh et al., 2023b) fine-tuned various multilingual models mBERT (Devlin et al., 2019), XLM-r (Kalyan et al., 2021) and LaBSE (Feng et al., 2022) to tackle the problem as a classification task. They observed a similar trend of low performance of the BM25 algorithm in multilingual settings, and LaBSE has consistently been the best model. Following these observations, fine-tuning LaBSE (Feng et al., 2022) was used as an effective solution for claim matching in the literature (Nielsen and McConville, 2022).

Different from these approaches, Larraz et al. (2023) integrated both a semantic similarity-based technique and classification to perform cross-lingual claim matching for political discourse. The K-Nearest Neighbors (KNN) algorithm (Peterson, 2009) was used to find the top N neighbors of the input claim first, and finally, a BERT-based classifier

Table 4
Summary of models used for claim detection and prioritization research.

Modeltype	Model	Reference
Machine learning	Logistic regression	Kartal (2020), Panda and Levitan (2021), Konstantinovskiy et al. (2021) and Beltrán et al. (2021)
	Random forest	Smeros et al. (2021) and Tarannum et al. (2022)
	Support vector machine	Konstantinovskiy et al. (2021), Schlicht et al. (2021), Beltrán et al. (2021) and Tarannum et al. (2022)
Deep learning	LSTM	Prabhakar et al. (2020)
	Bi-LSTM	Martinez-Rico et al. (2020) and Rony et al. (2020)
	CNN	Martinez-Rico et al. (2020)
	Feed forward NN	Martinez-Rico et al. (2020) and Konstantinovskiy et al. (2021)
	Falir	Woloszyn et al. (2021)
Transformer family	BERT	Kartal (2020), Prabhakar et al. (2020), Cheema et al. (2020), Alam et al. (2020), Panda and Levitan (2021), Pathak and Srihari (2021), Zhou et al. (2021), Uyangodage et al. (2021), Smeros et al. (2021), Alam et al. (2021), Zengin et al. (2021), Hasanain and Elsayed (2022), Savchev (2022), Kartal and Kutlu (2022), Suri and Dudeja (2022), Eyuboglu et al. (2023) and Sadouk et al. (2023)
	mBERT	Hasanain and Elsayed (2020), Alam et al. (2020), Panda and Levitan (2021), Uyangodage et al. (2021), Alam et al. (2021), Zengin et al. (2021), Hasanain and Elsayed (2022), Kartal and Kutlu (2022), Tarannum et al. (2022), Schlicht et al. (2023) and Sadouk et al. (2023)
	XLM-r	Alam et al. (2020, 2021), Beltrán et al. (2021), Hüsünbeyi et al. (2022), Tarannum et al. (2022), Du et al. (2022), Schlicht et al. (2023) and Aziz et al. (2023)
	RoBERTa	Alam et al. (2020), Zhou et al. (2021), Alam et al. (2021), Woloszyn et al. (2021), Williams et al. (2021), Savchev (2022), Sadouk et al. (2023) and Eyuboglu et al. (2023)
	DistilBERT	Prabhakar et al. (2020), Savchev (2022) and Eyuboglu et al. (2023)
	BERTweet	Zhou et al. (2021), Woloszyn et al. (2021), Hüsünbeyi et al. (2022) and Aziz et al. (2023)
	SentenceBERT	Schlicht et al. (2021)
	ConvBERT	Hüsünbeyi et al. (2022)
	ALBERT	Alam et al. (2021), Eyuboglu et al. (2023) and Sadouk et al. (2023)
	XL-Net	Kartal and Kutlu (2022) and Sadouk et al. (2023)
	Transformer	Panda and Levitan (2021), Du et al. (2022) and Gollapalli et al. (2023)
	FastText	Kartal (2020) and Alam et al. (2021)
	AraBERT	Alam et al. (2020), Hussein et al. (2021), Henia et al. (2021), Zengin et al. (2021), Hasanain and Elsayed (2022), Aziz et al. (2023) and Eyuboglu et al. (2023)
	Spanish BERT	Uyangodage et al. (2021)
	BERTurk	Uyangodage et al. (2021), Zengin et al. (2021), Hasanain and Elsayed (2022), Hüsünbeyi et al. (2022) and Eyuboglu et al. (2023)
	DutchBERT	Alam et al. (2020)
	SalvicBERT	Hasanain and Elsayed (2022)
Large language models	GPT	Agrestia et al. (2022) and Sadouk et al. (2023)

Table 5
Multilingual claim matching datasets.

Dataset	Label	Language	Language pair	Topic	Source	Size	Evaluation
Kazemi et al. (2021a)	Claim pairs	English Hindi Bengali Malayalam Tamil	Monolingual	Covid-19 Politics	WhatsApp	300–650 pairs	Accuracy Precision Recall F1 Score
Shaar et al. (2021b)	Claim-Tweets Pairs	English Arabic	Monolingual	Multitopic	Twitter Snopes AraFact (Ali et al., 2021) ClaimsKG (Tchechmedjiev et al., 2019)	2.5K pairs	MRR MAP@K Precision@K
Kazemi et al. (2022)	Claim-Report Pairs	English Hindi Spanish Portuguese	Monolingual Hindi-English	Multitopic	Twitter Google Fact Check Tools	400–4.8K pairs	Accuracy F1 Score MRR MAP@K
MuMiN (Nielsen and McConville, 2022)	Claim-Tweet Pairs	41 languages	Multilingual	Multitopic	Twitter Google Fact Check Tools	13K claims 21M tweets	F1 Score
MultiClaim (Pikuliak et al., 2023)	Claim-Post Pairs	27 languages	Multilingual	Multitopic	Face book, Twitter Instragam Google Fact Check Tools	31K Ppairs	Precision@K Recall@K
MMTweets (Singh et al., 2023b)	Claim-Misinformation Tweet Pairs	English Hindi Spanish Portuguese	Multilingual	Covid-19	Twitter Fact-checking Organizations	1.6K pairs	MRR MAP@K

Table 6
Summary of existing claim matching and claim clustering research.

Research	Task: Claim matching	Task: Clustering	Language: Cross-lingual	Language: Monolingual	Model: Transformer	Technique: Similarity function	Technique: Classification	Technique: Regression	Technique: Clustering algorithm
Kazemi et al. (2021b)	✓	✓	✓	–	✓	✓	✓	–	✓
Nielsen and McConville (2022)	✓	✓	✓	–	✓	–	✓	–	✓
Kazemi et al. (2022) and Larraz et al. (2023)	✓	–	✓	–	✓	✓	✓	–	–
Pikuliak et al. (2023)	✓	–	✓	✓	✓	✓	–	–	–
Singh et al. (2023b)	✓	–	✓	–	✓	–	✓	–	–
Shaar et al. (2022)	✓	–	–	✓	✓	✓	–	–	–
Shaar et al. (2020)	✓	–	–	✓	✓	✓	–	✓	–
Bouziane et al. (2020), Mansour et al. (2022), Singh et al. (2023a) and Mansour et al. (2023)	✓	–	–	✓	✓	–	✓	–	–
Adler and Boscaini-Gilroy (2019) and Hale et al. (2024)	–	✓	–	✓	✓	–	–	–	✓
Smeros et al. (2021)	–	✓	–	✓	–	–	–	–	✓

was employed to classify the claim pairs. The authors observed that the XLM-r (Kalyan et al., 2021) model was performing better among the BERT-based models compared to identifying the claim pairs among the top N neighbors.

Kazemi et al. (2022) solved a similar problem of matching a tweet to a report containing a similar claim as both classification and ranking tasks. The authors fine-tuned the multilingual transformer model XLM-r (Kalyan et al., 2021) to solve the classification problem. Ranking similar reports for the given tweet was achieved by developing a word embedding-based similarity search system using the sentence embedding representations (Reimers and Gurevych, 2019; Feng et al., 2022).

5.3. Multi-lingual claim matching

Interestingly, the claim matching problem is widely experimented in either cross-lingual settings or monolingual settings, and limited effort has been made to develop multi-lingual claim detection. In the only such effort to date, Kazemi et al. (2021b) solved claim matching as a classification problem in a multilingual environment, by training XLM-r (Kalyan et al., 2021) model for each monolingual language pair in the dataset they released, and then applying the BM25 ranking algorithm on language-specific embedding representation of claims. The authors observed an increase in performance in claim matching, when combining BM25 with XLM-r embedding representation, compared to performing BM25 ranking independently.

5.4. Mono-lingual claim matching

One of the pioneering works in this direction was experimented by Shaar et al. (2020). Given an unverified claim, the authors retrieved a ranked list of fact-checked claims. Both unverified claims and verified claims were represented using the sentence embedding representation obtained via the BERT model (Devlin et al., 2019) and its variations. Cosine similarity between the sentence embedding representations was used to retrieve a set of matching verified claims. Finally, a Support Vector Machine (SVM) model (Suthaharan and Suthaharan, 2016) was trained to rank the retrieved claim list. Shaar et al. (2022) applied the same technique with a wide range of sentence embedding representations combined with the SVM model for ranking. Apart from the integration of regression models with semantic similarity approaches, recent studies (Bouziane et al., 2020; Mansour et al., 2022, 2023) have focused on solving the verified claim retrieval problem as a classification task by fine-tuning BERT models.

A more recent study on monolingual claim matching (Singh et al., 2023a) shows the importance of having a larger scale training data for accurately retrieving the fact-checked claims. The authors generated a large number of synthetic claims from fact-checked claims using text-to-text transfer transformer (T5) and Chat-GPT models.[6] A large-scale synthetic dataset was used to fine-tune transformer models for matching claim pairs. Experiment results show that the proposed unsupervised method yields similar or slightly improved retrieval performance compared to the state-of-the-art transformer models directly trained on the claim-matching training data. This encourages generating synthetic data using large language models at a lower cost of computational efficiency to overcome the challenges associated with annotating large-scale training data.

5.5. Evaluation

Evaluating claim matching depends on whether the problem is treated as a classification or ranking task. Evaluation metrics discussed in Section 3.5 for the classification tasks and in Section 4.5 for the ranking tasks can be applied according to the nature of the problem. Table 5 shows the evaluation metrics used in the existing claim-matching datasets. It can be observed that Mean Reciprocal Ranking (MRR) and Mean Average Precision@K (MAP@K) are widely used for evaluation. Interestingly MultiClaim (Pikuliak et al., 2023) reports Precision@K and Recall@K as the ranking measures.

5.6. Discussion

Table 6 summarizes the claim-matching tasks in the literature. This problem is often solved as either a classification or regression or a semantic similarity task in the literature. In all these cases, transformer models are widely used as either the classification model or the embedding representation of the claims due to their superior performance in understanding the language and the task. Few studies experimented with ranking algorithms such as BM25 (Robertson et al., 2009) for retrieving similar claims in cross-lingual settings, and the experiment results reveal the algorithms perform at their optimal when combined with embedding representations such as LaBSE (Feng et al., 2022). Further, the difficulties associated with annotating large-scale training data have been addressed through the utilization of large language models for synthetic data generation.

[6] https://platform.openai.com/docs/models.

6. Claim clustering

This section summarizes existing work on the claim clustering task. The objective of clustering claims is to identify a set (including a pair or more) of claims expressing similar claims or similar topics. The former can be seen as a generalized problem of claim matching to identify the set of claims that can be fact-checked together. However, more coarse-grained clustering of claims can also be performed to further analyze claims belonging to the same topic. According to our knowledge, there is no training data annotated for identifying claim clusters existing even for the monolingual setting.

6.1. Cross-lingual claim clustering

Very little effort has been made in the direction of identifying multi-lingual clusters in the literature. The challenge of this task escalates with the unavailability of datasets annotated for claim clusters and makes it further difficult to evaluate the solutions proposed. Kazemi et al. (2021b) trained an XLM-R transformer model (Kalyan et al., 2021) to obtain the sentence embedding of the claims. Once the embedding representations were obtained, a single-link hierarchical clustering technique was applied to verify the existence of multi-lingual clusters. While the authors observed meaningful multilingual clusters in the dataset, they were enabled to evaluate the accuracy of the clusters obtained. A similar approach was used by Nielsen and McConville (2022) to cluster the claims using sentence embedding obtained via the pre-trained Sentence BERT model (Reimers and Gurevych, 2019). The authors applied the HDBSCAN clustering approach (McInnes et al., 2017) to the UMAP projection (McInnes et al., 2018) of the sentence embedding. While the authors observed the existence of topic clusters among the claims, they were enabled to evaluate the generated clusters due to the unavailability of the annotated dataset.

6.2. Monolingual claim clustering

The issue of the non-existence of a relevant dataset persists for monolingual claim clustering as well. Adler and Boscaini-Gilroy (2019) overcame this issue with the assumption that claims pertaining to the same news article should be clustered together. The authors used the Google USE Large pre-trained model (Cer et al., 2018) to obtain the embedding representation of the sentences and applied the DBSCAN clustering technique (Ester et al., 1996) to identify the claim clusters. They configured the clustering algorithm to allow even one claim to be a cluster to support sequential clustering (building the clusters dynamically by adding one element at a time). However, this could result in a huge amount of clusters, and the authors resolved this issue by applying the Louvain Community Detection algorithm (Blondel et al., 2008) to the identified clusters to determine the final clusters. The average percentage of claims belonging to the same story within a cluster was reported as a quantitative measure of the accuracy of the clusters identified.

Smeros et al. (2021) identified scientific claim clusters and reported a modified version of the Silhouette score (Rousseeuw, 1987), a widely used clustering evaluation metric for unlabeled data as the evaluation metric. The authors applied a wide range of domain-specific techniques to cluster scientific claims and related research papers represented as a bipartite graph. A notable clustering technique includes topic extraction from claim statements using Latent Dirichlet Allocation (Jelodar et al., 2019), and applying clustering algorithms such as K-Means (Lloyd, 1982) on topic vectors of claims. A similar idea of applying K-Means clustering techniques on the embedding representation of social media posts was also experimented to cluster posts discussing similar claims (Hale et al., 2024). Even though the authors could not report any quantitative measure of the accuracy of the clusters identified, the manual analysis revealed the existence of similar claims across multiple social platforms in different formats, languages, and lengths. This shows the requirement of performing claim clustering for effective fact-checking and misinformation management.

6.3. Evaluation

Due to the unavailability of a claim clustering dataset with cluster annotation, cluster evaluation metrics proposed for unlabeled clusters can be adopted for this task as a quantitative measure. Following is the cluster evaluation metric adopted for the claim clustering task in the literature:

- Silhouette score (Rousseeuw, 1987) - Computed as an average score of Silhouette score of all the data points. For a given data point i, the score is computed as follows,

$$S_i = \frac{b_i - a_i}{max(b_i, a_i)} \tag{5}$$

  where a_i is the average distance of data point i to any other point in the same cluster, and b_i is the average distance of point i to all the other points in the nearest cluster. Here, the distance between two claims can be defined using a semantic similarity function.

6.4. Discussion

Table 6 summarizes the claim clustering tasks in the literature. Clustering multilingual claims has been attempted on very limited occasions in the literature, and the quality of the clusters identified could not be precisely measured due to the unavailability of annotated datasets. As an alternative, traditional cluster evaluation metrics used for unlabeled clusters have been reported as a quantitative measure of claim cluster detection. Compared to the claim clustering task, more effort has been devoted to identifying a pair of claims or retrieving a list of claims discussing the same or similar claims. While this task can be treated with different objectives such as grouping similar claims for the same fact-check or grouping claims discussing the same topic, this research direction is still in its infancy and demands a thorough definition of the task and annotated datasets.

7. Open challenges

This section discusses open challenges associated with the ongoing research on multilingual claim detection.

- **Limited Multilingual Datasets:** One of the key aspects hindering the progress of multilingual claim detection research is the unavailability of training data. Especially, comprehensive multitopic claim detection datasets, verifiable claim type detection datasets, claim clustering datasets, and explainable claim detection are yet to be developed for the research progress even in monolingual settings. Automated generative approaches may be used as an alternative to generating claim detection datasets (Bussotti et al., 2023; Veltri et al., 2023).
- **Validity of Data Sources:** Annotating a massive amount of factual statements for their verifiability, priority, and similarity is a tedious and expensive task. This resulted in relying on existing tools such as the Google Fact Checking tool to partially automate the creation of training datasets. Further, the transparency in the data gathering and annotation process often does not persist, and these factors question the credibility of the dataset as well as the solutions developed on it.
- **Consolidated Definition of the Tasks:** Defining verifiability, priority, and similarity of claims may depend on various factors such as source, topic, target audience, etc. Therefore, a wide range of definitions are used in the literature to tackle all three aspects of the claim detection problem. This highly hinders the research progress with unified agreement on the definition of the tasks.
- **Change of Claim Status with Time:** Both the true value and the requirement to determine the verifiability, priority, and similarity of claims may change over time. Further, incorporating this temporal nature of the problem is scarcely explored in the literature, mainly due to the unavailability of datasets meeting these objectives, and the challenges associated with simulating the real-time environment for accurate experiments.

- **Demand of Generalizable Solutions:** As previously discussed, the source of factual statements can be from various platforms and can be articulated in various formats, languages, and modalities. Recent studies (Hale et al., 2024) have shown evidence of the existence of the same claims across multiple platforms, written in multiple formats, lengths, and details. While this demands more generalizable solutions to identify claims regardless of these factors, most of the existing research focuses on developing solutions specific to a source, data format, language, and modality.
- **Language Imbalance in Datasets:** Most of the existing multilingual datasets are composed of a higher number of annotated samples for high-resource languages such as English, compared to lesser-resourced languages. While existing research tried to tackle this problem via sampling, and via augmenting data in the underrepresented languages through machine translation techniques, this could lead to biases in the training when the model is provided with more data on certain languages.

8. Conclusion

Within the timely research area of automated fact-checking, this survey presents a comprehensive review of the existing multilingual claim detection research. Specifically, we detail the state-of-the-art techniques used to identify the verifiability, priority, and similarity of multilingual claims in the literature. It can be observed that relatively more effort has been given to prioritizing claims compared to identifying verifiable claims and similar claims. Further, fine-tuning multilingual pre-trained transformer models is widely used as a solution in all three problems, due to their powerful nature in transferring knowledge across languages and tasks.

As discussed previously, various factors affect the progress of multilingual claim detection research. Notably, the unavailability of datasets meeting the task requirement serves as a key challenge. Promising future direction includes the development of credible and comprehensive datasets, generalized solutions, explainable claim detection, time-aware claim detection, and prompt-based fine-tuning using large language models.

CRediT authorship contribution statement

Rrubaa Panchendrarajan: Writing – review & editing, Writing – original draft, Investigation, Conceptualization. **Arkaitz Zubiaga:** Writing – review & editing, 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.

Acknowledgments

This project is funded by the European Union and UK Research and Innovation under Grant No. 101073351 as part of Marie Skłodowska-Curie Actions (MSCA Hybrid Intelligence to monitor, promote, and analyze transformations in good democracy practices).

References

Abumansour, A.S., Zubiaga, A., 2023. Check-worthy claim detection across topics for automated fact-checking. PeerJ Comput. Sci. 9, e1365.

Adler, B., Boscaini-Gilroy, G., 2019. Real-time claim detection from news articles and retrieval of semantically-similar factchecks. arXiv preprint arXiv:1907.02030.

Agrestia, S., Hashemianb, A., Carmanc, M., 2022. PoliMi-FlatEarthers at CheckThat! 2022: GPT-3 applied to claim detection. Work. Notes CLEF.

Akbik, A., Blythe, D., Vollgraf, R., 2018. Contextual string embeddings for sequence labeling. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 1638–1649.

Alam, F., Dalvi, F., Shaar, S., Durrani, N., Mubarak, H., Nikolov, A., Da San Martino, G., Abdelali, A., Sajjad, H., Darwish, K., et al., 2021. Fighting the COVID-19 infodemic in social media: a holistic perspective and a call to arms. In: Proceedings of the International AAAI Conference on Web and Social Media. Vol. 15, pp. 913–922.

Alam, F., Shaar, S., Dalvi, F., Sajjad, H., Nikolov, A., Mubarak, H., Martino, G.D.S., Abdelali, A., Durrani, N., Darwish, K., et al., 2020. Fighting the COVID-19 infodemic: modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. arXiv preprint arXiv:2005.00033.

Ali, Z.S., Mansour, W., Elsayed, T., Al-Ali, A., 2021. AraFacts: the first large arabic dataset of naturally occurring claims. In: Proceedings of the Sixth Arabic Natural Language Processing Workshop. pp. 231–236.

Aziz, A., Hossain, M., Chy, A., 2023. CSECU-DSG at CheckThat! 2023: transformer-based fusion approach for multimodal and multigenre check-worthiness. Work. Notes CLEF.

Beltrán, J., Míguez, R., Larraz, I., 2021. ClaimHunter: An unattended tool for automated claim detection on Twitter. In: KnOD@ WWW.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Statist. Mech.: Theory Exp. 2008 (10), P10008.

Bouziane, M., Perrin, H., Cluzeau, A., Mardas, J., Sadeq, A., 2020. Team buster. ai at CheckThat! 2020 insights and recommendations to improve fact-checking. In: CLEF (Working Notes).

Bussotti, J.-F., Veltri, E., Santoro, D., Papotti, P., 2023. Generation of training examples for tabular natural language inference. Proc. ACM Manag. Data 1 (4), 1–27.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al., 2018. Universal sentence encoder. arXiv preprint arXiv:1803.11175.

Cheema, G.S., Hakimov, S., Ewerth, R., 2020. Check square at CheckThat! 2020: Claim detection in social media via fusion of transformer and syntactic features.

Das, A., Liu, H., Kovatchev, V., Lease, M., 2023. The state of human-centered NLP technology for fact-checking. Inf. Process. Manag. 60 (2), 103219.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. http://dx.doi.org/10.18653/v1/N19-1423, URL: https://aclanthology.org/N19-1423.

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

Du, S., Gollapalli, S.D., Ng, S.-K., 2022. Nus-ids at checkthat! 2022: identifying check-worthiness of tweets using checkthat5. Work. Notes CLEF.

Dutta, S., Dhar, R., Guha, P., Murmu, A., Das, D., 2022. A multilingual dataset for identification of factual claims in Indian Twitter. In: Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation. pp. 88–92.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd. Vol. 96, pp. 226–231.

Eyuboglu, A.B., Altun, B., Arslan, M.B., Sonmezer, E., Kutlu, M., 2023. Fight against misinformation on social media: Detecting attention-worthy and harmful tweets and verifiable and check-worthy claims. In: International Conference of the Cross-Language Evaluation Forum for European Languages. Springer, pp. 161–173.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., Wang, W., 2022. Language-agnostic BERT sentence embedding. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 878–891.

Gollapalli, S.D., Du, M., Ng, S.-K., 2023. Identifying checkworthy CURE claims on Twitter. In: Proceedings of the ACM Web Conference 2023. pp. 4015–4019.

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

Hale, S.A., Belisario, A., Mostafa, A., Camargo, C., 2024. Analyzing misinformation claims during the 2022 Brazilian general election on WhatsApp, Twitter, and kwai. arXiv preprint arXiv:2401.02395.

Hardalov, M., Arora, A., Nakov, P., Augenstein, I., 2022. A survey on stance detection for mis-and disinformation identification. In: Findings of the Association for Computational Linguistics. NAACL 2022, pp. 1259–1277.

Hasanain, M., Elsayed, T., 2020. bigIR at CheckThat! 2020: Multilingual BERT for ranking arabic tweets by check-worthiness. In: CLEF (Working Notes).

Hasanain, M., Elsayed, T., 2022. Cross-lingual transfer learning for check-worthy claim identification over Twitter. arXiv preprint arXiv:2211.05087.

Henia, W., Rjab, O., Haddad, H., Fourati, C., 2021. Icompass at NLP4IF-2021–fighting the COVID-19 infodemic. In: Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda. pp. 115–118.

Hussein, A., Ghneim, N., Joukhadar, A., 2021. DamascusTeam at NLP4if2021: Fighting the arabic COVID-19 infodemic on Twitter using AraBERT. In: Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda. pp. 93–98.

Hüsünbeyi, Z.M., Deck, O., Scheffler, T., 2022. RUB-DFL at CheckThat! 2022: Transformer models and linguistic features for identifying relevant.

Jaradat, I., Gencheva, P., Barrón-Cedeño, A., Màrquez, L., Nakov, P., 2018. ClaimRank: Detecting check-worthy claims in arabic and english. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. pp. 26–30.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L., 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools Appl. 78, 15169–15211.
Kalyan, K.S., Rajasekharan, A., Sangeetha, S., 2021. Ammus: A survey of transformer-based pretrained models in natural language processing. arXiv preprint arXiv:2108.05542.
Kartal, Y.S., 2020. TOBB ETU at CheckThat! 2020: Prioritizing english and arabic claims based on check-worthiness.
Kartal, Y.S., Guvenen, B., Kutlu, M., 2020. Too many claims to fact-check: Prioritizing political claims based on check-worthiness. arXiv preprint arXiv:2004.08166.
Kartal, Y.S., Kutlu, M., 2022. Re-think before you share: A comprehensive study on prioritizing check-worthy claims. IEEE Trans. Comput. Soc. Syst. 10 (1), 362–375.
Kazemi, A., Garimella, K., Gaffney, D., Hale, S., 2021a. Claim matching beyond english to scale global fact-checking. In: Zong, C., Xia, F., Li, W., Navigli, R. (Eds.), 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). Association for Computational Linguistics, Online, pp. 4504–4517. http://dx.doi.org/10.18653/v1/2021.acl-long.347, URL: https://aclanthology.org/2021.acl-long.347.
Kazemi, A., Garimella, K., Gaffney, D., Hale, S., 2021b. 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 (Volume 1: Long Papers). pp. 4504–4517.
Kazemi, A., Li, Z., Peréz-Rosas, V., Hale, S.A., Mihalcea, R., 2022. Matching tweets with applicable fact-checks across languages.
Konstantinovskiy, L., Price, O., Babakar, M., Zubiaga, A., 2021. Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. Digit. Threats: Res. Pract. 2 (2), 1–16.
Kotonya, N., Toni, F., 2020a. Explainable automated fact-checking: A survey. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 5430–5443.
Kotonya, N., Toni, F., 2020b. Explainable automated fact-checking for public health claims. arXiv preprint arXiv:2010.09926.
Larraz, I., Míguez, R., Sallicati, F., 2023. Semantic similarity models for automated fact-checking: ClaimCheck as a claim matching tool. Prof. Inf. 32 (3).
LaValley, M.P., 2008. Logistic regression. Circulation 117 (18), 2395–2399.
Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L., 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp. 7871–7880. http://dx.doi.org/10.18653/v1/2020.acl-main.703, URL: https://aclanthology.org/2020.acl-main.703.
Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
Lloyd, S., 1982. Least squares quantization in PCM. IEEE Trans. Inf. Theory 28 (2), 129–137.
Mansour, W., Elsayed, T., Al-Ali, A., 2022. Did i see it before? detecting previously-checked claims over twitter. In: European Conference on Information Retrieval. Springer, pp. 367–381.
Mansour, W., Elsayed, T., Al-Ali, A., 2023. This is not new! spotting previously-verified claims over Twitter. Inf. Process. Manage. 60 (4), 103414.
Martinez-Rico, J.R., Araujo, L., Martinez-Romo, J., 2020. NLP&IR@ UNED at CheckThat! 2020: A preliminary approach for check-worthiness and claim retrieval tasks using neural networks and graphs.
McInnes, L., Healy, J., Astels, S., 2017. Hdbscan: Hierarchical density based clustering. J. Open Source Softw. 2 (11), 205.
McInnes, L., Healy, J., Melville, J., 2018. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.
Micallef, N., Armacost, V., Memon, N., Patil, S., 2022. True or false: Studying the work practices of professional fact-checkers. Proc. ACM Hum.-Comput. Interact. 6 (CSCW1), 1–44.
Nakov, P., Alam, F., Shaar, S., Da San Martino, G., Zhang, Y., 2021. A second pandemic? Analysis of fake news about COVID-19 vaccines in Qatar. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing. RANLP 2021, pp. 1010–1021.
Nakov, P., Barrón-Cedeño, A., Da San Martino, G., Alam, F., Míguez, R., Caselli, T., Kutlu, M., Zaghouani, W., Li, C., Shaar, S., et al., 2022. Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets. In: 2022 Conference and Labs of the Evaluation Forum. CLEF 2022, CEUR Workshop Proceedings (CEUR-WS. org), pp. 368–392.
Nakov, P., Barrón-Cedeno, A., Elsayed, T., Suwaileh, R., Màrquez, L., Zaghouani, W., Atanasova, P., Kyuchukov, S., Da San Martino, G., 2018. Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings 9. Springer, pp. 372–387.
Nguyen, D.Q., Vu, T., Nguyen, A.T., 2020. BERTweet: A pre-trained language model for english tweets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 9–14.
Nielsen, D.S., McConville, R., 2022. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 3141–3153.
Panda, S., Levitan, S.I., 2021. Detecting multilingual COVID-19 misinformation on social media via contextualized embeddings. In: Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda. pp. 125–129.
Pathak, A., Shaikh, M.A., Srihari, R.K., 2020. Self-supervised claim identification for automated fact checking. In: Proceedings of the 17th International Conference on Natural Language Processing. ICON, pp. 213–227.
Pathak, A., Srihari, R.K., 2021. Assessing effectiveness of using internal signals for check-worthy claim identification in unlabeled data for automated fact-checking. arXiv preprint arXiv:2111.01706.
Peterson, L.E., 2009. K-nearest neighbor. Scholarpedia 4 (2), 1883.
Pfeiffer, J., Rücklé, A., Poth, C., Kamath, A., Vulić, I., Ruder, S., Cho, K., Gurevych, I., 2020. AdapterHub: A framework for adapting transformers. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 46–54.
Pikuliak, M., Srba, I., Moro, R., Hromadka, T., Smolen, T., Melisek, M., Vykopal, I., Simko, J., Podrouzek, J., Bielikova, M., 2023. Multilingual previously fact-checked claim retrieval. arXiv preprint arXiv:2305.07991.
Prabhakar, A.A., Mohtaj, S., Möller, S., 2020. Claim extraction from text using transfer learning.. In: Proceedings of the 17th International Conference on Natural Language Processing. ICON, pp. 297–302.
Radford, A., Narasimhan, K., 2018. Improving language understanding by generative pre-training. URL: https://api.semanticscholar.org/CorpusID:49313245.
Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. 21 (1), 5485–5551.
Reimers, N., Gurevych, I., 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.
Robertson, S., Zaragoza, H., et al., 2009. The probabilistic relevance framework: BM25 and beyond. Found. Trends® Inf. Retr. 3 (4), 333–389.
Rony, M.M.U., Hoque, E., Hassan, N., 2020. ClaimViz: Visual analytics for identifying and verifying factual claims. In: 2020 IEEE Visualization Conference. VIS, IEEE, pp. 246–250.
Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65. http://dx.doi.org/10.1016/0377-0427(87)90125-7, URL: https://www.sciencedirect.com/science/article/pii/0377042787901257.
Sadouk, H.T., Sebbak, F., Zekiri, H.E., 2023. ES-VRAI at CheckThat! 2023: Analyzing checkworthiness in multimodal and multigenre.
Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.
Savchev, A., 2022. AI rational at CheckThat! 2022: using transformer models for tweet classification. Work. Notes CLEF.
Schlicht, I.B., de Paula, A.F.M., Rosso, P., 2021. UPV at checkthat! 2021: mitigating cultural differences for identifying multilingual check-worthy claims. arXiv preprint arXiv:2109.09232.
Schlicht, I.B., Flek, L., Rosso, P., 2023. Multilingual detection of check-worthy claims using world languages and adapter fusion. In: European Conference on Information Retrieval. Springer, pp. 118–133.
Shaar, S., Alam, F., Da San Martino, G., Nikolov, A., Zaghouani, W., Nakov, P., Feldman, A., 2021a. Findings of the NLP4IF-2021 shared task on fighting the COVID-19 infodemic and censorship detection. In: Proceedings of the Fourth Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda. In: NLP4IF@NAACL' 21, Association for Computational Linguistics, Online.
Shaar, S., Babulkov, N., Da San Martino, G., Nakov, P., 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.
Shaar, S., Georgiev, N., Alam, F., Da San Martino, G., Mohamed, A., Nakov, P., 2022. Assisting the human fact-checkers: Detecting all previously fact-checked claims in a document. In: Findings of the Association for Computational Linguistics. EMNLP 2022, pp. 2069–2080.
Shaar, S., Hasanain, M., Hamdan, B., Ali, Z.S., Haouari, F., Nikolov, A., Kutlu, M., Kartal, Y.S., Alam, F., Da San Martino, G., et al., 2021b. Overview of the CLEF-2021 CheckThat! lab task 1 on check-worthiness estimation in tweets and political debates. In: CLEF (Working Notes). pp. 369–392.
Singh, I., Scarton, C., Bontcheva, K., 2023a. UTDRM: unsupervised method for training debunked-narrative retrieval models. EPJ Data Sci. 12 (1), 59.
Singh, I., Scarton, C., Song, X., Bontcheva, K., 2023b. Finding already debunked narratives via multistage retrieval: Enabling cross-lingual, cross-dataset and zero-shot learning. arXiv preprint arXiv:2308.05680.
Smeros, P., Castillo, C., Aberer, K., 2021. Sciclops: Detecting and contextualizing scientific claims for assisting manual fact-checking. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. pp. 1692–1702.
Suri, P.M., Dudeja, S., 2022. Asatya at CheckThat! 2022: multimodal BERT for identifying claims in tweets. Work. Notes CLEF.

Suthaharan, S., Suthaharan, S., 2016. Support vector machine. In: Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning. Springer, pp. 207–235.

Tarannum, P., Hasan, M.A., Alam, F., Noori, S.R.H., 2022. Z-index at CheckThat! lab 2022: Check-worthiness identification on tweet text.

Tchechmedjiev, A., Fafalios, P., Boland, K., Gasquet, M., Zloch, M., Zapilko, B., Dietze, S., Todorov, K., 2019. ClaimsKG: A knowledge graph of fact-checked claims. In: The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18. Springer, pp. 309–324.

Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A., 2018. FEVER: a large-scale dataset for fact extraction and VERification. In: Walker, M., Ji, H., Stent, A. (Eds.), 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, pp. 809–819. http://dx.doi.org/10.18653/v1/N18-1074, URL: https://aclanthology.org/N18-1074.

Uyangodage, L., Ranasinghe, T., Hettiarachchi, H., 2021. Can multilingual transformers fight the COVID-19 infodemic? In: Proceedings of the International Conference on Recent Advances in Natural Language Processing. RANLP 2021, pp. 1432–1437.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Veltri, E., Badaro, G., Saeed, M., Papotti, P., 2023. Data ambiguity profiling for the generation of training examples. In: 2023 IEEE 39th International Conference on Data Engineering. ICDE, IEEE, pp. 450–463.

Williams, E., Rodrigues, P., Tran, S., 2021. Accenture at CheckThat! 2021: Interesting claim identification and ranking with contextually sensitive lexical training data augmentation.

Woloszyn, V., Kobti, J., Schmitt, V., 2021. Towards automatic green claim detection. In: Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation. pp. 28–34.

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

Zengin, M.S., Kartal, Y.S., Kutlu, M., 2021. TOBB ETU at CheckThat! 2021: Data engineering for detecting check-worthy claims. In: CLEF (Working Notes). pp. 670–680.

Zhou, X., Wu, B., Fung, P., 2021. Fight for 4230 at CheckThat! 2021: Domain-specific preprocessing and pretrained model for ranking claims by check-worthiness. In: CLEF (Working Notes). pp. 681–692.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., Procter, R., 2018. Detection and resolution of rumours in social media: A survey. ACM Comput. Surv. 51 (2), 1–36.