ARTICLE

WILEY

Automated fact-checking: A survey

Xia Zeng[1] ⓘ | **Amani S. Abumansour**[1,2] | **Arkaitz Zubiaga**[1] ⓘ

[1]School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK

[2]Department of Computer Science, Taif University, Taif, Saudi Arabia

Correspondence

Arkaitz Zubiaga, School of Electronic
Engineering and Computer Science,
Queen Mary University of London,
London E1 4NS, UK.
Email: a.zubiaga@qmul.ac.uk

Abstract

As online false information continues to grow, auto-
mated fact-checking has gained an increasing amount
of attention in recent years. Researchers in the field of
Natural Language Processing (NLP) have contributed
to the task by building fact-checking datasets, devis-
ing automated fact-checking pipelines and proposing
NLP methods to further research in the development
of different components. This article reviews relevant
research on automated fact-checking covering both the
claim detection and claim validation components.

1 | INTRODUCTION

While online content continues to grow unprecedentedly, the spread of false information online
increases the potential of misleading people and causing harm. This leads to an increasing de-
mand on fact-checking, that is, a task consisting in assessing the truthfulness of a claim (Vlachos
& Riedel, 2014), where a claim is defined as 'a factual statement that is under investigation'
(Hanselowski, 2020). A number of fact-checking organisations have been founded in recent years,
for example, FactCheck, PolitiFact, Full Fact, Snopes, Poynter and NewsGuard. Fact-checkers
have conducted laborious manual fact-checking, that is, familiarise with the topic, identify the
claim, aggregate evidence, check source credibility, verify the claim and its reasoning chain and
check for fallacies (Hanselowski, 2020). However, the speed and efficiency of manual fact-check-
ing cannot keep up with the pace at which online information is posted and circulated. The
journalism community can benefit from tools that, at least partially, automate the fact-checking

process (Cohen et al., 2011; Hassan, Arslan, et al., 2017; Konstantinovskiy et al., 2021; Thorne & Vlachos, 2018), particularly by automating more mechanical tasks, so that human effort can instead be dedicated to more labour-intensive tasks (Babakar & Moy, 2016). Restricting claims to those that are objectively fact-checkable makes the automation task more realistically achievable while reducing the volume of content needing manual fact-checking. Furthermore, recent progress in the fields of natural language processing (NLP), information retrieval (IR) and big data mining has demonstrated the potential for efficiently processing large-scale textual information online, which inspires automated fact-checking.

Researchers have developed valuable fact-checking datasets, pipelines and models, an effort which has also been supported by shared tasks, including RumourEval (Derczynski et al., 2017; Gorrell et al., 2018), CLEF CheckThat! (Barron-Cedeno et al., 2020; Elsayed et al., 2019; Nakov et al., 2018), ClaimBuster (Hassan, Zhang et al., 2017), FEVER (Thorne et al., 2019; Thorne, Vlachos, Christodoulopoulos, & Mittal, 2018), SCIVER (Wadden et al., 2020), Fake News Challenge (Pomerleau & Rao, 2017) and HeroX fact checking challenge (Francis & Fact, 2016).

With different major concerns, proposed pipelines take various forms. For instance, ClaimBuster (Hassan, Zhang et al., 2017) designed a comprehensive pipeline of four components to verify web documents: a claim monitor that performs document retrieval, a claim spotter that performs claim detection, a claim matcher that matches a detected claim to fact-checked claims and a claim checker that performs evidence extraction and claim validation. A similar pipeline was proposed by CLEF CheckThat! (Nakov et al., 2021), which in its 2021 edition included three subtasks: firstly, perform claim detection to detect claims that are check-worthy; secondly, determine whether a claim has been previously fact-checked; and thirdly, perform claim validation to determine the factuality of the detected claims. While some pipelines include claim detection, some are only designed to tackle claim validation, for example, FEVER (Thorne et al., 2019; Thorne, Vlachos, Christodoulopoulos, & Mittal, 2018) and SCIVER (Wadden et al., 2020),[1] assuming check-worthy claims are already at hand. Figure 1 depicts a comprehensive fact-checking pipeline as discussed in this survey and consisting of two components: (1) a claim detection component, which looks for claims that need checking and tries to find matches between claims when they are related to the same fact-check and (2) a claim validation component, which retrieves the documents and rationales that can serve as evidence to fact-check a claim and ultimately performs the verification task, producing a verdict.

In this article, we present a survey on automated fact-checking with special focus on claim detection and claim validation, and is structured as follows. Section 2 focuses on the task of detecting check-worthy claims, which is the very first task of a comprehensive automated fact-checking pipeline. It also contains a brief overview of claim matching. Section 3 presents the task of claim validation, which typically involves addressing evidence retrieval and claim verification together. In Section 4, we discuss advantages and drawbacks of current automated fact-checking pipelines, with a focus on current challenges and promising future directions. Section 5 presents closely related NLP tasks. Conclusions are drawn in Section 6.

2 | CLAIM DETECTION

Claim detection plays a crucial role in automated fact-checking systems as all other components need to rely on the output of this stage. It aims to relief the burden of identifying claims for fact-checkers and help them by minimising the volume of online content they need to deal with.
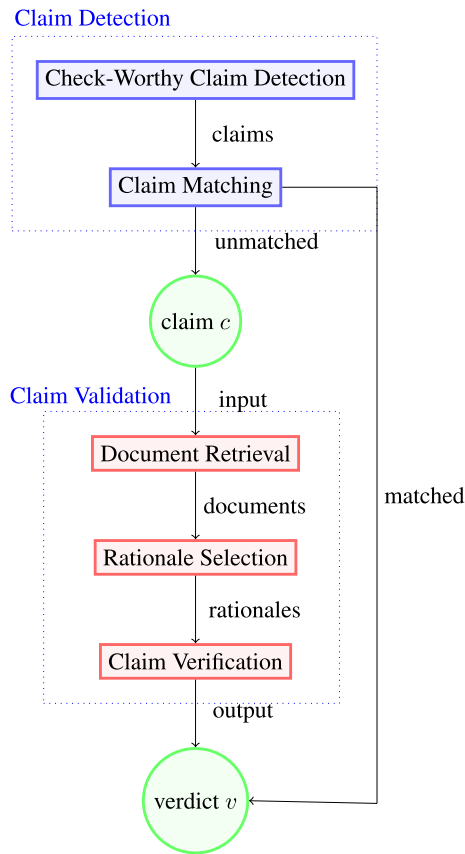
FIGURE 1 An overview of automated fact-checking system

2.1 | Approaches

The claim detection component is responsible for selecting claims that need to go through the rest of the fact-checking pipeline due to needing to be checked, that is, needing verification. For instance, a factual statement such as 'He voted against the first gulf war' can be deemed a claim that should be fact-checked. In contrast, a piece of opinion such as 'I think it's time to talk about the future' is not a claim that should be fact-checked (Hassan, Arslan, et al., 2017).

Going further, one can also distinguish between check-worthy versus non-check-worthy claims (Nakov et al., 2021), as cases that being claims are worthy or not fact-checking. For example, one could argue that 'the government invested more than 10 billion last year in education' is a claim that is worthy of fact-checking, whereas a claim such as 'my friend had a coffee this morning for breakfast' may not be worthy of fact-checking.

Researchers typically formulate the problem as one having a set of sentences as input (e.g., originating from a debate or conversation), and is tackled as a classification task, where a binary decision is made on whether each input sentence constitutes a claim or not, or a ranking task, where input sentences are ranked by check-worthiness, hence prioritising top claims on top positions of the list.

2.2 | Datasets

In recent studies, several datasets were built with the purpose of enabling training machine learning models to predict check-worthy claims, as shown in Table 1. The vast majority of datasets cover sentences pertaining to the political domain, as a result of events that synchronously occur with the United States elections. In contrast, the CheckThat! Lab released English and Arabic datasets that contain a small number of instances related to COVID-19 in early 2020.

In addition, publicly available datasets have a variety of sizes. For instance, ClaimBuster (Arslan et al., 2020) and CT19-T1 (Atanasova, Nakov, Karadzhov, et al., 2019) are the largest datasets, while CW-USPD-2016 (Gencheva et al., 2017), CT-CWC-18 (Atanasova et al., 2018), CT20-AR (Hasanain et al., 2020) and FactRank (Berendt et al., 2020) are a degree of magnitude smaller, followed by other smaller datasets.

Datasets have binary-class for either single-label or multi-labels, depending on the annotation process. The annotation process comes in diverse types. Some datasets are automatically built by collecting claims from fact-checking websites, while other datasets rely on manual annotations given specific definitions of check-worthiness. Crowd-sourcing platform has also been demonstrated to be helpful (Hassan et al., 2015).

Moreover, the majority of datasets are available in the English language as opposed to a smaller number of datasets in the Arabic language. Most of these Arabic datasets are generated from

TABLE 1 Check-worthiness claim detection datasets

Name	Size	Annotation type	Language
ClaimBuster (Arslan et al., 2020)	23,533 sentences	Manual	English
CW-USPD-2016 (Gencheva et al., 2017)	5415 sentences	From existing annotation	English
TATHYA (Patwari et al., 2017)	15,735 sentences	From existing annotation	English
Konstantinovskiy et al. (2021)	5571 sentences	Based on the annotation scheme, sentences labelled into seven categories then grouped into two categories	English
CT-CWC-18 (Atanasova et al., 2018)	8946 (En), 7254 (Ar) sentences	From existing annotation	English/partially translated to Arabic
CT19-T1 (Atanasova, Nakov, Karadzhov, et al., 2019)	23,500 sentences	From existing annotation	English
Shaar, Nikolov et al. (2020)	962 tweets	Manual	English
CT20-AR (Hasanain et al., 2020)	7.5k tweets	Manual	Arabic
TrClaim-19 (Kartal & Kutlu, 2020)	2287 tweets	Manual	Turkish
FactRank (Berendt et al., 2020)	7037 sentences	Manual	Dutch

translations of originally English datasets, except for CT20-AR, which is originally Arabic content. In addition, there is one dataset in Dutch and another one in Turkish, while datasets in other languages are not yet available.

2.3 | Check-worthy claim detection

ClaimBuster is the first automated fact-checking system that consists of integrated components tackling the entire fact-checking pipeline, starting off from the claim detection component. Its claim detection component called 'claim spotter' classifies input sentences into one of (1) a factual claim, (2) an unimportant factual claim or (3) a non-factual claim. This in turn assists fact checkers by prioritising the most check-worthy claims by ranking them based on accuracy measures such as Precision at k (P@K; Hassan, Zhang et al., 2017). To develop this, a multi-class support vector machine (SVM) classifier was built which used features such as bags-of-words, Part-Of-Speech (POS) tags and entity types (ET). The model achieved competitive performance and was considered as the baseline to beat in subsequent works (Hassan, Arslan, et al., 2017).

Another model called 'CNC' (i.e., 'Claim/not Claim'; Konstantinovskiy et al., 2021) builds on top of InferSent embeddings (Conneau et al., 2017), combining them with POS tags and named entities found in texts, which are fed to a Logistic Regression classifier. Authors of CNC had as their main goal the improvement of the recall score achieved by their system, arguing that fact-checkers do not want to miss out any claims (no false negatives) while they can deal with some false positives. While improving in terms of recall, CNC also achieved superior performance in $F1$ score.

Apart from traditional machine learning models, neural networks have also been studied for the claim detection task. For example, in the CheckThat! Lab 2019 shared task, LSTM neural networks and Feed forward neural networks were the most effective models used by the top two participants. Along with the use of neural networks, top participants also showed the usefulness of context (i.e., surrounding sentences) in improving claim detection performance (Elsayed et al., 2019). The use of context was studied in more detail in another work conducted outside the shared task, in this case by (Atanasova, Nakov, Màrquez, et al., 2019). They studied the inclusion of context and discourse features along with sentence-level features. They used a Feed-Forward Neural Network (FNN) as the model, which was then evaluated as a ranking task, proving the effectiveness of context and discourse features.

While all aforementioned works focused on English claims, there have also been efforts in other languages. For example, the ClaimRank model (Jaradat et al., 2018) was tested on Arabic claims (translated from claims originally in English). The Arabic claim detection model used Farasa (Abdelali et al., 2016) for tokenisation, POS tagging, as well as MUSE embeddings. The first experiments on original Arabic data (rather than translated) were conducted in the CheckThat! 2020 shared task. Most participants proposed approaches fine-tuning pre-trained language models. For instance, the top-performing participant fine-tuned AraBERT v0.1 with neural networks (Williams et al., 2020). Likewise Hasanain and Elsayed (2020) fine-tuned multilingual BERT (mBERT) with different classification models. Another recent effort, called FactRank (Berendt et al., 2020), focused on claim check-worthiness detection for the Dutch language, in this case using a convolutional neural network (CNN) along with Platt scaling for an SVM model and a softmax to obtain the degree of check-worthiness.

Given that methods for claim detection have been applied in different settings or on different datasets, it is difficult to establish what the state-of-the-art model is today. However, the best way

of determining the best-performing model is to look at the leaderboard in the CheckThat! shared task of the most recent edition.

2.4 | Claim matching

Another task that has recently emerged is claim matching, also referred to as identifying previously fact-checked claims. For a claim spotted in the claim detection component, claim matching consists in determining whether this is a claim that exists in the database and can be resolved by a previous fact-check. The task is formulated as follows: given a check-worthy claim as input, and having a database of previously fact-checked claims, it consists in determining if any of the claims in the database is related to the input; in this case, the new claim would not need fact-checking again, as it was fact-checked in the past. It is normally framed as a ranking task, where claims in the database are ranked based on their similarity to the input claim (Shaar, Babulkov et al., 2020). This task comes right after the claim detection component, to determine if the claim is new, and can help avoid the need for running the claim validation component for a particular claim when it is found in the database.

There are two released datasets: one based on PolitiFact and the other based on Snopes. Initial explorations were conducted on using BM25 (Robertson et al., 1994) and BERT-based models respectively as well as building a SVM reranker with features from both approaches (Shaar, Babulkov et al., 2020). Otherwise, CLEF2020-CheckThat! held a shared task on Verified Claim Retrieval which uses the Snopes dataset. While the baseline system is a simple BM25 system, shared task participants explored various scoring functions, including unsupervised approaches such as Terrier and Elastic Search scores, classic supervised models such as SVM and various BERT-based models (Shaar, Nikolov et al., 2020). Buster.ai, the winning team, fine-tune a RoBERTa (Liu, Ott, et al., 2019) model on the task which was first fine-tuned on other fact-checking datasets (Bouziane et al., 2020). Team UNIPI-NLE, achieving close performance to the winning team, performed two cascade fine-tunings on a sentenceBERT (Reimers & Gurevych, 2019) model (Passaro et al., 2020).

3 | CLAIM VALIDATION

As a component of the automated fact-checking pipeline, claim validation is formulated as 'the assignment of a truth value to a claim made in a particular context' (Vlachos & Riedel, 2014).

3.1 | Approaches

In order to fulfil the task of claim validation, two different major approaches to verification have emerged: (1) the claim is verified against textual references such as documents from Wikipedia (Thorne et al., 2019; Thorne, Vlachos, Christodoulopoulos, & Mittal, 2018); (2) the claim is verified against existing knowledge bases (Shi & Weninger, 2016; Syed et al., 2019). Both approaches assume their references are reliable. The first approach may limit evidence to only trusted resource such as Wikipedia, fact-checking websites, peer-reviewed academic papers and government documents, achieving substantial coverage of information. However, the second approach faces bigger challenges in terms of coverage of reliable information. Existing knowledge

bases tend to be too small to cover sufficient information for claim validation purposes (Azmy et al., 2018; Mendes et al., 2012; Pellissier Tanon et al., 2020). Attempts have been made to automatically populate knowledge bases (Adel, 2018; Balog, 2018; Mesquita et al., 2019; Nakashole & Weikum, 2012) but this method has the risk of further introducing unreliable noise and makes it harder to maintain the knowledge bases. Due to its maturity and reliability, our survey focuses on the first approach.

There have been a number of shared tasks focused on claim validation in slightly different ways. One of the major differences is whether the final verification step is reliant on previously identified pieces of evidence (such as Wikipedia documents or scientific articles) or it is instead reliant on the stances expressed by users (e.g., by aggregating supporting or opposing stances towards a story in social media). Of those relying on evidence, well-known shared tasks include FEVER (Thorne, Vlachos, Christodoulopoulos, & Mittal, 2018) and SCIVER (Wadden et al., 2020), both of which perform different forms of evidence retrieval first and then perform claim validation based on that evidence. On the other hand, both UKP Snopes (Hanselowski et al., 2019) and RumourEval (Derczynski et al., 2017; Gorrell et al., 2018) proposed to tackle the task by retrieving texts relevant to a story, determining the stance of those texts afterwards, to ultimately classify the veracity value of the story.

3.2 | Datasets

The NLP community has developed valuable datasets to progress research in automated claim validation, though with common issues of being synthetic and imbalanced. As shown in Table 2, recent datasets are not only growing in size, but they also attempt to capture naturally occurring sentences, include context and metadata, cover different domains and offer evidence chains.

3.3 | Evidence retrieval

Evidence retrieval is conventionally addressed in two steps: document retrieval and rationale selection. Document retrieval is the task of retrieving relevant documents that supports the prediction of a claim's veracity. Rationale selection is the task of selecting directly relevant sentences out of the retrieved documents to get final supporting evidence for claim verification.

3.3.1 | Document retrieval

Deeply influenced by IR research, the majority of work in the literature addresses it as a ranking problem consisting in retrieving the top k documents. Various combination of Named Entities, Noun Phrases and Capitalised Expressions from the claim were used to query search APIs such as Google or Wikipedia and search servers (Thorne, Vlachos, Cocarascu, et al., 2018), when participating in the FEVER shared task. Metadata, such as page viewership statistics is helpful to rank webpages (Nie et al., 2019). However, when search engines are not available, such as in the SCIVER shared task, the majority of effort goes into exploring similarity metrics that are used as a proxy to determine the documents' relevance to a claim. TF-IDF similarity is a common baseline (Malon, 2018; Wadden et al., 2020) and BM25 (Robertson et al., 1994) is demonstrated to be effective (Pradeep et al., 2021). When dealing with a specific domain, in-domain word

TABLE 2 Claim validation datasets

Name	Number of claims or claim-evidence pairs	Domains	Details
PolitiFact (Vlachos & Riedel, 2014)	106 claims	Politics	Very small; metadata and evidence of various forms
Emergent (Ferreira & Vlachos, 2016)	300 claims	News	Very small; 2595 associated documents
LIAR (Wang, 2017)	12,836 claims	Politics	Medium; metadata
Snopes (Popat et al., 2017)	4956 claims	Snopes website	Medium; 30 Google retrieved documents for each claim
FEVER (Thorne, Vlachos, Christodoulopoulos, & Mittal, 2018)	185,445 claims	Wikipedia	Big; associated Wikipedia evidence
LIAR-PLUS (Alhindi et al., 2018)	12,836 claims	Politics	Medium; automatically extracted justifications
Perspectrum (Chen, Khashabi, et al., 2019)	907 claims	Debates	Small; evidence and perspectives
UKP Snopes (Hanselowski et al., 2019)	6422 claims	Snopes website	Medium; associated evidence
MultiFC (Augenstein et al., 2019)	34,918 claims	Fact-checking websites	Medium; metadata and 10 Google retrieved webpages for each claim
Scifact (Wadden et al., 2020)	1409 claims	Scientific papers	Small; associated documents
PolitiHop (Ostrowski et al., 2020)	500 claims	Politics	Very small; evidence chains for multi–hop reasoning
WikiFactCheck-English (Sathe et al., 2020)	124,821 claims	Wikipedia	Big; context and evidence
Climate-FEVER (Diggelmann et al., 2021)	1535 claims	Climate	Medium; 7675 claim-evidence pairs with climate related claims verified against Wikipedia evidence
COVID-fact (Saakyan et al., 2021)	4086 claims	COVID-19	Medium; 1296 supported claims from r/COVID19 subreddit and 2790 automatically generated refuted claims
Vitamin-C (Schuster et al., 2021)	488,904 pairs	Wikipedia	Big; contrastive evidence from Wikipedia edits
FEVEROUS (Aly et al., 2021)	87,026 claims	Wikipedia	Biggest; evidence collected from both structured and unstructured information on whole Wikipedia

embeddings are also a promising option, for example, BioSentVec (Chen, Peng, & Lu, 2019) for the SCIFACT dataset (Li et al., 2021).

Instead of completely relying on unsupervised methods, improvements have been achieved by reranking based on supervised learning on top of a large number of retrieved documents (Pradeep et al., 2021).

3.3.2 | Rationale selection

Keyword matching, sentence similarity scoring and supervised ranking are common approaches to rationale selection (Thorne, Vlachos, Cocarascu, et al., 2018). Similar to document retrieval, attempts typically use one of these approaches or a combination of them to get a ranking score and select top k sentences as rationale with a manual choice of the k value (Pradeep et al., 2021).

Most of studies in the literature conduct evidence retrieval by addressing document retrieval and rationale selection in a pipeline manner, which ignores valuable information across sentences.

3.4 | Claim verification

Claim verification is commonly addressed as a text classification task by NLP researchers. Given a claim under investigation and its retrieved evidence, models need to reach a verdict of the claim, which may be 'SUPPORT', 'CONTRADICTION' or 'NOT ENOUGH INFORMATION'. Some other datasets (Hanselowski et al., 2019; Wang, 2017) include other labels such as 'mostly-true', 'half-true', 'pants-fire', 'most false', 'most true' and 'other', whose finer granularity is more difficult to tackle through automated means and are sometimes collapsed into fewer labels. An important observation here is the difference in the types of labels used by different studies. Some studies rely on truth values (e.g., true, false and half-true), determining the veracity value of a claim. Others refer to the concept of support instead (i.e., support and contradict), which instead determine whether there is an agreement between the claim and the reference. The latter avoids making an explicit connection with truthfulness, looking instead at the alignment of a claim with respect to a given reference.

The task of claim verification may be essentially addressed as a recognising textual entailment (RTE) task, that is, 'deciding, given two text fragments, whether the meaning of one text is entailed (can be inferred) from another text' (Dagan et al., 2009) or a natural language inference (NLI) task, that is, 'characterising and using semantic concepts of entailment and contradiction in computational systems' (Bowman et al., 2015).

Given that claim verification is predominantly addressed as a RTE or NLI task, we present a brief overview of them below. The RTE task, which dates back to 2005, focuses on detecting whether the hypothesis h is entailed by a given text t or not, which corresponds to 'SUPPORT' or not. Proposed models may take a linguistic approach, a statistical approach, a machine learning approach or a hybrid version of them. The NLI task, equipped with many large-scale labelled datasets, has powered large neural models to be the dominant approach. State-of-the-art models are large pre-trained language models that are fine-tuned on large NLI datasets.

3.4.1 | Recognising textual entailment

Textual entailment is defined as a relation between a text T and a hypothesis H. Formal semantics defines that a text t entails hypothesis h if h is true in every possible circumstance where t is true (Chierchia & McConnell-Ginet, 2000). This definition, as well as many other formal linguistic theories, is theoretically sound but practically too rigid to handle uncertainty. In practical NLP context, we define entailment to include cases where the truth value of hypothesis h is highly plausible given text t, rather than absolutely certain (Dagan et al., 2009). In other words, 'text t entails hypothesis h if, typically, a human reading t would infer that h is most likely true' (Bar-Haim et al., 2014). In contrast to a formal theoretical definition, this somewhat informal definition of entailment allows and requires common sense background knowledge.

The task of RTE started as a two-way classification of deciding whether hypothesis h is entailed/supported by text t or not (Bar-Haim et al., 2006; Dagan et al., 2005; Giampiccolo et al., 2007). After the notion of 'contradiction', that is, 'the negation of the hypothesis h is entailed by the text t' is introduced (de Marneffe et al., 2008), the RTE task became a three-way classification task of predicting labels of a text pair out of 'ENTAILMENT', 'CONTRADICTION' and 'UNKNOWN' (Giampiccolo et al., 2008).

Driven by a yearly RTE challenge from 2005 to 2011, the NLP community developed some useful datasets for the task, specifically the RTE1–RTE7 datasets. Despite being relatively small and imbalanced, these datasets enabled developing and evaluating of various approaches. While lexical-based and syntax-based approaches struggled to achieve good results (Bar-Haim et al., 2014; Dagan et al., 2009), machine learning approaches achieved reasonable performance, often combined with logical or probabilistic methods.

One of the earlier models attempted to feed deep semantic features generated by first-order theorem prover and finite model builder into a machine learning classified to make predictions (Bos & Markert, 2006). Surprisingly, the deep semantic features failed to outperform shallow semantic features. This is likely due to the models' naïve and rigid representation of sentences, lack of background knowledge and flawed sample distribution of the dataset.

Another intuitive approach is to first induce representations of text snippets into a hierarchical knowledge representation and then use a sound inferential mechanism to prove semantic entailment (de Salvo Braz et al., 2006). Despite its sound and tangible system design, this model only achieved an overall accuracy of 65.9%.

Furthermore, the NatLog system deals with the problem in three stages. It first conducts linguistic analysis, then aligns the dependency graph of the text t and the hypothesis h, finally uses a decision tree classifier to perform entailment inference based on antonyms, polarity, graph structure and semantic relations (Chambers et al., 2007). This NatLog system trades low recall (31.71 on RTE3 test set) for higher precision (68.06 on RTE3 test set).

To help address the low recall achieved by first-order rules, the class of pair feature spaces was introduced (Zanzotto et al., 2009). It allowed the model to enrich the sentence-pair with 'placeholders' and then generate first-order rewrite rules to relax the rigidness. This model achieved around 68% overall accuracy on RTE3.

Moreover, COGEX developed a system that first transforms the text into three-layered semantically-rich logic form representations, then generates a set of linguistic and world knowledge axioms, and searches for a proof of entailment (Tatu & Moldovan, 2007). This system achieved an overall accuracy of 72.25%.

Overall, many inspiring hybrid models of logical inference and machine learning methods were developed for RTE challenges. Though they did not achieve perfect performance, we believe

they have great potential once equipped with better text representations and more powerful neural models.

3.4.2 | Natural language inference

More recently, NLI is proposed as 'the problem of determining whether a natural language hypothesis *h* can reasonably be inferred from a given premise *p*' (Bowman et al., 2015; MacCartney & Manning, 2009). Noticeably, the definition of NLI is very similar to RTE and researchers tend to mention them together when addressing the problem.

Despite that, NLI datasets have improvements over RTE datasets. Earlier RTE datasets, published before the notion of 'CONTRADICTION' attracted enough attention, only have binary labelling of 'ENTAILMENT'. In contrast, NLI datasets all include three-way labelling that includes 'ENTAILMENT', 'CONTRADICTION' and 'UNKNOWN'. Furthermore, recent NLI datasets have larger size, more balanced label distribution and cover various domains. Table 2 presents NLI datasets that are potentially useful for claim verification.

With their large size and balanced design, NLI datasets have powered large neural network models, which has become the dominant approach. The common practice is to fine-tune a large pre-trained language model on the target NLI dataset, which may or may not be coupled with small task-specific techniques. Compared with traditional approaches, this approach improved text representations, achieved better generalisability and allowed more complex computing without relying on hand-crafted rules.

Current state-of-the-art models on NLI datasets are BERT (Devlin et al., 2019), RoBERTa (Liu, Ott, et al., 2019), MT-DNN (Liu, He, et al., 2019), ALBERT (Lan et al., 2020) and T5 (Raffel et al., 2020).

4 | DISCUSSION AND CHALLENGES

In this section, we discuss current progress in each of the components of the automated fact-checking task, as well as highlight the main open challenges.

4.1 | Claim detection

4.1.1 | Conceptual definition of claim

The definition of claim check-worthiness is brief (Allein & Moens, 2020). Full Fact describe it as 'an assertion about the world that can be checked'. In contrast (Konstantinovskiy et al., 2021), mentioned this definition is not enough to decide whether this claim is worthy for check or not. Similarly, Berendt et al. (2020) declared that not every factual claim will be verified by fact checkers.

4.1.2 | Narrow domains

Claims in the political domain are dominating the interest of journalists and researchers, as can be seen in existing datasets. As an example Wright and Augenstein (2020) investigated the development of a claim check-worthiness detection method that would consistently perform over different domains, in this case rumours on Twitter, Wikipedia citations and political speeches. However, the method showed important challenges in trying to perform well across domains. Recent research in claim detection has expanded to focus on health claims too, particularly owing to COVID-19.

4.1.3 | Annotation issues

Labelling of sentences as claims or non-claims is generally done manually by non-experts (see Table 1). An alternative to this is to derive labels from previously fact-checked claims collected from fact-checking websites. The main caveat of this approach is that fact-checking websites only list claims, rather than non-claims, which means that one needs to develop models that only leveraged instances of the positive class, that is, positive unlabelled learning (Wright & Augenstein, 2020).

4.1.4 | Imbalanced datasets

The majority of datasets are imbalanced where not check-worthy claims outnumber check-worthy claims. While this is possibly due to the nature of the task, existing models can have a tendency to overfit due to this imbalance, which calls for more research to tackle the problem. For example, in the CheckThat! Lab 2020 (Williams et al., 2020), attempted to mitigate the problem of overfitting by retraining the model by resampling the larger number of positive instances that were augmenting data through translation between Arabic and English (Williams et al., 2020).

4.2 | Claim validation

Despite the noticeable progress, current automated claim validation systems also face unique challenges and desire improvements over several key aspects: datasets quality, system integrity and model interpretability.

4.2.1 | Datasets quality

State-of-the-art systems heavily rely on training large language models, which require large-scale, high-quality, labelled datasets that are expensive and may be unrealistic for specific domains. Despite the great contributions recent datasets have made, they tend to be imbalanced and somewhat synthetic, which are not ideal for model training. We believe future high-quality datasets will continue to help progress the field.

4.2.2 | System scalability and integrity

Proposed automated claim validation systems cover a range of relevant tasks. Though a few of them try to jointly handle rationale selection and claim verification (Hidey & Diab, 2018; Li et al., 2021), most of them are pipeline systems that deal with subtasks separately. Improved scalability and integrity is desired.

Large pre-trained language models, the current dominant approach of various relevant tasks, requires lots of computing resources to train and inference. The scalability and accessibility of the proposed systems remain inferior.

Otherwise, increased system integrity is desired. Pipeline design has its inevitable disadvantage: downstream components can only make inferences on upstream results and errors accumulate throughout the pipeline. For instance, a claim verification component that takes the retrieved evidence and the detected claim as input will perform poorly with low-quality evidence or claims that are not checkable. Furthermore, the popular three-way label prediction approach is not the best approach for claim verification. Models struggle particularly to predict contradiction relation due to a lack of training data in this class, which accumulates errors across classes. For example, a model may predict a claim to be 'NO INFO' while it should be 'CONTRADICT', which makes it a false positive for the 'NO INFO' class and a true negative for the 'CONTRADICT' class. Preliminary research splitting the three-way classification into two binary classifications (Zeng & Zubiaga, 2021) is likely to help avoid such errors. Moreover, current approaches leave no space for aggregating evidence across sentences.

We believe a more compact overall system design is desired for automated claim validation such that it handles subtasks in a systematic way. We believe a promising direction is to train a model to learn all involved tasks in a multitask learning manner so that it may optimise for better overall performance.

4.2.3 | Model interpretability

Neural networks are robust but struggle with interpretability and generalisability (Duan et al., 2020), which is of particular importance for automated claim validation. Underwhelming model interpretability may induce an increased probability of models making the right prediction based on the wrong evidence. In contrast, symbolic systems that are unfortunately fragile and inflexible have strong interpretability and abstraction. Naturally, building a neural-symbolic system that integrates neural networks with symbolic logic becomes an interesting direction. In a nutshell, *neural-symbolic systems = connectionist machine + logical abstractions* (Besold et al., 2017).

Researchers have proposed various architectures that incorporate first-order logic into neural networks. A recent study proposed a general framework capable of enhancing neural networks with declarative first-order logic (Hu et al., 2016). Another study explored a symbolic intermediate representation for neural surface realisation (Elder et al., 2019), that is similar to first-order logic. Moreover, a recent attempt adapted module networks to model natural logic operations, which is enhanced with a memory component to model contextual information (Feng et al., 2020). Furthermore, RuleNN (Sen et al., 2020) is developed to tackle sentence classification where models are in the form of first-order logic, and achieved performance, that is comparable to some neural models.

Neural-symbolic methods have a fascinating potential of attaining interpretability from symbolic models and robustness from neural models makes. We believe that designing and implementing neural-symbolic methods for various tasks of automated fact-checking is promising and of particular interest to our society.

5 | RELATED TASKS

There are some other popular tasks in NLP which are also related to the accuracy, verifiability and credibility of information, which we briefly discuss next as topics recommended for further reading:

5.1 | Fake news detection

It is the task of determining whether a news article on the web is accurate or not (Shu et al., 2017). Proposed classification approaches are typically centred on shallow features of the articles: n-grams, characters, stop words, POS tags, readability scores, term frequency, etc. Some more advanced approaches use additional metadata. However, these approaches are more likely to merely capture patterns of different article styles, rather than to sensibly distinguish reliable and unreliable articles (Hanselowski, 2020).

5.2 | Rumour detection

It is the task of identifying unverified reports circulating on social media. Predictions are typically made on language subjectivity and metadata on social media (Zubiaga et al., 2018). Despite the relevance of these features, the truth value of a claim does not directly depend on these features.

5.3 | Clickbait detection

Being considerably different from automated fact-checking, clickbait detection does not require external evidence. Approaches with relatively shallow linguistic features (Chakraborty et al., 2016; Chen et al., 2015; Potthast et al., 2016) have yielded reasonable performance.

5.4 | Commonsense reasoning

To perform commonsense reasoning (Storks et al., 2019), the model needs to be able to do reasoning beyond the explicit information given in sentence pairs, which is highly valued in automated fact-checking (Thorne & Vlachos, 2018). As a new frontier of artificial intelligence, novel studies have investigated learned knowledge in pre-trained language models, commonsense integration from external knowledge bases, symbolic knowledge incorporation, etc. However, these tasks are currently under investigation and the field calls for major breakthroughs. For more information, we refer to a recent survey (Storks et al., 2019) and a tutorial (Sap et al., 2020).

6 | CONCLUSIONS

In this article, we present a survey on automated fact-checking with special focus on claim detection and claim validation. Substantial progress has been made by applying large pre-trained language models through designed pipelines, but numerous open challenges still need further research. Claim Detection faces challenges from conceptual definition, narrow domains, annotation issues and imbalanced datasets. In addition, improvements over datasets quality, system integrity and model interpretability are desired for claim validation.

CONFLICT OF INTEREST
The authors acknowledge that there are no conflicts of interest to declare.

ORCID
Xia Zeng https://orcid.org/0000-0002-7535-6904
Arkaitz Zubiaga https://orcid.org/0000-0003-4583-3623

ENDNOTE
[1] The task of claim validation is referred to as fact-checking by some papers in the literature.

REFERENCES
Abdelali, A., Darwish, K., Durrani, N., & Mubarak, H. (2016). Farasa: A fast and furious segmenter for Arabic. In *Proceedings of the 2016 conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (pp. 11–16). Qatar Computing Research Institute.

Adel, H. (2018). *Deep learning methods for knowledge base population* [PhD Thesis]. Ludwig-Maximilians-Universität München.

Alhindi, T., Petridis, S., & Muresan, S. (2018). Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)* (pp. 85–90). Association for Computational Linguistics.

Allein, L., & Moens, M.-F. (2020). Checkworthiness in automatic claim detection models: Definitions and analysis of datasets. In *Multidisciplinary International Symposium on Disinformation in open online media* (pp. 1–17). Springer.

Aly, R., Guo, Z., Schlichtkrull, M., Thorne, J., Vlachos, A., Christodoulopoulos, C., Cocarascu, O., & Mittal, A. (2021). FEVEROUS: Fact extraction and VERification over unstructured and structured information. *arXiv:2106.05707 [cs]*.

Arslan, F., Hassan, N., Li, C., & Tremayne, M. (2020). A benchmark dataset of check-worthy factual claims. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*(1), 821–829. https://ojs.aaai.org/index.php/ICWSM/article/view/7346

Atanasova, P., Barron-Cedeno, A., Elsayed, T., Suwaileh, R., Zaghouani, W., Kyuchukov, S., Martino, G. D. S., & Nakov, P. (2018). Overview of the CLEF-2018 CheckThat! Lab on automatic identification and verification of political claims. Task 1: Check-worthiness. *CLEF 2018 Working Notes*. http://ceur-ws.org/Vol-2125/

Atanasova, P., Nakov, P., Karadzhov, G., Mohtarami, M., & Da San Martino, G. (2019). *Overview of the CLEF-2019 CheckThat! Lab: Automatic identification and verification of claims. Task 1: Checkworthiness*. Springer.

Atanasova, P., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Karadzhov, G., Mihaylova, T., Mohtarami, M., & Glass, J. (2019). Automatic fact-checking using context and discourse information. *Journal of Data and Information Quality*, *11*(3), 1–27.

Augenstein, I., Lioma, C., Wang, D., Chaves Lima, L., Hansen, C., Hansen, C., & Simonsen, J. G. (2019). MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. 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)* (pp. 4685–4697). Association for Computational Linguistics.

Azmy, M., Shi, P., Lin, J., & Ilyas, I. (2018). Farewell Freebase: Migrating the Simple Questions dataset to DBpedia. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 2093–2103). Association for Computational Linguistics.

Babakar, M., & Moy, W. (2016). *The state of automated factchecking* (Technical Report Full Fact). https://fullfact.org/media/uploads/full_fact-the_state_of_automated_factchecking_aug_2016.pdf

Balog, K. (2018). Populating knowledge bases. In K. Balog (Ed.), *Entity-oriented search, the information retrieval series* (pp. 189–222). Springer International Publishing.

Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., & Giampiccolo, D. (2006). The second PASCAL recognising textual entailment challenge. In *Proceedings of the second PASCAL Challenges Workshop on Recognising Textual Entailment* (pp. 177–190). LNCS.

Bar-Haim, R., Dagan, I., & Szpektor, I. (2014). Benchmarking applied semantic inference: The PASCAL recognising textual entailment challenges. In N. Dershowitz & E. Nissan (Eds.), *Language, culture, computation. Computing—Theory and technology: Essays dedicated to Yaacov Choueka on the occasion of his 75th birthday, Part I, lecture notes in computer science* (pp. 409–424). Springer.

Barron-Cedeno, A., Elsayed, T., Nakov, P., Martino, G., Hasanain, M., Suwaileh, R., Haouari, F., Babulkov, N., Hamdan, B., Nikolov, A., Shaar, S., & Sheikh Ali, Z. (2020). Overview of CheckThat! 2020: Automatic identification and verification of claims in social media. *CLEF 2020 Working Notes*. https://link.springer.com/chapter/10.1007/978-3-030-58219-7_17

Berendt, B., Burger, P., Hautekiet, R., Jagers, J., Pleijter, A., & Aelst, P. V. (2020). *FactRank: Developing automated claim detection for Dutch-language fact-checkers.* https://people.cs.kuleuven.be/~bettina.berendt/FactRank/FactRank.pdf

Besold, T. R., Garcez, A. d., Bader, S., Bowman, H., Domingos, P., Hitzler, P., Kuehnberger, K.-U., Lamb, L. C., Lowd, D., Lima, P. M. V., de Penning, L., Pinkas, G., Poon, H., & Zaverucha, G. (2017). Neural-symbolic learning and reasoning: A survey and interpretation. *arXiv:1711.03902 [cs]*.

Bos, J., & Markert, K. (2006). When logical inference helps determining textual entailment (and when it doesn't). In *Proceedings of the Second PASCAL RTE Challenge*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.138.5503&rep=rep1&type=pdf

Bouziane, M., Perrin, H., Cluzeau, A., Mardas, J., & Sadeq, A. (2020). *Team Buster.ai at CheckThat! 2020: Insights and Recommendations to Improve fact-checking* (p. 12). http://ceur-ws.org/Vol-2696/paper_134.pdf

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 632–642). Association for Computational Linguistics.

Chakraborty, A., Paranjape, B., Kakarla, S., & Ganguly, N. (2016, August 18–21). Stop clickbait: Detecting and preventing clickbaits in online news media. In R. Kumar, J. Caverlee, & H. Tong (Eds.), *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*. IEEE Computer Society.

Chambers, N., Cer, D., Grenager, T., Hall, D., Kiddon, C., MacCartney, B., de Marneffe, M.-C., Ramage, D., Yeh, E., & Manning, C. D. (2007). Learning alignments and leveraging natural logic. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing* (pp. 165–170). Association for Computational Linguistics.

Chen, Q., Peng, Y., & Lu, Z. (2019). BioSentVec: Creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)* (pp. 1–5).

Chen, S., Khashabi, D., Yin, W., Callison-Burch, C., & Roth, D. (2019). Seeing things from a different angle: Discovering diverse perspectives about claims (Long and Short Papers). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 542–557). Association for Computational Linguistics.

Chen, Y., Conroy, N. J., & Rubin, V. L. (2015). Misleading online content: Recognizing clickbait as "false news". In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection WMDD '15* (pp. 15–19). Association for Computing Machinery.

Chierchia, G., & McConnell-Ginet, S. (2000). *Meaning and grammar: An introduction to semantics*. MIT Press.

Cohen, S., Li, C., Yang, J., & Yu, C. (2011). Computational journalism: A call to arms to database researchers. In *CIDR 2011, Fifth Biennial Conference on Innovative Data Systems Research, Asilomar, CA, United States* (pp. 148–151). https://asu.pure.elsevier.com/en/publications/computational-journalism-a-call-to-arms-to-database-researchers

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 670–680). Association for Computational Linguistics.

Dagan, I., Dolan, B., Magnini, B., & Roth, D. (2009). Recognizing textual entailment: Rational, evaluation and approaches. In *Natural Language Engineering* (Vol. 15, Special issue 4, pp. i–xvii). Cambridge University Press.

Dagan, I., Glickman, O., & Magnini, B. (2005). The PASCAL recognising textual entailment challenge. In J. Quiñonero-Candela, I. Dagan, B. Magnini, & F. d'Alché Buc (Eds.), *Machine learning challenges. Evaluating predictive uncertainty, visual object classification, and recognising tectual entailment, lecture notes in computer science* (pp. 177–190). Springer.

de Marneffe, M.-C., Rafferty, A. N., & Manning, C. D. (2008). Finding contradictions in text. In *Proceedings of ACL-08: HLT* (pp. 1039–1047). Association for Computational Linguistics.

de Salvo Braz, R., Girju, R., Punyakanok, V., Roth, D., & Sammons, M. (2006). *An inference model for semantic entailment in natural language* (pp. 261–286). Springer.

Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Hoi, G. W. S., & Zubiaga, A. (2017). Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 69–76). Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)* (pp. 4171–4186). Association for Computational Linguistics.

Diggelmann, T., Boyd-Graber, J., Bulian, J., Ciaramita, M., & Leippold, M. (2021). CLIMATE-FEVER: A dataset for verification of real-world CLIMATE claims. *arXiv:2012.00614 (cs)*.

Duan, N., Tang, D., & Zhou, M. (2020). Machine reasoning: Technology, dilemma and future. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts* (pp. 1–6). Association for Computational Linguistics.

Elder, H., Foster, J., Barry, J., & O'Connor, A. (2019). Designing a symbolic intermediate representation for neural surface realization. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation* (pp. 65–73). Association for Computational Linguistics.

Elsayed, T., Nakov, P., Barrón-Cedeño, A., Hasanain, M., Suwaileh, R., Martino, G., & Atanasova, P. (2019). Overview of the CLEF-2019 CheckThat! Lab: Automatic identification and verification of claims. In *Experimental IR meets multilinguality, multimodality, and interaction* (pp. 301–321). Springer.

Feng, Y., Zheng, Z., Liu, Q., Greenspan, M., & Zhu, X. (2020). Exploring end-to-end differentiable natural logic modeling. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 1172–1185). International Committee on Computational Linguistics.

Ferreira, W., & Vlachos, A. (2016). Emergent: A novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1163–1168). Association for Computational Linguistics.

Francis, D., & Fact, F. (2016). *Fast & furious fact check challenge*. https://www.herox.com/factcheck/teams

Gencheva, P., Nakov, P., Màrquez, L., Barrón-Cedeño, A., & Koychev, I. (2017). A context-aware approach for detecting worth-checking claims in political debates. In *Proceedings of the International Conference Recent Advances in Natural Language Processing* (pp. 267–276). INCOMA Ltd.

Giampiccolo, D., Dang, H., Magnini, B., Dagan, I., Cabrio, E., & Dolan, W. (2008). *The fourth PASCAL recognizing textual entailment challenge*. TAC.

Giampiccolo, D., Magnini, B., Dagan, I., & Dolan, B. (2007). The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing* (pp. 1–9). Association for Computational Linguistics.

Gorrell, G., Bontcheva, K., Derczynski, L., Kochkina, E., Liakata, M., & Zubiaga, A. (2018). Determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 845–854). Association for Computational Linguistics.

Hanselowski, A. (2020). *A machine-learning-based pipeline approach to automated fact-checking* (PhD Thesis). Technische Universität.

Hanselowski, A., Stab, C., Schulz, C., Li, Z., & Gurevych, I. (2019). A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)* (pp. 493–503). Association for Computational Linguistics.

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

Hasanain, M., Haouari, F., Suwaileh, R., Ali, Z. S., Hamdan, B., Elsayed, T., Barrón-Cedeño, A., Martino, G. D. S., & Nakov, P. (2020). *Overview of CheckThat! 2020 Arabic: Automatic identification and verification of claims in social media*. CLEF.

Hassan, N., Arslan, F., Li, C., & Tremayne, M. (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* (pp. 1803–1812). Association for Computing Machinery.

Hassan, N., Li, C., & Tremayne, M. (2015). Detecting check-worthy factual claims in presidential debates. In J. Bailey, A. Moffat, C. C. Aggarwal, M. de Rijke, R. Kumar, V. Murdock, T. K. Sellis, & J. X. Yu (Eds.), *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015* (pp. 1835–1838). Association for Computing Machinery.

Hassan, N., Zhang, G., Arslan, F., Caraballo, J., Jimenez, D., Gawsane, S., Hasan, S., Joseph, M., Kulkarni, A., Nayak, A. K., Sable, V., Li, C., & Tremayne, M. (2017). ClaimBuster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, *10*(12), 1945–1948.

Hidey, C., & Diab, M. (2018). Team SWEEPer: Joint sentence extraction and fact checking with pointer networks. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)* (pp. 150–155). Association for Computational Linguistics.

Hu, Z., Ma, X., Liu, Z., Hovy, E., & Xing, E. (2016). Harnessing deep neural networks with logic rules. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2410–2420). Association for Computational Linguistics.

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). Association for Computational Linguistics.

Kartal, Y. S., & Kutlu, M. (2020). TrClaim-19: The first collection for Turkish check-worthy claim detection with annotator rationales. In *Proceedings of the 24th Conference on Computational Natural Language Learning* (pp. 386–395). Association for Computational Linguistics.

Konstantinovskiy, L., Price, O., Babakar, M., & Zubiaga, A. (2021). Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digital Threats: Research and Practice*, *2*(2), 1–16.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). Albert: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia*. OpenReview.net.

Li, X., Burns, G., & Peng, N. (2021). A paragraph-level multi-task learning model for scientific fact-verification. *Proceedings of the Workshop on Scientific Document Understanding*.

Liu, X., He, P., Chen, W., & Gao, J. (2019). Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4487–4496). https://aclanthology.org/P19-1441/

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:1907.11692 (cs)*.

MacCartney, B., & Manning, C. D. (2009). An extended model of natural logic. In *Proceedings of the Eight International Conference on Computational Semantics* (pp. 140–156). Association for Computational Linguistics.

Malon, C. (2018). Team Papelo: Transformer networks at FEVER. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)* (pp. 109–113). Association for Computational Linguistics.

Mendes, P., Jakob, M., & Bizer, C. (2012). DBpedia: A multilingual cross-domain knowledge base. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 1813–1817). European Language Resources Association (ELRA).

Mesquita, F., Cannaviccio, M., Schmidek, J., Mirza, P., & Barbosa, D. (2019). KnowledgeNet: A benchmark dataset for knowledge base population. 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)* (pp. 749–758). Association for Computational Linguistics.

Nakashole, N., & Weikum, G. (2012). Real-time population of knowledge bases: Opportunities and challenges. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction (AKBC-WEKEX)* (pp. 41–45). Association for Computational Linguistics.

Nakov, P., Barrón-Cedeño, A., Elsayed, T., Suwaileh, R., Màrquez, L., Zaghouani, W., Atanasova, P., Kyuchukov, S., & Martino, G. (2018). Overview of the CLEF-2018 CheckThat! Lab on automatic identification and verification of political claims. In *9th International Conference of the CLEF Association* (pp. 372–387). CLEF.

Nakov, P., Da San Martino, G., Elsayed, T., Barrón-Cedeno, A., Míguez, R., Shaar, S., Alam, F., Haouari, F., Hasanain, M., Babulkov, N., & Nikolov, A. (2021). The CLEF-2021 CheckThat! Lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *European Conference on Information Retrieval* (pp. 639–649). Springer.

Nie, Y., Chen, H., & Bansal, M. (2019). Combining fact extraction and verification with neural semantic matching networks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27–February 1, 2019* (pp. 6859–6866). AAAI Press.

Ostrowski, W., Arora, A., Atanasova, P., & Augenstein, I. (2020). Multi-hop fact checking of political claims. *arXiv:2009.06401 [cs]*.

Passaro, L. C., Bondielli, A., & Lenci, A. (2020). UNIPI-NLE at CheckThat! 2020: Approaching fact checking from a sentence similarity perspective through the lens of transformers. *CLEF 2020 Working Notes*.

Patwari, A., Goldwasser, D., & Bagchi, S. (2017). Tathya: A multi-classifier system for detecting check-worthy statements in political debates. In E. Lim, M. Winslett, M. Sanderson, A. W. Fu, J. Sun, J. S. Culpepper, E. Lo, J. C. Ho, D. Donato, R. Agrawal, Y. Zheng, C. Castillo, A. Sun, V. S. Tseng, & C. Li (Eds.), *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017* (pp. 2259–2262). Association for Computing Machinery.

Pellissier Tanon, T., Weikum, G., & Suchanek, F. (2020). Yago 4: A reason-able knowledge base. In A. Harth, S. Kirrane, A.-C. Ngonga Ngomo, H. Paulheim, A. Rula, A. L. Gentile, P. Haase, & M. Cochez (Eds.), *The semantic web. Lecture notes in computer science* (pp. 583–596). Springer.

Pomerleau, D., & Rao, D. (2017). *The fake news challenge: Exploring how artificial intelligence technologies could be leveraged to combat fake news*. http://www.fakenewschallenge.org/

Popat, K., Mukherjee, S., Strötgen, J., & Weikum, G. (2017). Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion, WWW '17 Companion* (pp. 1003–1012). Republic and Canton of Geneva, CHE International World Wide Web Conferences Steering Committee.

Potthast, M., Köpsel, S., Stein, B., & Hagen, M. (2016). *Clickbait detection*. Springer.

Pradeep, R., Ma, X., Nogueira, R., & Lin, J. (2021). Scientific claim verification with VERT5ERINI. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis* (pp. 94–103). Association for Computational Linguistics.

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 version 3. *Journal of Machine Learning Research*, *21*, 1–67. https://www.jmlr.org/beta/papers/v21/20-074.html

Reimers, N., & Gurevych, I. (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) (pp. 3982–3992). Association for Computational Linguistics.

Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M., & Gatford, M. (1994). *Okapi at TREC-3. undefined.* TREC.

Saakyan, A., Chakrabarty, T., & Muresan, S. (2021). COVID-Fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. *arXiv:2106.03794 [cs].*

Sap, M., Shwartz, V., Bosselut, A., Choi, Y., & Roth, D. (2020). Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts* (pp. 27–33). Association for Computational Linguistics.

Sathe, A., Ather, S., Le, T. M., Perry, N., & Park, J. (2020). Automated fact-checking of claims from Wikipedia. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 6874–6882). European Language Resources Association.

Schuster, T., Fisch, A., & Barzilay, R. (2021). Get your vitamin C! Robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 624–643). Association for Computational Linguistics.

Sen, P., Danilevsky, M., Li, Y., Brahma, S., Boehm, M., Chiticariu, L., & Krishnamurthy, R. (2020). Learning explainable linguistic expressions with neural inductive logic programming for sentence classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4211–4221). Association for Computational Linguistics.

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). Association for Computational Linguistics.

Shaar, S., Nikolov, A., Babulkov, N., Alam, F., Barron-Cedeno, A., Elsayed, T., Hasanain, M., Suwaileh, R., & Haouari, F. (2020). *Overview of CheckThat! 2020 English: Automatic identification and verification of claims in social media* (p. 24). CLEF.

Shi, B., & Weninger, T. (2016). Discriminative predicate path mining for fact checking in knowledge graphs. *Knowledge-Based Systems*, *104*, 123–133. https://www.sciencedirect.com/science/article/pii/S0950705116300570

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36.

Storks, S., Gao, Q., & Chai, J. Y. (2019). Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches version 1. *arXiv:1904.01172 [cs].*

Syed, Z. H., Röder, M., & Ngomo, A.-C. N. (2019). Unsupervised discovery of corroborative paths for fact validation. In C. Ghidini, O. Hartig, M. Maleshkova, V. Svátek, I. Cruz, A. Hogan, J. Song, M. Lefrançois, & F. Gandon (Eds.), *The semantic web – ISWC 2019. Lecture notes in computer science* (pp. 630–646). Springer.

Tatu, M., & Moldovan, D. (2007). COGEX at RTE 3. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing* (pp. 22–27). Association for Computational Linguistics.

Thorne, J., & Vlachos, A. (2018). Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 3346–3359). Association for Computational Linguistics.

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). Fever: A large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 809–819). Association for Computational Linguistics.

Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., & Mittal, A. (2018). The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)* (pp. 1–9). Association for Computational Linguistics.

Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., & Mittal, A. (2019). The FEVER2.0 shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)* (pp. 1–6). Association for Computational Linguistics.

Vlachos, A., & Riedel, S. (2014). Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science* (pp. 18–22). Association for Computational Linguistics.

Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., & Hajishirzi, H. (2020). Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 7534–7550). Association for Computational Linguistics.

Wang, W. Y. (2017). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 422–426). Association for Computational Linguistics.

Williams, E., Rodrigues, P., & Novak, V. (2020). Accenture at CheckThat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models. *CLEF 2020 Working Notes*.

Wright, D., & Augenstein, I. (2020). Claim check-worthiness detection as positive unlabelled learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings* (pp. 476–488). Association for Computational Linguistics.

Zanzotto, F. M., Pennacchiotti, M., & Moschitti, A. (2009). A machine learning approach to textual entailment recognition. *Natural Language Engineering*, *15*, 551–582.

Zeng, X., & Zubiaga, A. (2021). QMUL-SDS at SCIVER: Step-by-step binary classification for scientific claim verification. In *Proceedings of the Workshop on Scientific Document Processing* (pp. 116–123). Association for Computational Linguistics.

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

AUTHOR BIOGRAPHIES

Xia Zeng is a PhD candidate on Computer Science at Queen Mary University of London, where she is a member of the Social Data Science lab. She holds a Master's degree on Linguistics with Distinction from University College London. Previously she was a visiting student at UC San Diego and a research assistant at TU Darmstadt. She is interested in applying natural language processing (NLP) methods and linguistic theories on claim validation.

Amani S. Abumansour is a computer science PhD candidate at Queen Mary University of London, where she is taking a part in the social data science lab. She is also a lecturer in the computer science department at Taif University, Saudi Arabia. She has a master degree in computer science from California Lutheran University, USA. She is interested in utilizing Machine Learning and Natural Language Processing techniques for claim check-worthiness detection.

Arkaitz Zubiaga is a Lecturer at Queen Mary University of London, where he leads the Social Data Science lab. He is also a co-lead of the Social Data Science interest group at the Alan Turing Institute in London, UK. He is interested in linking online data with events in the real world, among others for tackling problematic issues on the Web and social media that can have a damaging effect on individuals or society at large, such as hate speech, misinformation, inequality, biases and other forms of online harm.