Automated Fact Checking:
Task formulations, methods and future directions

James Thorne
Department of Computer Science
University of Sheffield, UK
j.thorne@sheffield.ac.uk

Andreas Vlachos
Department of Computer Science
University of Sheffield, UK
a.vlachos@sheffield.ac.uk

Abstract

The recently increased focus on misinformation has stimulated research in fact checking, the task of assessing the truthfulness of a claim. Research in automating this task has been conducted in a variety of disciplines including natural language processing, machine learning, knowledge representation, databases, and journalism. While there has been substantial progress, relevant papers and articles have been published in research communities that are often unaware of each other and use inconsistent terminology, thus impeding understanding and further progress. In this paper we survey automated fact checking research stemming from natural language processing and related disciplines, unifying the task formulations and methodologies across papers and authors. Furthermore, we highlight the use of evidence as an important distinguishing factor among them cutting across task formulations and methods. We conclude with proposing avenues for future NLP research on automated fact checking.

Title and Abstract in Greek

Αναγνώριση Ψευδών Ισχυρισμών:
Ορισμοί, μέϑοδοι και κατευϑύνσεις για μελλοντική έρευνα

Το πρόσφατα αυξημένο ενδιαφέρον για το φαινόμενο της παραπληροφόρησης έχει κινητοποιήσει ερευνητικές προσπάϑειες για μεϑόδους αναγνώρισης ψευδών ισχυρισμών. Η έρευνα σε αυτοματοποιημένες μεϑόδους για το πρόβλημα αυτό διεξάγεται σε διάφορους επιστημονικούς κλάδους, όπως επεξεργασία φυσικής γλώσσας, μηχανική μάϑηση, αναπαράσταση γνώσης, βάσεις δεδομένων και δημοσιογραφία. Αν και έχει γίνει σημαντική πρόοδος, τα άρϑρα σχετικά με το αντικείμενο δημοσιεύονται στα συνέδρια και τα περιοδικά του κάϑε κλάδου και χρησιμοποιούν διαφορετική ορολογία, με συνέπεια να δυσχεραίνεται η περαιτέρω πρόοδος. Στο παρόν άρϑρο παρουσιάζουμε μια ανασκόπηση της έρευνας στον αυτόματη αναγνώριση ψευδών ισχυρισμών με κεντρικό άξονα την επεξεργασίας φυσικής γλώσσας, ενοποιώντας ορισμούς και μεϑόδους που έχουν προταϑεί στη βιβλιογραφία. Μια βασική διάκριση που διαχωρίζει τις ορισμούς και μεϑόδους που έχουν προταϑεί σε διαφορετικούς κλάδους είναι η χρήση ή μη αποδεικτικών στοιχείων. Η ανασκόπηση στο άρϑρο αυτό καταλήγει με κατευϑύνσεις για μελλοντική έρευνα στην επεξεργασία φυσικής γλώσσας σχετική με το πρόβλημα αυτό.

1 Introduction

The ability to rapidly distribute information on the internet presents a great opportunity for individuals and organizations to disseminate and consume large amounts of data on almost any subject matter. As many of the barriers required to publish information are effectively removed through the advent of social media, it is possible for content written by any individual or organization to reach an audience of hundreds of millions of readers (Allcott and Gentzkow, 2017). This enhanced ability of reaching audiences can be used to spread both true and false information, and has become an important concern as it is speculated that this has affected decisions such as public votes, especially since recent studies have shown that false information reached greater audiences (Vosoughi et al., 2018). Consequently, there has been an increased demand for fact checking of online content.

In the domain of journalism, fact checking is the task of assessing whether claims made in written or spoken language are true. This is a task that is normally performed by trained professionals: the fact checker must evaluate previous speeches, debates, legislation and published figures or known facts, and use these, combined with reasoning to reach a verdict. Depending on the complexity of the claim, this process may take from less than one hour to a few days (Hassan et al., 2015a). Due to the large volume of content to be checked, this process has been the subject of calls from the journalism community to develop tooling to automate parts of this task (Cohen et al., 2011; Babakar and Moy, 2016; Graves, 2018). The process of fact checking requires researching and identifying evidence, understanding the context of information and reasoning about what can be inferred from this evidence. The goal of automated fact checking is to reduce the human burden in assessing the veracity of a claim.

In this paper we survey automated fact checking work from the viewpoint of natural language processing (NLP) and related fields such as machine learning, knowledge representation, databases and social media analysis. While it is important to consider social aspects, including how fact checking is communicated and incorporated into social media platforms, these are considered out of scope for this survey. There is a vast body of related works with different motivations and approaches that we can learn from to develop natural language technologies for fact checking. However, because these are conducted in a siloed manner and often without wider awareness of the issue, there is inconsistency in the definitions and lack of communication between disciplines, impeding understanding and further progress. For example, the classification of whether an article title can be supported by content in the body has been interchangeably referred to as both fake news detection (Pomerleau and Rao, 2017), stance classification (Ferreira and Vlachos, 2016) and incongruent headline detection (Chesney et al., 2017).

In this survey we aim to unify the definitions presented in related works and identify common concepts, datasets and modelling approaches. In this process, we summarize key results, highlight limitations and propose some open research challenges. In particular, we detail the different types of evidence used by different approaches and how they affect the NLP requirements on the systems developed. For example, many approaches rely on textual entailment/natural language inference (Dagan et al., 2009), while others rely on knowledge base construction techniques (Ji and Grishman, 2011).

The structure of this survey is as follows: we first discuss fact checking in the context of journalism as this provides definitions and distinctions on key terminology that will be used throughout in the remainder. We then proceed to discussing previous research in automated fact checking in terms of what inputs they expect, what outputs they return and the evidence used in this process. Following this, we provide an overview of the most commonly used datasets and the models developed and evaluated on them. Subsequently, we discuss work related to automated fact checking and conclude by proposing avenues for future NLP research.

2 Fact checking in journalism

Fact checking is an essential component in the process of news reporting. Journalism has been defined by scholars as a "discipline of verification" to separate it from "entertainment, propaganda, fiction, or art" (Shapiro et al., 2013). While fact checking and verification are often used interchangeably, recent work has sought to define them as two distinct but complementary processes (Silverman, 2016). In particular, verification is defined as "scientific-like approach of getting the fact and also the right facts" (Kovach and Rosenstiel, 2014), which often involves verifying the source, date and the location of materials. Fact-checking on the other hand "addresses the claim's logic, coherence and context" (Mantzarlis, 2015). Consequently, verification is a necessary and crucial first step in the process of fact checking as it assesses the trustworthiness of the contexts considered. We adopt this distinction for the remainder of this survey as it helps highlight key differences between various automated approaches.[1]

A term that has become strongly associated with fact checking is "fake news", especially since its use in the context of the 2016 US presidential elections. However its popularity has resulted in its meaning becoming diluted, as it is used to label claims on a number of aspects not necessarily related to veracity (Vosoughi et al., 2018). The most prominent example of such usage of the term of fake news is its application to media organizations of opposing political sides. Furthermore, it is often grouped together with the term "hate speech" as in the case of recent legislation (Deutsche Welle, 2017), despite the latter being more related to the use of emotive language instead of truth assessment (Rahman, 2012). Therefore we avoid using this term in this survey.

A further consideration is the relation of fact checking with the terms misinformation and disinformation. While the former is the distribution of information that may not be accurate or complete, the latter additionally assumes malicious motives to mislead the reader (Jowett and O' Donnell, 2006). Due to this additional consideration, disinformation is considered a subset of misinformation. Fact checking can help detect misinformation, but not distinguish it from disinformation.

3 Towards Automated Fact Checking

The increased demand for fact checking has stimulated a rapid progress in developing tools and systems to automate the task, or parts thereof (Babakar and Moy, 2016; Graves, 2018). Thus, there has been a diverse array of approaches tailored to specific datasets, domains or subtasks. While they all share the same goal, these approaches utilize different definitions of what the task being automated is. In the following discussion, we highlight the differences between the task definitions used in previous research in the following axes: input, i.e. what is being fact checked, output, i.e. what kinds of verdicts are expected, and the evidence used in the fact checking process.

3.1 Inputs

We first consider the inputs to automated fact checking approaches as their format and content influences the types of evidence used in this process. A frequently considered input to fact checking approaches is subject-predicate-object triples, e.g. `(London, capital_of, UK)`, and is popular across different research communities, including NLP (Nakashole and Mitchell, 2014), data mining (Ciampaglia et al., 2015) and Web search (Bast et al., 2015). The popularity of triples as input stems from the fact that they facilitate fact checking against (semi-)structured knowledge bases such Freebase (Bollacker et al., 2008). However, it is important to acknowledge that approaches using triples as input implicitly assume a non-trivial level of processing in order to convert text, speech or other forms of claims into triples, a task falling under the broad definition of natural language understanding (Woods, 1973).

A second type of input often considered in automated fact checking is textual claims. These tend to be short sentences constructed from longer passages, which is a practice common among human fact checkers on dedicated websites such as PolitiFact[2] and Full Fact[3] with the purpose of including only the context relevant to the claim from the original passage. The availability of fact checked claims on such websites has rendered this format very popular among researchers in NLP.

A useful taxonomy of textual claims was proposed in the context of the HeroX fact checking challenge (Francis and Full Fact, 2016), in which four types of claims were considered:

- **numerical claims** involving numerical properties about entities and comparisons among them
- **entity and event properties** such as professional qualifications and event participants
- **position statements** such as whether a political entity supported a certain policy
- **quote verification** assessing whether the claim states precisely the source of a quote, its content, and the event at which it supposedly occurred.

[1] Note that Mantzarlis and Silverman disagree on whether fact-checking should be considered a subset of verification, or just overlapping with it. While this is in an important question, we do not address it in this survey as it does not affect our analysis.

[2] http://www.politifact.com/

[3] http://www.fullfact.org/

While some of these claims could be represented as triples, they typically require more complex representations; for example, events typically need to be represented with multiple slots (Doddington et al., 2004) to denote their various participants. Regardless of whether textual claims are verified via a subject-predicate-object triple representation or as text, it is often necessary to disambiguate the entities and their properties. For example, the claim 'Beckham played for Manchester United' is true for the soccer player 'David Beckham', but not true (at the time of writing) for the American football player 'Odel Beckham Jr'. Correctly identifying, disambiguating and grounding entities is the task of Named Entity Linking (McNamee and Dang, 2009). While this must be performed explicitly if converting a claim to a subject-predicate-object triple with reference to a knowledge base, it may also be performed implicitly through the retrieval of appropriate textual evidence if fact checking against textual sources.

Finally, there have been approaches that consider an entire document as their input. These approaches must first identify the claims and then fact check them. This increases the complexity of the task, as approaches are required to extract the claims, either in the form of triples by performing relation extraction (Vlachos and Riedel, 2015) or through a supervised sentence-level classification (Hassan et al., 2015b).

3.2 Sources of evidence

The type of evidence that is used for fact checking influences the model and the types of outputs that the fact checking system can produce. For example, whether the output is a label or whether a justification can be produced depends largely on the information available to the fact checking system.

We first consider task formulations that do not use any evidence beyond the claim itself when predicting its veracity such as the one by Rashkin et al. (2017). In these instances, surface-level linguistic features in the claims are associated with the predicted veracity. We contrast this to how journalists work when fact checking, where they must find knowledge relating to the fact and evaluate the claim given the evidence and context when making the decision as to whether a claim is true or false. The predictions made in task formulations that do not consider evidence beyond the claim are based on surface patterns of how the claim is written rather than considering the current state of the world.

Wang (2017) incorporate additional metadata in fact checking such as the originator of the claim, speaker profile and the media source in which the claim is presented. While these do not provide evidence grounding the claim, the additional context can act as a prior to improve the classification accuracy, and can be used as part of the justification of a verdict.

Knowledge graphs provide a rich collection of structured canonical information about the world stored in a machine readable format that could support the task of fact checking. We observe two types of formulations using this type of evidence. The first approach is to identify/retrieve the element in the knowledge graph that provides the information supporting or refuting the claim at question. For example, Vlachos and Riedel (2015) and Thorne and Vlachos (2017) identify the subject-predicate-object triples from small knowledge graphs to fact check numerical claims. Once the relevant triple had been found, a truth label is computed through a rule based approach that considers the error between the claimed values and the retrieved values from the graph. The key limitation in using knowledge graphs as evidence in this fashion is that it assumes that the true facts relevant to the claim are present in them. However, it is not feasible to capture and store every conceivable fact in the graph in advance of knowing the claim.

The alternative use of a knowledge graph as evidence is to consider its topology in order to predict how likely a claim (expressed as an edge in the graph) is to be true (Ciampaglia et al., 2015). While graph topology can be indicative of the plausibility of a fact, nevertheless if a fact is unlikely to occur that does not negate its truthfulness. Furthermore, improbable but believable claims are more likely to become viral and thus in greater need of verification by fact checkers.

Text, such as encyclopedia articles, policy documents, verified news and scientific journals contain information that can be used to fact check claims. Ferreira and Vlachos (2016) use article headlines (single sentences) as evidence to predict whether an article is for, against or observing a claim. The Fake News Challenge (Pomerleau and Rao, 2017) also formulated this part of the fact checking process in the same way, but in contrast to (Ferreira and Vlachos, 2016), entire documents are used as evidence, thus allowing for evidence from multiple sentences to be combined.

The Fact Extraction and VERification (FEVER) task (Thorne et al., 2018) requires combining information from multiple documents and sentences for fact checking. Unlike the aforementioned works which use text as evidence, the evidence is not given but must be retrieved from Wikipedia, a large corpus of encyclopedic articles. Scaling up even further, the triple scoring task of the WSDM cup (Bast et al., 2017) required participants to assess knowledge graph triples considering both Wikipedia and a portion of the web-scale ClueWeb dataset (Gabrilovich et al., 2013).

A different source of text-based evidence is repositories of previously fact checked claims (Hassan et al., 2017b). Systems using such evidence typically match a new claim to claim(s) in such a repository and return the label of the match if one is found. This type of evidence enables the prediction of veracity labels instead of only whether a claim is supported or refuted. However, this evidence is limiting fact checking only to claims similar to the ones already existing in the repository.

As an alternative to using knowledge stored in texts or databases as evidence, aggregate information on the distribution of posts on social networks can provide insights into the veracity of the content. Rumor veracity prediction is the assessment of the macro level behaviors of users' interactions with and distribution of content to predict whether the claims in the content are true or false (Derczynski and Bontcheva, 2015; Derczynski et al., 2017). This crowd behavior provide useful insights, especially in cases where textual sources or structured knowledge bases may be unavailable.

The trustworthiness of the sources used as evidence, i.e. the verification aspect of fact checking in journalistic terms, is rarely considered by automated approaches, despite its obvious importance. Often the sources of evidence are considered given, e.g. Wikipedia or Freebase, as this facilitates development and evaluation, especially when multiple models are considered. Nevertheless, approaches relying on Twitter metadata often consider credibility indicators for the tweets used as evidence in their fact checking process (Liu et al., 2015). Similar to journalism, trustworthiness assessment is often considered as a separate task and we discuss it further in the Related Work section.

3.3 Output

The simplest model for fact checking is to label a claim as true or false as a binary classification task (Nakashole and Mitchell, 2014). However, we must also consider that it is possible in natural language to be purposefully flexible with the degree of truthfulness of the information expressed or express a particular bias using true facts. Journalistic fact checking agencies such as Politifact model the degree of truthfulness on a multi-point scale (ranging from true, mostly-true, half-true, etc). Rather than modeling fact checking as a binary classification, Vlachos and Riedel (2014) suggested modeling this degree of truthfulness as an ordinal classification task. However, the reasoning behind why the manual fact checking agencies have applied these more fine-grained labels is complex, sometimes inconsistent[4] and likely to be difficult to capture and express in our models. Wang et al. (2017) and Rashkin et al. (2017) expect as output multiclass labels following the definitions by the journalists over a multi-point scale but ignoring the ordering among them.

The triple scoring task of the WSDM cup (Bast et al., 2017) expected as output triples scored within a numerical range indicating how likely they are to be true. The evaluation consisted of calculating both the differences between the predicted and the manually annotated scores, as well as the correlation between the rankings produced by the systems and the annotators.

Ferreira and Vlachos (2016) expect as output whether a claim is supported, refuted or just reported by a news article headline. Pomerleau and Rao (2017) added an extra label for the article being irrelevant to the claim, and consider the full article instead of the headline. While this output can be used in the process of fact checking, it is not a complete fact check.

In FEVER (Thorne et al., 2018) the output expected consists of two components: a 3-way classification label about whether a claim is supported/refuted by Wikipedia, or there is not enough information in the latter to Wikipedia it, and in the case of the first two labels, the sentences forming the evidence

[4] http://www.poynter.org/news/can-fact-checkers-agree-what-true-new-study-doesnt-point-answer

to reach the verdict. In effect this combines the multiclass labeling output with a ranking task over Wikipedia sentences. If the label predicted is correct but the evidence retrieved is incorrect, then the answer is considered incorrect, highlighting the importance of evidence in this task formulation. Nevertheless, this output cannot be considered a complete fact check though, unless we restrict world knowledge to Wikipedia. The use of full world knowledge in the justifications was considered in the HeroX shared task (Francis and Full Fact, 2016); while this allowed for complete fact checks, it also necessitated manual verification of the outputs.

4 Fact Checking Datasets

There are currently a limited number of published datasets resources for fact checking. Vlachos and Riedel (2014) collected and released 221 labeled claims in the political domain that had been fact checked by Politifact and Channel4.[5] For each labeled claim, the speaker or originator is provided alongside hyperlinks to the sources used by the fact checker as evidence. The sources range from statistical tables and Excel spreadsheets to PDF reports and documents from The National Archives. While these sources can be readily assessed by human fact checkers, the variety and lack of structure in the evidence makes it difficult to apply machine learning-based approaches to select evidence as separate techniques may be required for the each different document format. Where documents are provided as evidence, we do not know which portions of the document pertain to the claim, which compounds the difficulty of assessing and evaluating a fact checking system, given the need for evidence and accountability. Furthermore, with such a limited number of samples, the scale of this dataset precludes its use for developing machine learning-based fact checking systems.

Wang (2017) released a dataset similar to but an order of magnitude larger than that of Vlachos and Riedel (2014) containing 12.8K labeled claims from Politifact. In addition to the claims, meta-data such as the speaker affiliation and the context in which the claim appears in (e.g. speech, tweet, op-ed piece) is provided. At this scale, this dataset can support training and evaluating a machine learning-based fact checking system. However, the usefulness of the dataset may be limited due the claims being provided without machine-readable evidence beyond originator metadata, meaning that systems can only resort to approaches to fact checking such as text classification or speaker profiling.

Rashkin et al. (2017) collated claims from Politifact without meta-data. In addition, the authors published a dataset of 74K news articles collected from websites deemed as Hoax, Satire, Propaganda and Trusted News. The prediction of whether a news article is true or false is modeled as the task of predicting whether an article originates from one of the websites deemed to be “fake news” according to a US News & World report.[6] This allows systems to fact check using linguistic features, but does not consider evidence, or which aspects of a story are true or false.

Ferreira and Vlachos (2016) released a dataset for rumor debunking using data collected from the Emergent project (Silverman, 2015). For 300 claims, 2,595 corresponding news articles were collected and their stances were labeled as for, against or observing a claim. This dataset was extended for the 2017 Fake News Challenge (Pomerleau and Rao, 2017) dataset which consisted of approximately 50K headline and body pairs derived from the original 300 claims and 2,595 articles.

The HeroX Fast and Furious Fact Checking Challenge (Francis and Full Fact, 2016) released 90 (41 practice and 49 test) labeled claims as part of the competition. Because part of the challenge required identification of appropriate source material, evidence for these claims was not provided and therefore manual evaluation was needed on case-by-case basis. As with the claims collected by Vlachos and Riedel (2014), the size of the dataset prevents its use for training a machine learning-based fact checking system. However, the broad range of types of claims in this dataset highlights a number of forms of misinformation to help identify the requirements for fact checking systems.

Thorne et al. (2018) introduced a fact checking dataset containing 185K claims about properties of entities and concepts which must be verified using articles from Wikipedia. While this restricts the pool

[5] http://channel4.com/news/factcheck

[6] www.usnews.com/news/national-news/articles/2016-11-14/avoid-these-fake-news-sites-at-all-costs

of evidence substantially compared to HeroX, it allowed for a challenging evidence selection subtask that could be feasibly annotated manually at a large scale in the form of sentences selected as evidence. Combined with the labels on the claims, these machine readable fact checks allow training machine learning-based fact checking models.

5 Methods

The majority of the methods used for automated fact checking are supervised: learning a text classifier from some form of labeled data that is provided at training. This trait that is independent of the task input or what sources of evidence are considered. Vlachos and Riedel (2014) suggested fact checking using supervised models, making use of existing statements that had been annotated with verdicts by fact checking agencies and journalists. Wang (2017) and Rashkin et al. (2017) applied this approach to classify the veracity of fake news stories. The main limitation of text classification approaches is that fact checking a claim requires additional world knowledge, typically not provided with the claim itself. While language can indicate whether a sentence is factual (Nakashole and Mitchell, 2014), credible sounding sentences may also be inherently false. Text classification on claims alone has been used for the related task of detecting fact-check worthy claims (Hassan et al., 2017a).

Ciampaglia et al. (2015) use network analysis to predict whether an unobserved triple is likely to appear in a graph by modeling the task as a path ranking problem (Lao et al., 2011). The truth verdict is derived from the cost of traversing a path between the two entities under transitive closure, weighted by the degree of connectedness of the path. Nakashole and Mitchell (2014) combine linguistic features on the subjectivity of the language used with the co-occurrence of a triple with other triples from the same topic, a form of collective classification.

Ferreira and Vlachos (2016) modeled fact checking as a form of Recognizing Textual Entailment (RTE) (Dagan et al., 2009; Bowman et al., 2015), predicting whether a premise, typically (part of) a news article, is *for*, *against*, or *observing* a given claim. The same type of model was used by most of the 50 participating teams in the Fake News Challenge (Pomerleau and Rao, 2017), including most of the top entries (Riedel et al., 2017; Hanselowski et al., 2017).

The RTE-based models assume that the textual evidence to fact check a claim is given. Thus they are inapplicable in cases where this is not provided (as in HeroX), or it is a rather large textual resource such as Wikipedia (as in FEVER). For the latter, Thorne et al. (2018) developed a pipelined approach in which the RTE component is preceded by a document retrieval and a sentence selection component. There is also work focusing exclusively on retrieving sentence-level evidence from related documents for a given claim (Hua and Wang, 2017).

Vlachos and Riedel (2015) and Thorne and Vlachos (2017) use distantly supervised relation extraction (Mintz et al., 2009) to identify surface patterns in text which describe relations between two entities in a knowledge graph. Because these fact checking approaches only focus on statistical properties of entities, identification of positive training examples is simplified to searching for sentences containing numbers that are approximately equal to the values stored in the graph. Extending this approach to entity-entity relations would pose different challenges, as a there may be many relations between the same pair entities that would need to be accurately distinguished for this approach to be used.

A popular type of model often employed by fact checking organizations in their process is that of matching a claim with existing, previously fact checked ones. This reduces the task to sentence-level textual similarity as suggested by Vlachos and Riedel (2014) and implemented in ClaimBuster (Hassan et al., 2017b), Truthteller by The Washington Post[7] and one of the two modes of Full Fact's Live platform.[8] However, it can only be used to fact check repeated or paraphrased claims.

Long et al. (2017) extend the models produced by Wang (2017) and improve accuracy of a simple fact checking system through more extended profiling of the originators of the claims. The most influential feature in this model is the *credit history* of the originator, a metric describing how often the originator's

[7] https://www.knightfoundation.org/articles/debuting-truth-teller-washington-post-real-time-lie-detection-service-your-service-not-quite-yet

[8] https://fullfact.org/blog/2017/jun/automated-fact-checking-full-fact/

claims are classified as false. This feature introduces a bias in the model that fits with the adage "never trust a liar". In the case of previously unannotated sources which may haven no recorded history, this feature would be unavailable. Furthermore, the strong reliance on credit history has some important ethical implications that need to be carefully considered. Despite these limitations, speaker profiling has been shown to be effective in other related studies (Gottipati et al., 2013; Long et al., 2016) and is discussed further in Section 6.

6 Related Tasks

Verification In Section 2, we highlighted the difference between verification and fact checking. While the models considered in Section 4 operate under a closed-world paradigm where we assume that all evidence provided is true, for these fact checking technologies to scale-to and operate on the web, we must also consider information that is of unknown veracity as evidence.

Methods of predicting the authoritativeness of web pages such as PageRank (Brin and Page, 1998) only consider the hyperlink topology. TrustRank (Gyöngyi et al., 2004) provides a framework which incorporates annotated information and predicts the trustworthiness of pages based on graph-connectedness to known-bad nodes rather than the information content. An alternative is Knowledge-based Trustworthiness scoring (Dong et al., 2015), which allows predicting whether the facts extracted from a given document page are likely to be accurate given the method used to extract the facts and the website in which the document is published.

Common Sense Reasoning Fact checking requires the ability to reason about arguments with common sense knowledge. This requires developing systems that go beyond recognizing semantic phenomena more complex than those typically considered in textual entailment tasks. Habernal et al. (2018) introduced a new task and dataset for predicting which implicit *warrant* (the rationale of an argument) is required to support a claim from a given premise. Angeli and Manning (2014) proposed a method of extracting common sense knowledge from WordNet for reasoning about common sense knowledge using Natural Logic and evaluated their approach on a subset of textual entailment problems in the FraCaS test suite (Cooper et al., 1996). It is important to build systems that can reason about both explicit world knowledge and implicit common sense knowledge is an essential step towards automating fact checking.

Subjectivity and Emotive Language Rashkin et al. (2017) assess the reliability of entire news articles by predicting whether the document originates from a website classified as Hoax, Satire or Propaganda. This work is an instance of subjective language detection and does not represent evidence-based fact checking. The authors used supervised classifiers augmented with lexicons including the Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015), a sentiment lexicon (Wilson et al., 2005), a hedging lexicon (Hyland, 2015), and a novel 'dramatic' language lexicon to identify emotive, subjective and sensational language in the article bodies. Furthermore, analysis of the lexical features using a logistic regression classifier shows that the highest weighted (most distinguishing) features for the unreliable sources included the use of hedge words (such as 'reportedly') or words pertaining to divisive topics (such as 'liberals' or 'Trump'). The authors apply a similar model to the prediction of claims made by politicians from claims collected from Politifact. The addition of the LIWC lexicon which provides an indication of emotive tone and authenticity marginally improved the classification accuracy for simple lexical models.

Deceptive Language Detection There are linguistic cues and features in written text that are useful in identifying deceptive language (Zhou et al., 2004). In the context of detecting deceptive user-generated content - a specific form of disinformation, Mihalcea and Strapparava (2009) use a simple lexical classification model without further feature engineering. Analysis of the model identifies a number of word classes of the LIWC lexicon which pertain only to the deceptive texts. Ott et al. (2011) incorporate the use of psycholinguistic cues to improve classification accuracy. Mihalcea and Strapparava (2009) found that truthful texts were more likely to contain words belonging to the 'optimistic' LIWC class such as 'best', 'hope', and 'determined'. This is corroborated by the study of sentiment in deceptive texts

(Ott et al., 2013) which also identified that texts with negative sentiment were more likely to be deceptive. These feature classes however may be an artifact of the data generation process as crowd-sourced volunteers were first asked to write an honest text and rewrite it to make it deceptive.

Feng et al. (2012) detect deceptive texts and customer-generated reviews through the use of syntactic style rather word-based content. Non-terminal nodes of the constituency parse trees are used as features in conjunction with a lexical model to increase the accuracy over using words alone. Hai et al. (2016) identify deceptive reviews also using lexical features. However, rather than relying on labeled data, the authors induce labels over an unlabeled dataset through a semi-supervised learning approach that exploits a minimal amount labeled data from related tasks in a multi-task learning set up.

Even though linguistic content, emotive language and syntax are useful indicators for detecting deceit, the truthfulness of a statement depends also on the context. Without considering these factors these approaches cannot be used to fact check information alone.

Rumor Detection Rumor detection (Qazvinian et al., 2011) is the task of identifying unverified reports circulating on social media. A prediction is typically based on language subjectivity and growth of readership through a social network. While these are important factors to consider, a sentence can be true or false regardless of whether it is a rumor (Zubiaga et al., 2018).

Speaker Profiling Identifying claims that do not fit with the profile of the originator may provide insights as to whether the information is truthful or not. Furthermore, determining which topics the originator is truthful about may allow for generation of a risk-based approach to fact checking. As mentioned earlier, Long et al. (2017) introduced a notion of credit history which increases the classification accuracy for fake news detection. However, this notion doesn't account for which topics the originator lies about. Furthermore, the assumption that each source has an overall trustworthiness score to be attached to every claim from there is not a valid one, since inaccurate information may be found even on the most reputable of sources.

An alternative is to consider the compatibility of a claim with respect to the originator's profile. This remains an open research area. Feng and Hirst (2013) perform the inverse of this task for deceptive language detection in product reviews by creating an average profile for the targets (products) and using the distance between a review and the target as a feature. The shortcomings of this method are that number of reviews are required for each target. Considering the tasks of verifying automatically extracted information or fact checking for politics, for a new topic the challenge is that there may be insufficient data to create a profile for it.

Pérez-Rosas and Mihalcea (2015) identify author characteristics (such as age and gender) that influence the linguistic choices made by the authors when fabricating information in product reviews. As the affiliation, age and gender of most politicians is public-domain knowledge, it is conceivable that these features may assist fact checking political claims. While the use of meta-data does improve classification accuracy (Wang et al., 2017; Long et al., 2017), for fact checking we must consider the meaning of the claim with respect to ground truth rather than based on the linguistic style of the originator.

Click Bait Detection Intentionally misleading headlines or titles that are designed specifically to encourage a user to click through and visit a website are called click bait. Studies into the detection of click bait have have yielded positive results from relatively simple linguistic features (Chen et al., 2015; Potthast et al., 2016; Chakraborty et al., 2016). These approaches only consider the article headline and do not make use of evidence. We contrast this to the task of detecting headlines which are incongruent to the document body (Chesney et al., 2017), where existing methods for recognizing textual entailment such as those used for the Fake News Challenge (Pomerleau and Rao, 2017) can be applied.

7 Open Research Challenges

In this paper, we provided an NLP-motivated overview of fact checking, considering the case for evidence - similar to how the task is performed by journalists. We highlighted existing works that consider fact checking of claims expressed in text or via knowledge base triples and pointed out their shortcomings, given the need to justify their decisions using evidence.

We did not identify approaches that make use of open-world knowledge. Assessing the ability of a system in an open-world setting will be difficult; the HeroX challenge resorted to fully manual evaluation, which is difficult to scale up, especially for the purposes of developing machine learning-based approaches. Thus we need to consider how to address the information retrieval challenge of the task, including its evaluation. We must also consider the verification of the evidence used, which is ignored under the closed world assumption. Ideally we should consider verification jointly with fact checking, which is in fact how it is conducted in journalism.

The relative scarcity of resources designed explicitly for fact checking, highlights the difficulties in capturing and formalizing the task, especially considering the strong interest in it. While recently published large scale resources such as FEVER can stimulate progress, they only consider simple short sentences (8 words long on average). Thus, there is scope to fact check compound information or complex sentences, and scale up to fact checking at the document level.

Furthermore, in FEVER the justification for the labels is restricted to sentences selected from Wikipedia. This is much unlike the rationales produced by human fact checkers, who synthesize information. The generation of such rationales has attracted attention only recently in NLP (Ling et al., 2017), and automated fact checking could provide an ideal testbed to develop it further.

While text is often used to make a claim, often the evidence need for fact checking appears in other modalities, such as images and videos. Given the recent interest in multi-modal NLP, we argue that this would be an important direction for future research, especially when considering that a lot of the verification efforts in journalism are focused on identifying forged images and footage.

Finally, it is important to acknowledge that the complexity of the fact checking conducted by journalists is for the moment beyond the abilities of the systems we can develop due to the rather complex reasoning needed. Even a simple short statement such as that the "UK, by leaving the EU, will take back control of roughly £350 million per week" takes a substantial amount of work to be checked.[9] In this fact check for example, the first step is to adjust the claim to render it more accurate in terms of its meaning so that the actual fact check can proceed. While this complexity can help stimulate progress in NLP and related fields, it should also calibrate our expectations and promises to society.

Acknowledgements

Andreas Vlachos is supported by the EU H2020 SUMMA project (grant agreement number 688139). The authors would like to thank Dhruv Ghulati and Alexios Mantzarlis for their feedback.

References

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. Technical report, National Bureau of Economic Research.

Gabor Angeli and Christopher D. Manning. 2014. NaturalLI : Natural Logic Inference for Common Sense Reasoning. *Proceedings of EMNLP*, pages 534–545.

Mevan Babakar and Will Moy. 2016. The State of Automated Factchecking. Technical report.

Hannah Bast, Björn Buchhold, and Elmar Haussmann. 2015. Relevance scores for triples from type-like relations. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 243–252. ACM.

Hannah Bast, Björn Buchhold, and Elmar Haussmann. 2017. Overview of the Triple Scoring Task at the WSDM Cup 2017. *WSDM Cup*.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. *SIGMOD 08 Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

[9] https://fullfact.org/europe/350-million-week-boris-johnson-statistics-authority-misuse/

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference.

S Brin and L Page. 1998. The anatomy of a large scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1/7):107–17.

Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. 2016. Stop Clickbait: Detecting and preventing clickbaits in online news media. *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*, pages 9–16.

Yimin Chen, Niall J. Conroy, and Victoria L. Rubin. 2015. Misleading Online Content. *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection - WMDD '15*, (November):15–19.

Sophie Chesney, Maria Liakata, Massimo Poesio, and Matthew Purver. 2017. Incongruent Headlines: Yet Another Way to Mislead Your Readers. In *Natural Language Processing meets Journalism workshop at EMNLP 2017*, number 1, pages 56–61.

Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M. Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PLoS ONE*, 10(6):1–13.

Sarah Cohen, Chengkai Li, Jun Yang, and Cong Yu. 2011. Computational Journalism: a call to arms to database researchers. *Proceedings of the 5th Biennial Conference on Innovative Data Systems Research (CIDR 2011) Asilomar, California, USA.*, (January):148–151.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, et al. 1996. Using the framework. Technical report, Technical Report LRE 62-051 D-16, The FraCaS Consortium.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 15(4):i–xvii.

Leon Derczynski and Kalina Bontcheva. 2015. Veracy in Digital Social Networks. *Pheme.eu*.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours.

Deutsche Welle. 2017. Germany to force facebook, twitter to delete hate speech. `http://www.dw.com/en/germany-to-force-facebook-twitter-to-delete-hate-speech/a-379270` Accessed: 2018-09-30.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation. In *Proceedings of the Fourth Conference on Language Resources and Evaluation*.

Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. 2015. Knowledge-based trust: Estimating the trustworthiness of web sources. *Proc. VLDB Endow.*, 8(9):938–949, May.

Vanessa Wei Feng and Graeme Hirst. 2013. Detecting Deceptive Opinions with Profile Compatibility. *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, (October):338–346.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2. Association for Computational Linguistics*, (July):171–175.

William Ferreira and Andreas Vlachos. 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*, pages 1163–1168, San Diego, California, June. Association for Computational Linguistics.

Diane Francis and Full Fact. 2016. Fast & furious fact check challenge. `https://www.herox.com/factcheck/`. Accessed: 2018-09-30.

Evgeniy Gabrilovich, Michael Ringgaard, and Amarnag Subramanya. 2013. FACC1: Freebase annotation of ClueWeb corpora, Version 1 (Release date 2013-06-26, Format version 1, Correction level 0), June.

Swapna Gottipati, Minghui Qiu, Liu Yang, Feida Zhu, and Jing Jiang. 2013. Predicting User ' s Political Party. pages 177–191.

Lucas Graves. 2018. Understanding the Promise and Limits of Automated Fact-Checking. Technical report, Reuters Institute, University of Oxford.

Z Gyöngyi, H Garcia-Molina, and J Pedersen. 2004. Combating web spam with TrustRank. *Proceedings of the Thirtieth international conference on Very large data bases*, 30:576–587.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 1930–1940.

Zhen Hai, Peilin Zhao, Peng Cheng, Peng Yang, Xiao-Li Li, and Guangxia Li. 2016. Deceptive Review Spam Detection via Exploiting Task Relatedness and Unlabeled Data. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-16)*, pages 1817–1826.

Andreas Hanselowski, Avinesh P.V.S, Benjamin Schiller, and Felix Caspelherr. 2017. Description of the system developed by team athene in the fnc-1. Technical report.

Naeemul Hassan, Bill Adair, James Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. 2015a. The quest to automate fact-checking. In *Proceedings of the 2015 Computation+Journalism Symposium*.

Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015b. Detecting Check-worthy Factual Claims in Presidential Debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM*.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017a. 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*, pages 1803–1812. ACM.

Naeemul Hassan, Anil Kumar Nayak, Vikas Sable, Chengkai Li, Mark Tremayne, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, and Aaditya Kulkarni. 2017b. ClaimBuster: the first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12):1945–1948.

Xinyu Hua and Lu Wang. 2017. Understanding and detecting supporting arguments of diverse types. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 203–208. Association for Computational Linguistics.

Ken Hyland. 2015. Metadiscourse. *The International Encyclopedia of Language and Social Interaction*, (April 2015):1–11.

Heng Ji and Ralph Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1148–1158, Stroudsburg, PA, USA. Association for Computational Linguistics.

Garth S. Jowett and Victoria O' Donnell. 2006. What is propaganda, and how does it differ from persuasion? In *Propaganda and Misinformation*, chapter 1. Sage Publishers.

Bill Kovach and Tom Rosenstiel. 2014. *The elements of journalism: What newspeople should know and the public should expect*. Three Rivers Press (CA).

N Lao, T Mitchell, and WW Cohen. 2011. Random walk inference and learning in a large scale knowledge base. *Proceedings of the Conference ...*, (March):529–539.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 158–167.

Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. 2015. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1867–1870, New York, NY, USA. ACM.

Yunfei Long, Qin Lu, Yue Xiao, Minglei Li, and Chu Ren Huang. 2016. Domain-specific user preference prediction based on multiple user activities. *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016*, pages 3913–3921.

Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. 2017. Fake News Detection Through Multi-Perspective Speaker Profiles. *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2:252–256.

Alexios Mantzarlis. 2015. Will verification kill fact-checking? `https://www.poynter.org/news/will-verificatio` Accessed: 2018-09-30.

Paul McNamee and Hoa Trang Dang. 2009. Overview of the tac 2009 knowledge base population track.

Rada Mihalcea and Carlo Strapparava. 2009. The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language. *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, (August):309–312.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, (August):1003–1011.

Ndapandula Nakashole and Tom M Mitchell. 2014. Language-Aware Truth Assessment of Fact Candidates. *Acl*, pages 1009–1019.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 309–319, Stroudsburg, PA, USA. Association for Computational Linguistics.

Myle Ott, Claire Cardie, and Jeffrey T. Hancock. 2013. Negative deceptive opinion spam. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 497–501. Association for Computational Linguistics.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The Development and Psychometric Properties of LIWC2015. *Environment and Planning D: Society and Space*.

Verónica Pérez-Rosas and Rada Mihalcea. 2015. Experiments in Open Domain Deception Detection. (September):1120–1125.

Dean Pomerleau and Delip Rao. 2017. Fake News Challenge. \url{http://fakenewschallenge.org/}.

Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen. 2016. Clickbait Detection. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9626(1):810–817.

Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. 2011. Rumor has it : Identifying Misinformation in Microblogs. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1589–1599.

Jacquelyn Rahman. 2012. The n word: Its history and use in the african american community. *Journal of English Linguistics*, 40(2):137–171.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937. Association for Computational Linguistics.

Benjamin Riedel, Isabelle Augenstein, George Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. *CoRR*, abs/1707.03264.

Ivor Shapiro, Colette Brin, Isabelle Bedard-Brule, and Kasia Mychajlowycz. 2013. Verification as a strategic ritual. *Journalism Practice*, 7(6):657–673.

Craig Silverman. 2015. Lies , Damn Lies , and Viral Content: How News Websites Spread (and Debunk) Online Rumors, Unverified Claims, and Misinformation. *Columbia Journalism School*.

Craig Silverman. 2016. Verification handbook: Additional materials. `http://verificationhandbook.com/additionalmaterial/`. Accessed: 2018-09-30.

James Thorne and Andreas Vlachos. 2017. An Extensible Framework for Verification of Numerical Claims. In *European Chapter of the Association for Computational Linguistics (EACL)*, pages 37–40.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. In *NAACL-HLT*.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA, June. Association for Computational Linguistics.

Andreas Vlachos and Sebastian Riedel. 2015. Identification and Verification of Simple Claims about Statistical Properties. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, (September):2596–2601.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Shuohang Wang, Mo Yu, Jing Jiang, Wei Zhang, Xiaoxiao Guo, Shiyu Chang, Zhiguo Wang, Tim Klinger, Gerald Tesauro, and Murray Campbell. 2017. Evidence Aggregation for Answer Re-Ranking in Open-Domain Question Answering. 1:1–13.

William Yang Wang. 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)*, pages 422–426.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05*, (October):347–354.

W. A. Woods. 1973. Progress in natural language understanding: An application to lunar geology. In *Proceedings of the June 4-8, 1973, National Computer Conference and Exposition*, AFIPS '73, pages 441–450, New York, NY, USA. ACM.

Lina Zhou, Judee K. Burgoon, Jay F. Nunamaker, and Doug Twitchell. 2004. Automating Linguistics-Based Cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation*, 13(1):81–106.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.*, 51(2):32:1–32:36, February.