Credibility Analysis of Textual Claims with Explainable Evidence

A dissertation submitted towards the degree
Doctor of Engineering (Dr.-Ing.)
of the Faculty of Mathematics and Computer Science
of Saarland University

by
Kashyap Popat

Saarbrücken
September 2019

Defense Colloquium

Date:	November 26, 2019
Dean:	Prof. Dr. Sebastian Hack

Examination Committee

Chair:	Prof. Dr. Anja Feldmann
Reviewer and Advisor:	Prof. Dr. Gerhard Weikum
Reviewer:	Prof. Dr. Felix Naumann
Reviewer:	Dr. Andrew Yates
Academic Assistant:	Dr. Paramita Mirza

Abstract

Despite being a vast resource of valuable information, the Web has been polluted by the spread of false claims. Increasing hoaxes, fake news, and misleading information on the Web have given rise to many fact-checking websites that manually assess these doubtful claims. However, the rapid speed and large scale of misinformation spread have become the bottleneck for manual verification. This calls for credibility assessment tools that can automate this verification process. Prior works in this domain make strong assumptions about the structure of the claims and the communities where they are made. Most importantly, black-box techniques proposed in prior works lack the ability to explain why a certain statement is deemed credible or not.

To address these limitations, this dissertation proposes a general framework for automated credibility assessment that does not make any assumption about the structure or origin of the claims. Specifically, we propose a feature-based model, which automatically retrieves relevant articles about the given claim and assesses its credibility by capturing the mutual interaction between the language style of the relevant articles, their stance towards the claim, and the trustworthiness of the underlying web sources. We further enhance our credibility assessment approach and propose a neural-network-based model. Unlike the feature-based model, this model does not rely on feature engineering and external lexicons. Both our models make their assessments interpretable by extracting explainable evidence from judiciously selected web sources.

We utilize our models and develop a Web interface, CredEye, which enables users to automatically assess the credibility of a textual claim and dissect into the assessment by browsing through judiciously and automatically selected evidence snippets. In addition, we study the problem of stance classification and propose a neural-network-based model for predicting the stance of diverse user perspectives regarding the controversial claims. Given a controversial claim and a user comment, our stance classification model predicts whether the user comment is supporting or opposing the claim.

Kurzfassung

Das Web ist eine riesige Quelle wertvoller Informationen, allerdings wurde es durch die Verbreitung von Falschmeldungen verschmutzt. Eine zunehmende Anzahl an Hoaxes, Falschmeldungen und irreführenden Informationen im Internet haben viele Websites hervorgebracht, auf denen die Fakten überprüft und zweifelhafte Behauptungen manuell bewertet werden. Die rasante Verbreitung großer Mengen von Fehlinformationen sind jedoch zum Engpass für die manuelle Überprüfung geworden. Dies erfordert Tools zur Bewertung der Glaubwürdigkeit, mit denen dieser Überprüfungsprozess automatisiert werden kann. In früheren Arbeiten in diesem Bereich werden starke Annahmen gemacht über die Struktur der Behauptungen und die Portale, in denen sie gepostet werden. Vor allem aber können die Black-Box-Techniken, die in früheren Arbeiten vorgeschlagen wurden, nicht erklären, warum eine bestimmte Aussage als glaubwürdig erachtet wird oder nicht.

Um diesen Einschränkungen zu begegnen, wird in dieser Dissertation ein allgemeines Framework für die automatisierte Bewertung der Glaubwürdigkeit vorgeschlagen, bei dem keine Annahmen über die Struktur oder den Ursprung der Behauptungen gemacht werden. Insbesondere schlagen wir ein featurebasiertes Modell vor, das automatisch relevante Artikel zu einer bestimmten Behauptung abruft und deren Glaubwürdigkeit bewertet, indem die gegenseitige Interaktion zwischen dem Sprachstil der relevanten Artikel, ihre Haltung zur Behauptung und der Vertrauenswürdigkeit der zugrunde liegenden Quellen erfasst wird. Wir verbessern unseren Ansatz zur Bewertung der Glaubwürdigkeit weiter und schlagen ein auf neuronalen Netzen basierendes Modell vor. Im Gegensatz zum featurebasierten Modell ist dieses Modell nicht auf Feature-Engineering und externe Lexika angewiesen. Unsere beiden Modelle machen ihre Einschätzungen interpretierbar, indem sie erklärbare Beweise aus sorgfältig ausgewählten Webquellen extrahieren.

Wir verwenden unsere Modelle zur Entwicklung eines Webinterfaces, CredEye, mit dem Benutzer die Glaubwürdigkeit einer Behauptung in Textform automatisch bewerten und verstehen können, indem sie automatisch ausgewählte Beweisstücke einsehen. Darüber hinaus untersuchen wir das Problem der Positionsklassifizierung und schlagen ein auf neuronalen Netzen basierendes Modell vor, um die Position verschiedener Benutzerperspektiven in Bezug auf die umstrittenen Behauptungen vorherzusagen. Bei einer kontroversen Behauptung und einem Benutzerkommentar

sagt unser Einstufungsmodell voraus, ob der Benutzerkommentar die Behauptung unterstützt oder ablehnt.

To my spiritual guide

Acknowledgment

I express my sincere gratitude to my supervisor, Gerhard Weikum, for his excellent mentorship and guidance throughout my doctoral studies. His continuous support, constant motivation to do my best and the freedom to pursue my research interests have been of the utmost importance to make this dissertation possible.

Tremendous thanks to Prof. Felix Naumann and Dr. Andrew Yates for reviewing this thesis and giving their helpful feedback. Thanks also to Prof. Anja Feldmann and Dr. Paramita Mirza for agreeing to serve on my examination committee. I also thank Dr. Simon Razniewski and Petra Schaaf for translating the abstract of this dissertation to German.

Special thanks to my collaborators and co-authors Subhabrata Mukherjee and Jannik Strötgen for their valuable contributions towards shaping this thesis. I would also like to thank Fabrizio Silvestri for hosting me during my internship at Facebook London. I learned many valuable lessons about research and development during my internship.

I am grateful to all my friends and colleagues from MPI-INF for creating an excellent work environment and providing valuable feedback throughout my doctoral studies. I am thankful to the administrative staff at the institute for providing their assistance. Thanks to all my friends in Saarbrücken, especially Sarvesh Nikumbh, Arunav Mishra, Neha Agarwal, Satish Verma, Nikhil Upadhyaya, Visheet Arya, Guruprasad Hegde, Sivarajan Karunanithi, and Goutam Y G for making my life enjoyable.

My heartfelt thanks to Geetanjali Ram, Ram Mahalingam, and their lovely children, Sujana and Ramanuj, for making me a part of their family. I will always cherish their support and the good times I spent with them. I also thank my friends from the Heartfulness meditation group for helping me to grow spiritually and emotionally.

Last but not least, I want to thank my parents, Chandrika and Kirit Popat, and my sister Bansi Popat, for their continuous support and encouragement through all these years. Most importantly, I thank my wife Lila Kurse, for always supporting me, especially during the final stages of my doctoral studies. Thank you, Lila, for always being by my side and helping me to become a better person every day.

Kashyap Popat

Contents

List of Figures

List of Tables

1 Introduction

Contents

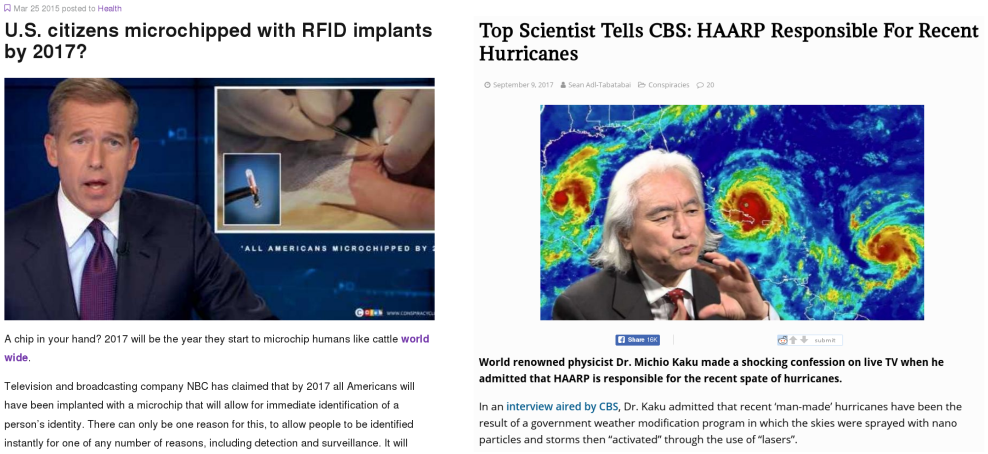
1.1 Motivation

The revolutionary invention of the World Wide Web has made sharing information to the world an extremely easy task. This explosive growth of the web, including online news and social media, has made significant changes in the consumption of the web content. More and more people tend to rely on news from the web rather than traditional news organizations. For instance, a recent survey discovers that 68% of U.S. adults get news on social media sites[1].

Despite being a vast resource of valuable information, in recent years, there is a significant increase in the spread of misinformation on the web [Shu et al., 2017; Kumar and Shah, 2018]. The World Economic Forum has identified "the rapid spread of misinformation online" as one of the top ten challenges the world faces[2]. This rampant spread of misinformation on the web and social media has made extremely negative impacts at both societal and individual level, such as, hindering the relief and response efforts during the crisis [Mendoza et al., 2010; Gupta et al., 2013], affecting stock market [Aggarwal and Wu, 2006; Bollen et al., 2011], affecting political attitudes [Brewer et al., 2013; Balmas, 2014], etc., to name a few. Studies on the misinformation effect have also shown impairment in human memory arising

[1] https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/ (accessed on 15 July, 2019)

[2] http://reports.weforum.org/outlook-14/ (accessed on 15 July, 2019)



Mar 25 2015 posted to Health

U.S. citizens microchipped with RFID implants by 2017?

A chip in your hand? 2017 will be the year they start to microchip humans like cattle **world wide**.

Television and broadcasting company NBC has claimed that by 2017 all Americans will have been implanted with a microchip that will allow for immediate identification of a person's identity. There can only be one reason for this, to allow people to be identified instantly for one of any number of reasons, including detection and surveillance. It will

(a) An article from `disclose.tv`[3]

Top Scientist Tells CBS: HAARP Responsible For Recent Hurricanes

September 9, 2017 Sean Adl-Tabatabai Conspiracies 20

World renowned physicist Dr. Michio Kaku made a shocking confession on live TV when he admitted that HAARP is responsible for the recent spate of hurricanes.

In an interview aired by CBS, Dr. Kaku admitted that recent 'man-made' hurricanes have been the result of a government weather modification program in which the skies were sprayed with nano particles and storms then "activated" through the use of "lasers".

(b) An article from `yournewswire.com`[4]

Figure 1.1: Examples of misinformation on the web.

after exposure to misleading information [Loftus, 2005; Morgan et al., 2013]. Given the widespread nature of this critical issue, words like "Post-truth" and "Fake news" are named as word of the year by Oxford dictionary in the year 2016 and by the American Dialect Society in the year 2017 respectively. Two examples of such web articles with misinformation are shown in Figure 1.1.

This societal challenge has given rise to many fact-checking and debunking websites such as Snopes (`snopes.com`), PolitiFact (`politifact.com`), FullFact (`fullfact.org`), etc., where trained professionals manually analyze such controversial claims, assess their credibility and provide analysis along with the supporting evidence such as, background articles, trustworthiness of the information source, quotations, etc. However, this manual verification is intellectually demanding and time-consuming. Depending on the complexity of the claim, this verification may take from few hours to few days [Hassan et al., 2015]. Hence, to keep up with the scale and the speed at which misinformation spreads, we need tools to automated this manual verification process. This has stimulated great research interest in addressing this arduous task of automated credibility assessment – also known as automated fact-checking. As fully objective and unarguable truth is often elusive or ill-defined, we use the term "credibility" instead of "truth".

The goal of the automated credibility assessment is to reduce the burden by assisting human in verifying the veracity of the factual information. However, considering the severity of the problem, it is not enough to build black-box systems which can only assess information to be credible (true) or dubious (false). We need

[3] `https://www.disclose.tv/us-citizens-microchipped-with-rfid-implants-by-2017-309943` (accessed on 15 July 2019)

[4] `https://archive.is/Kg9mV` (archived version; accessed on 15 July 2019)

systems which can also provide user-interpretable evidence and counter-evidence to support its automatic assessment. Information that needs to be assessed can be in different formats such as video, audio, text or their different combinations. However, in this thesis, we focus only on the textual information in the English language.

1.2 Challenges

Enabling machines to successfully perform any intellectual task that a human being can perform has been a long-term goal of the Artificial Intelligence (AI) research field [Goertzel and Pennachin, 2007]. However, as reported in Kumar et al. [2016], sometimes, even humans cannot easily distinguish hoax articles from authentic ones, and quite a few people have mistaken satirical articles (e.g., theonion.com) as truthful news. Hence, automatic assessment of information veracity is an extremely challenging task.

The problem of automated credibility assessment of textual claims comprises of several challenging problems spanning across multiple fields such as natural language processing (NLP), machine learning, social network analysis, etc. Primarily, there are three fundamental challenges for this task:

- **Understanding natural language:** One of the major challenges for automated credibility assessment is to understand what is conveyed by the natural language text as well as how it is conveyed. Even though there has been significant progress towards understanding natural language [Devlin et al., 2019], the problem remains far from completely solved.

- **Extracting evidence:** In this era of big data, the major challenge for automatic assessment is to collect relevant and sufficient evidence to verify facts. A vast amount of data is made available on the web at every second. Most of this data is in unstructured form. Technologies such as knowledge base repositories (e.g., YAGO [Suchanek et al., 2007], WikiData [Vrandečić and Krötzsch, 2014], etc.), information extraction and semantic web help in processing the unstructured text into a machine-readable format. However, the coverage of these repositories and technologies is limited compared to the available unstructured data on the web.

- **Estimating trustworthiness:** Content on the web is generated by various sources, for instance, news websites, blog posts, social media, discussion forums, etc. Unfortunately, not all the information sources are credible. Hence, another key challenge for automated credibility assessment is to assess the trustworthiness of the web sources.

In the following sections, we discuss these challenges in detail.

1.2.1 Understanding Natural Language

"Knowledge of languages is the doorway to wisdom."
– Roger Bacon

Language plays a key role in assessing information veracity. To understand what is conveyed by the natural language text as well as how it is conveyed (i.e., language style) is of crucial importance for automated credibility assessment. Numerous studies have validated the relationship between the quality of the information and the language style in which it is presented [Afroz et al., 2012; Chen et al., 2015; Rashkin et al., 2017].

One of the primary purposes behind spreading misinformation is to intentionally deceive people for financial or political gains. Articles producing misinformation often use exaggeration, scaremongering, and opinionated or sensational language to attract attention and encourage users to engage with the misinformation. For instance, consider the content of the article shown in Figure 1.1b:

Example 1.1

"World renowned physicist Dr. Michio Kaku made a shocking confession on live TV when he admitted that HAARP is responsible for the recent spate of hurricanes."

The above text tries to mislead the user by misquoting a famous scientist and evokes anger toward a particular government research program by scaremongering. The subjective phrases such as *"shocking confession"* and *"admitted"* give cues about the bias and deceiving language style. Hence, one of the challenges for automated credibility assessment is to capture this language stylistic cues.

Diverse Perspectives and Their Stance

Another consequence of information overload on the web is increasing diverse perspectives about the controversial information such as misleading statements from politicians, biased news reports, rumors, etc. People express their opinion about these controversial claims through various channels like editorials, blog posts, social media, and discussion forums. To achieve a deeper understanding of information credibility, it is essential to understand these diverse perspectives and their *stance* towards the claim. For instance, consider the content of an article[5] expressing its perspective about the claim that *"U.S. citizens are supposed to be microchipped with RFID implants by 2017"* (see Figure 1.1a):

[5] https://www.thatsnonsense.com/will-all-americans-be-microchipped-by-2017-debunked/ (accessed on 15 July, 2019)

Example 1.2

"The theory that the American government is actively looking to implant Americans with RFID tracking chips to help control the US population is a long running conspiracy that is persistent as it is utterly baseless. Despite the many different variants of this consistent conspiracy theory, no compelling evidence has ever been offered to support the baseless and paranoid claims."

The above text expresses the author's stance about the claim. Highlighted phrases such as, *"conspiracy"*, *"utterly baseless"*, *"no compelling evidence"* clearly indicate that the author is *refuting* the claim. Encountering such evidence that refutes the claim gives cues about the controversial nature of the claim and helps in understanding its credibility. Hence, the challenge here is to consider these diverse perspectives and understand their stance.

1.2.2 Extracting Evidence

> *"Extraordinary claims require extraordinary evidence."*
>
> – Carl Sagan

A fact is something which can be proven to be true with evidence. Hence, gathering evidence is a fundamental step in assessing the credibility of any claim or information. Automated verification of such claims requires machines to automatically collect the relevant evidence. Therefore, any such automated system will be restricted to the repository of evidence which are available digitally – in a machine-readable format.

The web is the embodiment of human knowledge. Majority of the textual data on the web is in unstructured form. Knowledge base repositories, such as YAGO or WikiData extract information from the unstructured web data and convert it into a structured format. However, such repositories are not up-to-date and their coverage is quite limited. Hence, they are not very helpful in providing evidence to verify controversial facts especially the ones which are arising out of current world affairs.

Another way for accessing evidence on the web is facilitated by search engines. An automated approach for credibility assessment can utilize these search engines and carry out a web search to retrieve the relevant evidence. However, typically search engines return a list of webpages which are relevant to the textual search query. These relevant webpages are in the different format following different structure, such as a news article, a collection of question answers, or a discussion on social media, etc. Extracting relevant evidence out of this chaotic jumble is another challenge for automated credibility assessment.

1.2.3 Estimating Trustworthiness

"Learning to trust is one of life's most difficult tasks."

– Isaac Watts

Even though the web is a vast resource of knowledge, not everything on the web is credible and not all the information sources are trustworthy. The trustworthiness of the information sources directly affects the credibility of the information [Flanagin and Metzger, 2008]. For instance, a fact reported in The New York Times (`nytimes.com`) is likely to credible – rigorously analyzed by the professional journalists. On the other hand, some report from The Onion (`theonion.com`) is most certainly not credible since it is a satire news organization. Hence, estimating the trustworthiness of the information sources is of utmost importance for assessing the credibility of the information.

Traditional approaches for estimating the quality of web sources, such as PageRank [Brin and Page, 1998] and authority-hub analysis [Kleinberg, 1999] rely on the hyperlink structure of the web graph. However, such approaches only capture the authority and popularity of the web-sources and not their trustworthiness from the information credibility perspective. For instance, the satirical news website The Onion has a very high PageRank score (7 out of 10). Hence, estimating the trustworthiness of information sources from the credibility perspective remains a challenge for automated credibility assessment.

1.3 Prior Work and Its Limitations

Prior approaches for truth-finding and data fusion (refer to Li et al. [2015b] for a survey) mainly focused on resolving conflicts among the structured facts. The facts are typically in the form of subject-predicate-object or relational tables from multiple sources. A classical example of such structured fact is *"Mahatma Gandhi was born in Delhi"* viewed as a triple $\langle$*Mahtma Gandhi, born in, Delhi*$\rangle$ where *"Delhi"* is the critical value. These approaches also assume that the alternative values for the questionable slot, e.g., *"Porbandar"*, *"Mumbai"*, or *"Goa"* in the above example, are already present. Given a set of these conflicting values, these approaches perform conflict resolution and find the true value (i.e., *"Porbandar"*). These truth-finding approaches can not work with unstructured facts in natural language text.

On the other hand, approaches for social media credibility analysis (refer to Shu et al. [2017]; Zubiaga et al. [2018] for surveys) have mainly focused on detecting rumors and misinformation in closed social media communities such as Twitter and Facebook. However, most of these approaches rely heavily on the platform-specific features, for instance, number of retweets, number of likes, etc. Moreover, some of these approaches also utilize the underlying network of the social media community. Such approaches are not suitable for assessing the credibility of facts in an open-domain setting without any assumptions about the community or website where these factual claims are made.

Moreover, most of the works for automated credibility assessment utilize machine learning models and classifiers to predict discrete decision labels as output, for instance, *"true"* or *"false"* in the case of credibility assessment. However, these black-box approaches rarely explain how the model reaches a particular decision. Interpretability of such models is very limited and it becomes extremely challenging to explain the final verdict of the model to the end-users.

1.4 Thesis Contributions

This dissertation addresses the challenges outlined in the previous sections. We overcome the limitations of the prior approaches and address the novel problem of automated credibility assessment of textual claims that are expressed freely in an open-domain setting. Moreover, we do not make any assumption about the structure of the claim, or characteristics of the community or website where the claim is made. In summary, this dissertations makes the following contributions:

Credibility Assessment Framework

This dissertation proposes a key framework for automated credibility assessment. Given a textual claim, we first search the articles from multiple web-sources which are relevant to the input claim. Then, we individually analyze these articles to estimate their opinions regarding the claim's credibility. Finally, we aggregate these individual opinions to predict how likely the input claim is *true* or *false*. As fully objective and unarguable truth is often elusive or ill-defined, instead of directly predicting the credibility labels, we return the probability scores associated with the credibility labels. Our preliminary model for assessing the credibility leverages the joint interaction between the language style of the evidence articles and the trustworthiness of the underlying web sources (based on PageRank and AlexaRank). Our experiments with two real-world datasets from Snopes[6] and reported cases of Wikipedia hoaxes[7,8] demonstrate the effectiveness of our framework. This work was published at CIKM 2016 [Popat et al., 2016].

[6] https://www.snopes.com/ (accessed 15 July, 2019)

[7] https://en.wikipedia.org/wiki/List_of_hoaxes#Proven_hoaxes (accessed 15 July, 2019)

[8] https://en.wikipedia.org/wiki/List_of_fictitious_people (accessed 15 July, 2019)

Feature-Based Credibility Assessment

Here, we address the limitations of our prior model for credibility assessment [Popat et al., 2016]. We propose that considering only the language style of the evidence articles is not adequate. Understanding the stance of these articles towards the claim is crucial for automated credibility assessment. Additionally, our initial approach for estimating the trustworthiness of underlying web-sources was based on PageRank and AlexaRank measures. However, these measures mostly capture the popularity of the web-sources. To address this, we propose a new methodology for estimating the trustworthiness of web-sources from the perspective of information credibility. We also incorporate the dynamics of how claims emerge, spread, and are supported or refuted (i.e., stance towards the claim) to further enhance our credibility assessment model. In addition to the final credibility verdict of the claim, we also provide explanations for interpreting the final verdict. These user-interpretable explanations are in the form of informative snippets from judiciously selected sources. This is another major contribution of this work. Our extensive experiments demonstrate the viability of our enhanced approach. This work was published at WWW 2017 [Popat et al., 2017].

Neural-Network-Based Credibility Assessment

This dissertation also proposes a neural-network-based approach for automated credibility assessment. Here, we address the limitations of our own prior approaches to further enhance our model for credibility assessment. The downside of our prior approaches [Popat et al., 2016, 2017] is that it requires substantial feature modeling and rich lexicons to detect bias and subjectivity in the language style. Our proposed end-to-end neural network model overcomes this limitation as it does not require any feature engineering, lexicons or other manual interventions. Moreover, we also propose an attention mechanism to capture the interaction between the claim and the evidence article. Automatically generated user-interpretable explanations enriched with informative features help users to understand the model predictions. Our experiments with four different datasets highlight the strength of our approach. This work was published at EMNLP 2018 [Popat et al., 2018b].

Web Interface for Credibility Assessment

In this work, we publicly release *CredEye*, a web interface for automated credibility assessment based on our prior work [Popat et al., 2017]. Given an input claim in textual form on an arbitrary topic, CredEye automatically retrieves relevant articles from the web, using a search engine. It assesses the credibility of the input claim by analyzing the language style and stance of these articles along with the trustworthiness of the underlying sources. CredEye enables users to dissect and drill down into the assessment by browsing through judiciously and automatically selected snippets with the markup of indicative words. These indicative words capture linguistic features that express bias and subjectivity (decreasing credibility) or neutral and objective language (increasing credibility). We also show the details of the analysis in the form of per-article and per-source scores. CredEye is available at `https://gate.d5.mpi-inf.mpg.de/credeye/`. This work was published as a demonstration paper at WWW 2018 [Popat et al., 2018a].

Determining Stance

In this work, we propose a neural network model for stance classification leveraging representations from the language representation model BERT (Bidirectional Encoder Representations from Transformers) [Devlin et al., 2019] and augmenting them with a novel consistency constraint. Given an input pair of a claim and a users perspective, our model predicts whether the perspective is *supporting* or *opposing* the claim. Experiments on the Perspectrum dataset [Chen et al., 2019], consisting of claims and users perspectives from various debate websites, demonstrate the effectiveness of our approach over state-of-the-art baselines. This work was published at EMNLP 2019 [Popat et al., 2019].

1.5 Prior Publications

The results of this thesis have been published in the following articles:

1. Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, 2016.

2. Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 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, 2017.

3. Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, 2018b.

4. Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. Credeye: A credibility lens for analyzing and explaining misinformation. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, 2018a.

5. Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. STANCY: Stance classification based on consistency cues. 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)*, EMNLP '19, 2019.

Additionally, a summary of this dissertation has appeared in the quarterly ACM SIGWEB Newsletter [Popat, 2018] and the thesis proposal has been presented at the PhD Symposium at WWW 2017 [Popat, 2017].

1.6 Organization

The remainder of the dissertation is organized as follows. Chapter 2 provides a summary of related approaches in this area. Chapter 3 lays the foundation of our credibility assessment framework. We gradually enhance our model for automated credibility assessment in the form of a feature-based approach in Chapter 4 and a neural-network-based approach in Chapter 5. We describe CredEye, our web interface in Chapter 6, which can automatically assess the credibility of natural language claims in a few seconds. For a better understanding of controversial claims from diverse perspectives, we explore the problem of stance classification in Chapter 7. Finally, Chapter 8 concludes this dissertation and describes future research directions.

2
Related Work

Contents

Our work is related to several overlapping domains: truth discovery, credibility analysis on social media and web, stance detection and explainable evidence extraction. In this background chapter, we provide an overview of various approaches in these areas. We discuss the state-of-the-art and present their limitations.

2.1 Truth Discovery

The goal of truth discovery approaches is to resolve conflicts among multi-source data [Yin et al., 2008; Dong et al., 2009; Galland et al., 2010; Pasternack and Roth, 2010; Zhao et al., 2012; Li et al., 2012; Pasternack and Roth, 2011, 2013; Dong and Srivastava, 2013; Li et al., 2014b,c, 2015c; Ma et al., 2015; Zhi et al., 2015; Gao et al., 2015; Lyu et al., 2017]. These approaches, starting with the seminal work of Yin et al. [2008], assume the input data to be in a structured format, for instance, an entity of interest (e.g., book) along with its potential conflicting values provided by different sources (e.g., the author). Li et al. [2015b] give a detailed survey of truth-finding approaches.

The assumption about the structured data is reflected by these approaches in different forms. Dong et al. [2009]; Zhao et al. [2012]; Pasternack and Roth [2011] assume the input facts to be in the form of a source, object, and questionable value. Similarly, Li et al. [2011b, 2012] assume that the input facts are in a particular form with a clear identification of questionable values. On the other hand, Nakashole and Mitchell [2014] assumes the input facts to be in the form of subject-predicate-object triples, e.g., <*Obama, born_in, Kenya*>, where *"Kenya"* is the critical value. The assumption about such structured input is crucial for these approaches in order to identify alternative facts. Models proposed in Dong et al. [2009]; Zhao et al. [2012]; Pasternack and Roth [2011]; Li et al. [2014b, 2015c] assume that such alternative facts are already given. On the other hand, Li et al. [2011b]; Nakashole and Mitchell [2014] go one step ahead and use a search engine to retrieve conflicts facts from multiple sources.

Algorithms proposed in Dong et al. [2009]; Galland et al. [2010]; Pasternack and Roth [2010]; Yin et al. [2008] estimate the truth values and source trustworthiness iteratively until the convergence. Li et al. [2014b,c, 2015c] propose optimization-based methods for truth-finding with the objective function of minimizing the weighted distance between the truth and the conflicting values from different sources. On the other hand, some truth discovery approaches [Zhao et al., 2012; Pasternack and Roth, 2013; Ma et al., 2015] are based on probabilistic graphical models (PGM). Nakashole and Mitchell [2014] propose a language features based model to determine whether the given subject-predicate-object triplet is objective or speculative. Vinod Vydiswaran et al. [2011] propose a ranking-based method to assess the trustworthiness of medical claims based on community knowledge in health portals.

Most of these truth-finding approaches address the problem of conflict resolution amongst multi-source data with an assumption about the structure of the input facts and the availability of conflicting facts. Due to these limitations, the majority of these methods do not take into account the natural language facts and the language in which these facts are reported by various sources.

The method in Samadi et al. [2016] jointly estimates the credibility of sources and correctness of the claims using the Probabilistic Soft Logic framework. However, it does not consider the deeper semantic aspects of article language. Vydiswaran et al. [2012] conducted a user study to understand how various factors such as, the impact of presenting contrasting viewpoints, source expertise ratings, etc., affect the truthfulness of controversial claims. Similarly, Rashkin et al. [2017]; Wang [2017] propose neural network-based approaches for determining the credibility of a textual claim, but it does not consider external sources like web evidence and claim sources.

In this thesis, we propose generic approaches for credibility assessment for natural language facts without making any assumption about their structure. Our models jointly capture mutual interactions between the language style of the articles reporting the fact, their stance towards the fact and the trustworthiness of underlying web-sources. Moreover, unlike many of the black-box approaches, we provide the user interpretable evidence for explaining the automatic verdict.

2.2 Information Credibility in Social Media

Methods for assessing the credibility of social media posts mainly exploit community-specific features, such as, number of likes or upvotes, popularity, who-replied-to-whom, etc. to detect rumors and deceptive content [Castillo et al., 2011; Qazvinian et al., 2011; Yang et al., 2012; Gupta et al., 2013; Yates et al., 2015; Zhao et al., 2015; Volkova et al., 2017]. A detailed survey of various social media-centric approaches for credibility assessment is given in Shu et al. [2017] and Kumar and Shah [2018]. These assessment approaches mainly target the following problems.

2.2.1 Rumor Detection

The seminal work of Castillo et al. [2011] proposes a supervised model for assessing the credibility of user posts on Twitter (`twitter.com`). Their approach is based on features from the text of the user postings, users' postings and re-posting (retweets) behavior, and references to external sources. A large corpus of tweets, topics, and events along with the associated human judgments about their credibility is released in Mitra and Gilbert [2015].

An unsupervised language model based method for detecting fake content is proposed in Lavergne et al. [2008]. Whereas, Qazvinian et al. [2011]; Gupta and Kumaraguru [2012] propose supervised models utilizing content-based and network-based features for detecting rumors on Twitter. Mitra et al. [2017] harness the language cues to model the credibility of tweets. Similarly, methods in Yang et al. [2012]; Wu et al. [2015] combine user, text, topics, and propagation-based features to detect rumors on Sina Weibo (`weibo.com`). Jin et al. [2016] propose a network-based iterative approach which utilizes conflicting viewpoints in microblogs to predict the credibility of news. Detecting fake images on Twitter based on user and tweet based features is addressed in Gupta et al. [2013].

On the other hand, Ma et al. [2016] propose a neural network-based model for rumor detection in microblogs. Similarly, a three-stage neural network approach in Ruchansky et al. [2017] jointly models the text of the article shared on the microblog, the response it receives, and the user sources to detect the fake news articles.

2.2.2 Identifying Social Media Bots

The spread of misinformation also involves “bad” actors. Many works address this problem of identifying such bad users on social media platforms. Ferrara et al. [2016] gives a detailed survey about this problem and highlights the methods to detect social media bots on Twitter. To address this problem, a Twitter bot challenge was also held recently by the U.S. agency, DARPA [Subrahmanian et al., 2016].

Studies in Shao et al. [2018]; Bessi and Ferrara [2016] analyze messages and articles shared on Twitter during and following the 2016 U.S. presidential campaign and election. They provide evidence for how social bots amplify the low-credibility content on social media. Similarly, Beutel et al. [2013] address the problem of detecting fraudulent user feedback on Facebook (`facebook.com`). Whereas, frameworks to study the impact and influence of bots on Twitter have been proposed in Gilani et al. [2016]; Varol et al. [2017].

Methods in Stein et al. [2011] and Alvisi et al. [2013] address this problem of detecting bots in an adversarial learning setting. Techniques proposed in Lee et al. [2011]; Chu et al. [2012]; Davis et al. [2016] utilize various network-based, user-based and temporal features to detect social media bots. Analysis by Dickerson et al. [2014] shows that sentiment related factors are crucial for identifying social media bots. On the other hand, Lee et al. [2010] propose a honeypot-based approach for uncovering spammers on the social media platform.

2.2.3 Detecting Spread of Misinformation

Several existing works also study the problem of how misinformation spreads in the social media network. For instance, work by Kwon et al. [2013] studies the propagation of rumors in social media by examining the temporal, structural and linguistic aspects of diffusion. They propose a time series model to detect rumors. A detailed study of rumor cascade is presented in Friggeri et al. [2014].

A study of user behavior and the propagation of rumors on Twitter during an emergency is presented in Mendoza et al. [2010]. Similarly, Starbird [2017] studies alternative media ecosystem on Twitter. It utilizes a network-based approach to expose how alternative narratives spread misinformation on Twitter.

To address the problem of misinformation spread, Kim et al. [2018] propose a temporal point processes based framework which efficiently selects which stories from Twitter and Sina Weibo to send for manual fact-checking and when to do so. Tripathy et al. [2010] models rumor spread as a diffusion process on a network and proposes anti-rumor strategies by embedding agents in the network to fight the spread of misinformation. Instead of classifying microblog information as credible or not, work by Nguyen et al. [2012] proposes a method for identifying a small set of influential users in the social media network to counter the spread of misinformation. A web-based service for real-time analysis of misinformation diffusion is demonstrated in Ratkiewicz et al. [2011a,b].

A recent study in Quattrociocchi et al. [2016] explores the polarization of social media and provides quantitative evidence to highlight the existence of echo chambers on social media. Similarly, Vicario et al. [2016]; Del Vicario et al. [2016] further study users' involvement inside the echo chamber and how it affects the spreading of misinformation on Facebook. Whereas, Garimella et al. [2018] study political echo chambers on Twitter.

However, all these methods are geared towards specific social media platforms and most of the times they rely heavily on the platform-specific features, such as the number of likes, tweets, shares, etc. Hence, it is difficult to generalize these methods in open-domain. Moreover, these approaches mainly propose black-box models which do not give any explanations for their final verdict.

In this thesis, we propose generic methods for automated credibility assessment of natural language claims without making any assumptions about the community where these claim are made.

2.3 Information Credibility in Communities

Prior research for credibility assessment in communities mainly address the problem of detecting deceptive content and harmful users in the community, for instance, identifying sockpuppets, detecting vandalism on Wikipedia, detecting opinion spams, etc. [Mukherjee and Weikum, 2015; Mukherjee et al., 2016; Mukherjee, 2017; Kumar et al., 2016, 2017].

2.3.1 Predicting Content Quality

Some existing works also address the problem of predicting the quality of the content shared on web communities. For instance, Probabilistic Graphical Models (PGMs) are proposed for detecting credible user statements in health forums [Mukherjee et al., 2014], news discussion forms [Mukherjee and Weikum, 2015], and product review forums [Mukherjee et al., 2016, 2017]. These methods jointly model the credibility of user statements, their language objectivity, and trustworthiness of community users.

Work by Kumar et al. [2016] studies the impact of misinformation on Wikipedia and propose a classification model to detect whether a given Wikipedia article is a hoax. Similarly, few methods [Nakov et al., 2017; Mihaylova et al., 2019] address the problem of predicting content quality in community question answering forums.

2.3.2 Opinion Spam Analysis

Seminal work of Jindal and Liu [2007, 2008] lays the foundation of opinion spam detection problem and presents a detailed study about the types of spam reviews and proposes supervised models to detect them. Methods proposed in Ott et al. [2011, 2013]; Harris [2012]; Xu and Zhao [2012]; Li et al. [2014a] employ linguistic analysis to separate the deceptive reviews from the truthful ones. A semi-supervised model for detecting spam reviews and spammers is proposed by Li et al. [2011a].

On the other hand, Akoglu et al. [2013] propose a network-based unsupervised method for detecting spam reviews on a large scale datasets. Rayana and Akoglu [2015] combines content-based meta-data and network-based features to build a joint-model. Some techniques [Xie et al., 2012; Li et al., 2015a, 2017] also rely on the temporal and spatial patterns to solve the problem of spam reviews. Bayesian model to identify fraudulent reviews, based on user rating behavior, is proposed in Hooi et al. [2016].

Mukherjee et al. [2012] targets the group of fake reviewers and proposes frequent itemset mining and user behavioral-based models to detect them. They further try to also dissect Yelp's algorithm for filtering spam reviews [Mukherjee et al., 2013].

2.3.3 Identifying Harmful Users

Existing works by Lim et al. [2010]; Kumar et al. [2018] propose methods for identifying fraudulent users using network and user behavior properties. Similarly, Wang et al. [2011] construct a heterogeneous review graph capturing relationships amongst the reviewers, reviews, and products. Their iterative model harnesses the interaction between graph nodes to detect the spam reviewers.

On the other hand, Yang et al. [2011]; Kumar et al. [2017] study sockpuppetry in which a single community user creates multiple identities to deceive other community users or manipulate discussions. Similarly, Cheng et al. [2017] study how a user's mood and the context of a discussion can lead to trolling behavior.

However, most of these approaches are limited to specific communities – utilizing community-specific features. They are not easily adaptable to the open-domain setting. Additionally, the lack of explanations from these methods also makes it extremely hard to explain the final verdict to the end-users.

2.4 Language-Based Text Analytics

Several research works analyze the text from a linguistic point of view to address various problems such as sentiment analysis, bias detection, satire or deception detection, sarcasm detection, etc. Works addressing these problems mainly harness different linguistic cues and propose supervised models.

Sentiment Analysis

Starting with the seminal work of Pang et al. [2002] several techniques have addressed the problem of sentiment analysis. These methods [Turney, 2002; Dave et al., 2003; Pang and Lee, 2004, 2008; Taboada et al., 2011; Liu, 2012] tap into linguistic features such as phrase and word-based linguistic lexicons, dependency relations, discourse analysis, etc. to classify customer reviews as positive, negative, or objective. On the other hand, works by Pak and Paroubek [2010]; Agarwal et al. [2011] have proposed to address the problem of sentiment analysis of Twitter data.

Bias and Subjectivity Detection

Linguistic cues for detecting biased language, such as factive verbs, implicatives, hedges, subjective words, etc. are identified in Recasens et al. [2013]. Similarly, methods proposed in Wiebe et al. [2004]; Wiebe and Riloff [2005]; Lin et al. [2011] use different linguistic features to address the problem of identifying subjective text.

Satire Detection

A novel task of detecting whether a news article is satire or not is proposed in Burfoot and Baldwin [2009]. A detailed analysis of various kinds of deceptive news articles, including satirical, fabricated and hoax news articles is provided in Rubin et al. [2015]. Other techniques, such as Ahmad et al. [2014]; Pilar Salas-Zárate et al. [2017]; Ravi and Ravi [2017] utilize various linguistic features for satire detection. Similarly, the model proposed in Afroz et al. [2012] uses linguistic cues to detect hoaxes, frauds, and deception in writing style.

2.5 Stance Detection

Ease of expressing opinions provided by the web has triggered a great research interest in mining these opinions and diverse perspectives. Especially for a better understanding of controversial claims, analyzing diverse perspectives becomes a crucial task [Chen et al., 2019]. Additionally, recent research (including our own) [FNC-1, 2016; Popat et al., 2017; Baly et al., 2018] has shown stance classification to be a critical step for information credibility and automated fact-checking.

Various methods for detecting user's stance in online debating platforms are proposed in Somasundaran and Wiebe [2009, 2010]; Anand et al. [2011]; Walker et al. [2012]; Hasan and Ng [2013]; Sridhar et al. [2015]. A method proposed in Sridhar et al. [2014] harnesses the structural and linguistic features of user posts to predict their stance towards the controversial topics. These methods mainly rely on the linguistic features, for instance, n-grams, dependency parse tree, opinion lexicons, sentiment, etc., to determine the stance of user perspectives about controversial topics discussed on various online debate websites. A method proposed by Ferreira and Vlachos [2016] further incorporates controversial claims in natural language form along with the users' perspectives. They propose a logistic regression model using the lexical and semantic features of claims and perspectives.

A stance classifier based on hand-crafted lexicons is proposed in Bar-Haim et al. [2017]. Their method identifies important phrases in perspectives and their consistency with the claim to predict the stance. However, their model assumes that the important phrases in claims are already identified.

Recently, many neural network-based approaches have been proposed for stance classification. These approaches learn the claim and perspective representations separately and later combine them with conditional LSTM encoding [Augenstein et al., 2016], attention mechanisms [Du et al., 2017] or memory networks [Mohtarami et al., 2018]. Additional lexical features are also incorporated in some neural network models [Riedel et al., 2017; Hanselowski et al., 2018; Zhang et al., 2018].

On the other hand, various SemEval tasks [Ebrahimi et al., 2016; Mohammad et al., 2016, 2017; Derczynski et al., 2017] and other approaches [Chen and Ku, 2016; Lukasik et al., 2016; Sobhani et al., 2017; Kochkina et al., 2017] have focused on determining stance in Twitter discussions.

A recent work [Chen et al., 2019] proposes a supervised method for stance detection based on a language representation model called BERT (Bidirectional Encoder Representations from Transformers) [Devlin et al., 2019]. However, it does not explicitly capture the agreement between the controversial claim and user perspective. In this thesis, we address this limitation and enhance the stance detection model by augmenting it with a novel consistency constraint to capture agreement between the controversial claim and user perspective.

2.6 Trust and Reputation Analysis

There has been a lot of work studying how to measure the trustworthiness and quality of the web-sources. The seminal algorithms for trust estimation, PageRank [Brin and Page, 1998] and Authority-hub analysis [Kleinberg, 1999] analyze links between various sources on the web to estimate their trustworthiness. Similarly, algorithms such as EigenTrust [Kamvar et al., 2003] and TrustMe [Singh and Liu, 2003] rely on source behavior in a P2P network to estimate their trustworthiness.

Similar methods have also studied source trust and reputation in Wikipedia [Adler and de Alfaro, 2007], P2P networks [Wang and Vassileva, 2003] and online interactions [Mui et al., 2002; De Alfaro et al., 2011]. On the other hand, algorithms proposed in Castillo et al. [2007]; Gyöngyi et al. [2004]; Li et al. [2014c] focus on detecting web spam. Vydiswaran et al. [2011] proposed an algorithm for trust propagation in a network of claims, articles, and article sources.

However, all these approaches mainly rely on the hyperlink structure of the web graph and do not capture the source trustworthiness from the information credibility perspective. To address this limitation, the work by Dong et al. [2015] goes beyond the hyperlink structure of the web graph and proposes a probabilistic graphical model to estimate the source trustworthiness based on the correctness of the factual information provided by different sources. A temporal point process model for estimating source trustworthiness in community question answering forums is proposed in Tabibian et al. [2017].

2.7 Interpretable Machine Learning

In the era of artificial intelligence, machine learning models have become the first choice for solving critical problems related to finance, health, recruitment, the justice system, etc. Due to this prime importance, it has become crucial to understand how and why the models make certain decisions. As defined in Miller [2017], interpretability is the degree to which a human can understand the cause of the decision. However, most of the current machine learning models are not interpretable since they do not explain their decisions. In general, interpretability also helps in detecting underlying biases in machine learning models. This problem has attracted significant attention from the research community [Wilson et al., 2017; Linzen et al., 2018]. A detailed discussion about the motivation of interpretability and different ways to achieve it is given in Lipton [2018].

Many classical machine learning models, such as regression, Naive Bayes, decision tree, random forest, etc. are naturally interpretable. For instance, coefficient weights in regression provide the importance of the features. Similarly, the classical feature selection approaches [Yang and Pedersen, 1997; Guyon and Elisseeff, 2003] also help in explaining model decisions since they provide the importance and contribution of individual features. Recent works by Wang and Rudin [2015]; Letham et al. [2015]; Lakkaraju et al. [2016, 2017] propose methods to generate decision lists which improve the interpretability over decision trees.

Many existing works have also explored the problem of interpreting black-box models. A method for explaining predictions of black-box models for individual instances is presented in Robnik-Šikonja and Kononenko [2008]. Similarly, another method proposed in Baehrens et al. [2010] explains the decision taken by arbitrary nonlinear classifiers. Another method for explaining the predictions of any classifier is proposed in Ribeiro et al. [2016]. It approximates the black-box model locally around the prediction. Further, Samek et al. [2017] propose techniques to explain predictions of deep learning models.

On the other hand, few methods for argument mining have focused on automatically extracting evidence which support the factual claims made in a debate. Supervised learning methods for achieving this for claims on social media and debate platforms are presented in Rinott et al. [2015]; Addawood and Bashir [2016]. Similarly, methods proposed in Cartright et al. [2011]; Bellot et al. [2013] address this problem from the information retrieval perspective. Their techniques use a collection of documents to retrieve evidence to support a claim.

In this thesis, we follow the direction similar to evidence retrieval approaches. Given a claim, our models judiciously extract snippets from the relevant articles which help in explaining their automated assessment.

3
Credibility Analysis Framework

Contents

3.1 Introduction

While prior work on truth discovery has focused on the case of checking factual statements, here we address the novel task of assessing the credibility of arbitrary claims made in natural-language – in an open-domain setting without any assumptions about the structure of the claim, or the community where it is made.

In this chapter, we propose a generic framework for credibility analysis. This framework is based on automatically finding relevant articles from the web (including news and social media), and analyzing them for assessing the credibility of a claim (i.e., *true* or *false*). Our preliminary model for credibility assessment leverages the joint interaction between the language of articles about the claim and the reliability of the underlying web sources. Experiments with claims from the popular website *snopes.com* and from reported cases of Wikipedia hoaxes demonstrate the viability of our framework and its superior accuracy over various baselines.

Claim: Solar panels drain the sun's energy, experts say
Assessment: False
Explanation: Solar panels do not suck up the Sun's rays of photons. Just like wind farms do not deplete our planet of wind. These renewable sources of energy are not finite like fossil fuels. Wind turbines and solar panels are not vacuums, nor do they divert this energy from other systems. (*iflscience.com*)

Table 3.1: A sample claim with assessment and manually extracted explanation.

State of the Art and its Limitations: As described in Chapter 2, prior work on credibility analysis (see Li et al. [2015b] for a survey) has focused on factual claims (e.g., Li et al. [2011b, 2012, 2015c]) and/or online communities with specific characteristics like user metadata, who-replied-to-whom, who-edited-what, etc. (e.g., Mukherjee et al. [2014]; Kumar et al. [2016]). Truth-finding methods of this kind, starting with the seminal work of Yin et al. [2008], assume that claims follow a structured template with clear identification of the questionable values [Li et al., 2011b, 2012], or correspond to subject-predicate-object triples obtained by information extraction [Nakashole and Mitchell, 2014]. A classic example is *"Obama is born in Kenya"* viewed as a triple $\langle$*Obama, born in, Kenya*$\rangle$ where *"Kenya"* is the critical value. The assumption of such a structure is crucial in order to identify alternative values for the questionable slot (e.g., *"Hawaii"*, *"USA"*, *"Africa"*), and is appropriate when checking facts for tasks like knowledge base curation. However, these approaches are limited in their coverage and cannot handle many kinds of claims found on news and social media, which are often in the form of long sentences or entire paragraphs.

Overview of our approach: To address these limitations, we present a novel framework to assess the credibility of *textual claims*, in an *open-domain* setting, where we do not assume any community-specific characteristics or structure in the input data. Given a claim in the form of a sentence, we first use a search engine to identify documents from multiple web-sources, which are relevant to the claim. We refer to these documents as *reporting articles*. Then, we individually analyze these evidence to determine their opinions regarding the credibility of the input claim and finally, we aggregate these individual opinions to determine the overall credibility of the claim (see Section 3.3 and Section 3.4). Figure 3.1 gives a pictorial overview of our framework.

We perform experiments with claims from the fact-checking website *snopes.com* and with data about hoaxes and fictitious persons in Wikipedia. The performance of our model demonstrates major improvements in accuracy over various baselines (see Section 3.5 and Section 3.6).

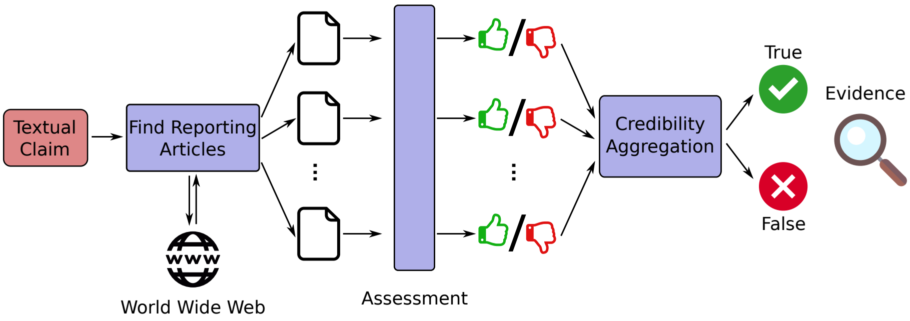


Figure 3.1: Overall system framework for credibility assessment.

3.2 Problem Statement

Given a natural language claim (or a factual statement) and a set of relevant web articles, our objective is to assess the credibility of the claim and determine whether it is *true* or *false*. Moreover, we want to automatically extract user-interpretable evidence which explain the automated assessment.

Table 3.1 shows an example of an input and output of our method. For the given example, we assess its credibility as *false* and provide user-interpretable explanation in the form of informative snippets – automatically extracted from relevant web-articles. However, in this chapter, we address this problem only partially. Our preliminary model is restricted to only computing the binary credibility verdict (*true* or *false*) without providing any explanations.

3.3 Components of Framework

Our framework for credibility assessment consists of the following components:

- **Claim (C):** A fact or an assertion in natural language form. For example, *"The use of solar panels drains the sun of energy"*.

- **Articles (A):** A set of relevant web-articles which discuss or report about the claim. For example, an article[1] from the *iflscience.com* website:

> *"An article has been circulating on the net for the last few days, released by National Report, entitled Solar Panels Drain the Suns Energy, Experts Say. While at first glance it might look genuine because it includes the names of institutions and quotes..."*

[1] https://www.iflscience.com/environment/no-solar-panels-will-not-drain-suns-energy/ (accessed July 8, 2019)

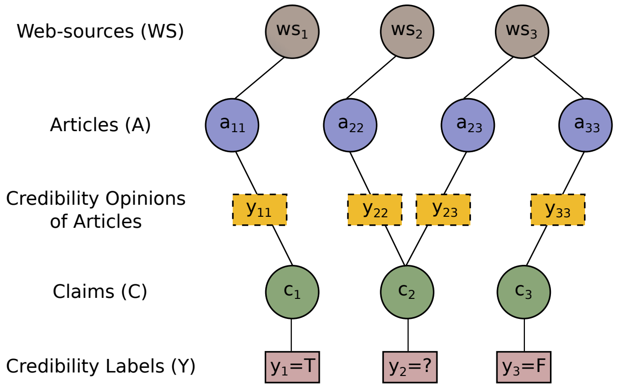


Figure 3.2: Components of the credibility analysis framework.

- **Article Sources (WS):** A set of web-sources publishing the relevant web-articles. For example, the website *iflscience.com* is the article source for the above-mentioned evidence article.

Consider a set of textual claims $\langle C \rangle$ in the form of sentences, and a set of web-sources $\langle WS \rangle$ containing relevant articles $\langle A \rangle$ that report on the claims. Let $a_{ij} \in A$ denote an article of web-source $ws_j \in WS$ about claim $c_i \in C$. Each claim c_i is associated with a binary random variable y_i that depicts its credibility label, where $y_i \in \{T, F\}$ (T stands for *True*, whereas F stands for *False*). Each article a_{ij} is associated with a random variable y_{ij} that depicts the opinion (*true* or *false*) of the article a_{ij} (from ws_j) regarding the credibility of c_i – when considering only this article. Figure 3.2 illustrates this model. Given the labels of a subset of the claims (e.g., y_1 for c_1, and y_3 for c_3), our objective is to predict the credibility label of the remaining claims (e.g., y_2 for c_2).

3.4 Credibility Assessment Model

Our preliminary model for credibility assessment incorporates the following factors that help in determining the credibility of a claim:
i) How is the claim reported? The *writing style* of the articles reporting the claim gives important clues about the credibility of the claim. For example, related work in detecting biased language [Recasens et al., 2013] and credibility analysis in closed communities [Mukherjee et al., 2014; Mukherjee and Weikum, 2015] leverage linguistic features like discourse, subjectivity, and modality.
ii) Who is reporting the claim? The *provenance* of the claim coupled with the *reliability* of the source plays a key role in understanding its credibility. For instance, *theonion.com* is known to publish satirical articles, whereas *wikipedia.org* usually provides objective information according to its *Neutral Point of View* policy.

To learn the parameters in our credibility assessment model, we use *Distant Supervision* to attach observed true/false labels of claims to corresponding reporting articles and learn a *Credibility Classifier*. In this process, we need to (a) understand the language of the articles, and (b) consider the *reliability* of the underlying web sources reporting these articles. Thereafter, we (c) compute the credibility opinion scores of individual articles, and finally, (d) *aggregate* these scores from all articles to obtain the overall credibility label of target claims. The following sections describe the features used in our model and how we learn the parameters.

3.4.1 Language Stylistic Features

The style in which a claim is reported in an article plays a critical role in understanding its credibility. A true claim is assumed to be reported in an objective and unbiased language. On the other hand, if a claim is reported in a highly subjective or a sensationalized style, then it is likely to be less credible. This hypothesis is validated in Nakashole and Mitchell [2014] through an experiment using Amazon Mechanical Turk.

In order to capture the linguistic style of the reporting articles to model the above hypothesis, we use the set of lexicons from Mukherjee and Weikum [2015], in particular, the following types of stylistic features:

- **Assertive verbs:** They capture the degree of certainty to which a proposition holds (e.g., “suppose”).
- **Factive verbs:** These words presuppose the truth of a proposition in a sentence (e.g., “know”).
- **Hedges:** These are mitigating words which soften the degree of commitment to a proposition (e.g., “may”).
- **Implicatives:** These words trigger presupposition in an utterance (e.g., “decline”).
- **Report verbs:** These words emphasize the attitude towards the source of the information (e.g., “argue”).
- **Discourse markers:** They capture the degree of confidence, perspective, and certainty in the set of propositions made (e.g., “therefore”).
- **Subjectivity and bias:** a list of positive and negative opinionated words, and an affective lexicon to capture the state of mind (like attitude and emotions) of the writer while writing an article.

Feature vector construction: For each article a_{ij}, we compute the normalized frequency of all the linguistic features $\langle f_k \rangle$. Given all the stylistic language features, we compute,

$$F^L(a_{ij}) = \langle freq_{a_{ij}}^{f_k} = n_{a_{ij}}^{f_k} / length(a_{ij}) \rangle$$

where $n_{a_{ij}}^{f_k} = number\ of\ times\ f_k\ occurs\ in\ a_{ij}$.

3.4.2 Source Reliability

Apart from the reporting style of the evidence article, the reliability of the web-source hosting the article also has a significant impact on the credibility of the claim. For instance, one should not believe a claim reported by an article from the "The UnRreal Times" website[2], as opposed to a claim on the "World Health Organization" website.

To capture the reliability of the web-source for each evidence article, we determine the AlexaRank and PageRank of its source and use them as proxies for the source reliability. AlexaRank[3] is based on a combined measure of unique visitors and page views of the website. PageRank determines the importance of the website by counting the number and quality of links to and from the website. To avoid modeling from sparse observations, we combine all the web-sources having less than 10 articles in the dataset to a single web-source.

Feature vector construction: For each article a_{ij}, we capture the identity of its web-source ws_j using a one-hot vector of dimension $cardinality(< WS >)$ (i.e., 1357 - after collapsing the "long-tail" sources to a single source) by setting the j^{th} element in the vector to 1, and the remaining ones to 0. We also use the AlexaRank and PageRank of the web-source as additional features capturing the source reliability.

$$F^{SR}(a_{ij}) = \langle 0 \ldots, ws_j = 1, 0 \ldots, \log PR_{ws_j}, \log AR_{ws_j} \rangle$$

where, PR and AR represent the PageRank, and the AlexaRank, respectively.

3.4.3 Credibility Classification Using Distant Supervision

Credibility labels are available *per-claim*, and not per-reporting-article. Thus, in our approach for credibility aggregation from multiple sources, we use *Distant Supervision* for *training* — whereby we attach the (observed) label y_i of each claim c_i to each article a_{ij} reporting the claim (i.e., setting labels $y_{ij} = y_i$). For instance, in Figure 3.1, $y_{11} = y_1 = T, y_{33} = y_3 = F$. Using these $\langle y_{ij} \rangle$ as the corresponding

[2] A satire, spoof, parody and humor portal: http://www.theunrealtimes.com/ (accessed July 8, 2019)

[3] https://support.alexa.com/hc/en-us/articles/200449744-How-are-Alexa-s-traffic-rankings-determined- (accessed July 8, 2019)

Type of Feature	Number of Features
Linguistic	
Assertive Verbs	66
Factive Verbs	27
Hedges	100
Implicatives	32
Report Verbs	181
Discourse Markers	13
Subjectivity and Bias	8770
Reliability	
Source Identity	1357
PageRank	1
AlexaRank	1

Table 3.2: Statistics of features used in our model.

training labels for $\langle a_{ij} \rangle$, with the corresponding feature vectors $\langle F^L(a_{ij}) \cup F^{SR}(a_{ij}) \rangle$, we train an L_1-regularized logistic regression model on the training data. Statistics of features used in our model are given in Table 3.2.

For any *test* claim c_i whose credibility label is unknown, and its corresponding reporting articles $\langle a_{ij} \rangle$, we use this *Credibility Classifier* to obtain the corresponding credibility opinions $\langle y_{ij} \rangle$ of the articles. We determine the overall credibility label y_i of c_i by considering a sum of *per-article* credibility probabilities:

$$y_i = \underset{l \in \{T,F\}}{\arg\max} \sum_{a_{ij}} Prob(y_{ij} = l) \tag{3.1}$$

3.5 Case Studies

3.5.1 Snopes

We performed experiments with data from a fact checking website: *snopes.com. Snopes* covers Internet rumors, hoaxes, urban legends, e-mail forwards, and other stories of unknown or questionable origin. It is a well-known resource for validating and debunking such stories, receiving around 300,000 visits a day. They typically collect rumors and claims from *Facebook*, *Twitter*, *Reddit*, news websites, e-mails by users, etc.

Each article verifies a single claim, e.g., *"North Carolina no longer considers the $20 bill to be legal tender"*. The Snopes editors assign a *manual* credibility verdict

Total claims	4856
True claims	1277 (26.3%)
False claims	3579 (73.7%)
Web articles	133272
Avg. articles per claim	27.44

Table 3.3: *Snopes* data statistics.

	Hoaxes	**Fictitious People**
Total Claims	100	57
Web articles	2813	1552
Avg. articles per claim	28.13	27.22

Table 3.4: *Wikipedia* data statistics.

to each such claim: *True* or *False*. Few of the claims have labels like *Mostly True* or *Mostly False*. We map *Mostly True* labels to *True*, and *Mostly False* labels to *False* — thereby considering only *binary* credibility labels for this work. Claims having labels like *Partially True* or *Partially False* are ignored. The credibility verdict is accompanied by a description how the editor(s) came across the claim (e.g., it was collected from a Facebook post, or received by email, etc.), an *Origin* section describing the origin of the claim, and an *Analysis* section justifying the verdict. Our model is agnostic of the structure of Snopes as we use only the claim and its credibility verdict, ignoring all other related information.

We collected data from *Snopes* published until February 2016. For each claim c_i, we fired the *claim text* as a *query* to the Google search engine and extracted the first three result pages (i.e., up to 30 articles) as a set of reporting articles $\langle a_{ij} \rangle$. We ignore the ranking information in the set of collected articles to have minimal dependency on the search engine. Other search engines or other means of evidence gathering can easily be used. We then crawled all these articles from their corresponding web-sources $\langle ws_j \rangle$. We removed search results from the *snopes.com* domain to avoid any kind of bias. Statistics of the data crawled from *snopes.com* is given in Table 3.3.

3.5.2 Wikipedia

We collected a set of 100 proven hoaxes reported on Wikipedia[4], e.g., *"Alien autopsy film by Ray Santilli"*, *"Disappearing blonde gene"* etc. All these hoaxes can be mapped to claims of types: "*<ENTITY> exists*", "*<ENTITY> is genuine*" or "*<EVENT> occurred*". While collecting the data, hoaxes not falling under these categories were ignored. Words related to hoaxes, e.g., *false*, *fictional*, *nonexistent*, etc. were removed from the claim description to avoid any kind of search bias while retrieving articles using a search engine. Since the dataset contains only hoaxes, the ground-truth label for all of these claims is *False*.

In addition, we also collected a set of 57 fictitious people as reported on the Wikipedia page[5], e.g., *"Ern Malley, an Australian poet"*, *"P. D. Q. Bach, a composer"* etc. All these entities can be mapped to claims of type: "*<ENTITY>*

[4] https://en.wikipedia.org/wiki/List_of_hoaxes#Proven_hoaxes (accessed 8 July, 2019)

[5] https://en.wikipedia.org/wiki/List_of_fictitious_people (accessed 8 July, 2019)

exists". The ground-truth label for all of these claims is *False* as the dataset contains only fictitious people.

Table 3.4 reports the statistics of the dataset. As described earlier, we used a search engine to get a set of reporting articles for these claims. Similar to the previous case, we removed results from the *wikipedia.org* domain. Note that we trained our *Credibility Classifier* on *Snopes* data, and tested it on this data from *Wikipedia* — thereby demonstrating that our model generalizes and can be easily applied to data from other domains.

3.6 Experiments

We conducted a set of experiments using data from *Snopes* and *Wikipedia* to test the performance of our method.

Evaluation Measures: We train our models with *Snopes* data, and report standard 10-fold cross-validation accuracy on both the datasets. *Snopes*, primarily being a hoax debunking website, is biased towards (refuting) the *False* claims. Therefore, we also report the per-class accuracy and the *macro-averaged accuracy* which is the average of *per-class* accuracy — giving equal weight to both classes irrespective of the data imbalance. We also report the Area-under-Curve (AUC) values of the ROC (Receiver Operating Characteristic) curve. To highlight the effectiveness of our model in identifying false claims (i.e., hoaxes, rumors, etc.), we also report the precision, recall and F1 score for the *False* claim class.

3.6.1 Credibility Assessment: Snopes

While performing 10-fold cross-validation on the claims, we trained on any 9-folds of the data — where the algorithm learned the *Credibility Classifier* and web-source reliabilities from the reporting articles and their corresponding sources present in the training split. In order to remove any training bias, we ignored all *Snopes*-specific references from the data and the search engine results.

For addressing the data imbalance issue, we adjust the classifier's loss function. We place a large penalty for misclassifying instances from the *true* class which boosts certain features from that class. The overall effect is that the classifier makes fewer mistakes for *true* instances, leading to balanced classification. We set the penalty for the *true* class to 2.8 — given by the ratio of the number of *false* claims to *true* claims in the *Snopes* data.

We compare to the following baselines:

- **ZeroR:** This is a trivial baseline, designed for imbalanced data, that always labels a claim as the class with the largest proportion, i.e., *false* in our case. The overall accuracy of this baseline is **73.69%**, and the macro-averaged accuracy is **50%**.

- **FactChecker:** Recent work on fact-checking [Nakashole and Mitchell, 2014] relies on the hypothesis that claims reported by objective articles are more likely to be true than those reported in subjective articles. The authors extracted *alternative* fact candidates for the given claim and used the hypothesis to rank all candidates. This approach works well in their use case of knowledge base curation, as all the claims are factual and have the form of Subject-Predicate-Object (SPO) triples. On the other hand, the claims in our case are textual snippets without any explicit alternative candidates. Therefore, we could only implement this method as a baseline "in spirit". To this end, we used the code[6] of Mukherjee and Weikum [2015] to construct an "Objectivity Detector". Given a claim and a set of reporting articles, the target claim was labeled *true* if the sum of the objectivity scores of its reporting articles — as determined by the Objectivity Detector — was higher than the sum of the subjective scores, and *false* otherwise. This approach resulted in **55.29%** overall accuracy and **56.27%** macro-averaged accuracy for credibility classification.

Along with the above baselines, we also report the results of our model with different feature configurations for linguistic style and web-source reliability:

- Model using only *language* (LG) features,

- Model using only *web-source reliability* (SR) features,

- Aggregated model with the combination of, *language* and *source reliability* (LG + SR) features.

Table 3.5 shows the 10-fold cross-validation accuracy of various baselines against different configurations of our model, with the ROC curves plotted in Figure 3.3. From the results, we observe that using only language stylistic features (LG) is not sufficient; it is important to understand the source reliability (SR) of the article as well. High precision score for the *False* claim class shows the strength of our model in detecting *False* claims.

[6] Code: http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/impact/credibilityanalysis/ (accessed 8 July, 2019)

Configuration	Overall Accuracy (%)	*True* Claims Accuracy (%)	*False* Claims Accuracy (%)	Macro-averaged Accuracy (%)	AUC	*False* Claims Precision	*False* Claims Recall	*False* Claims F1-Score
LG + SR	71.96	75.43	70.77	**73.10**	**0.80**	0.89	0.71	0.79
LG	69.43	66.47	70.55	68.51	0.75	0.85	0.71	0.77
SR	66.52	68.56	65.90	67.23	0.73	0.85	0.66	0.74
FactChecking	55.29	58.34	54.21	56.27	0.58	0.78	0.54	0.64
ZeroR	73.69	00.00	100	50.00	0.50	0.74	1.00	0.85

Table 3.5: Performance comparison of our model vs. related baselines with 10-fold cross-validation on Snopes dataset. LG: language stylistic features, SR: web-source reliability.

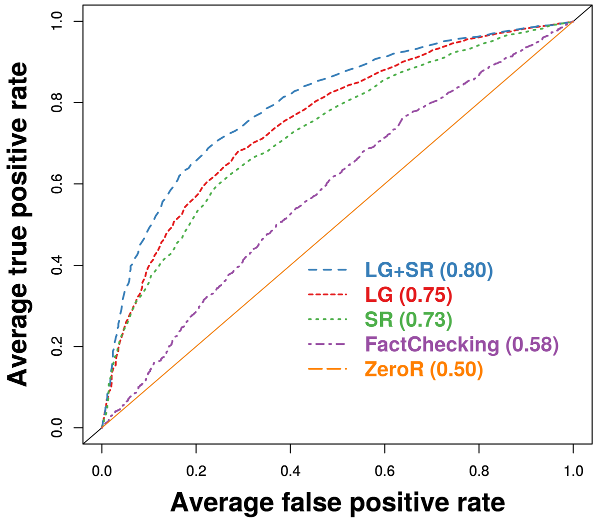


Figure 3.3: ROC curves for different model configurations for Snopes dataset.

Test Data	#Claims	Accuracy (%)
Wiki Hoaxes	100	84.00
Wiki Fictitious People	57	66.07

Table 3.6: Accuracy of credibility classification on *Wikipedia* dataset.

3.6.2 Credibility Assessment: Wikipedia

To demonstrate the generality of our approach, the model trained on the *Snopes* dataset was tested on the *Wikipedia* dataset of hoaxes and fictitious persons. The results are shown in Table 3.6. Similar to the *Snopes* setting, we removed all references to Wikipedia from the data and the search engine results. As we can see from the results, our system is able to detect hoaxes and fictitious people with high accuracy, although the claim descriptions here are stylistically quite different from those of *Snopes*.

Claim	Verdict & Evidence
A woman stabbed her boyfriend with a sharpened selfie stick because he didn't like her newest Instagram selfie quickly enough.	**[Verdict]: False** **[Evidence]:** A weird kind of story in heavy circulation online states ... No, the claim is not a fact.
90% of people in the U.S. marry their high school sweethearts.	**[Verdict]: False** **[Evidence]:** The school category resulted in only 14% of total respondent base. In analyzing these surveys, one must realize that potential biases in survey methods exist, such as ... It seems absolutely clear that these and other surveys conducted in early 1990s represent nowhere nearly close to 90% ...
A Facebook coupon offering 50% off at Target stores is real.	**[Verdict]: False** **[Evidence]:** The newest questionable offer to take hold of Facebook newsfeeds involves the false promise of a coupon ... A rep for Target HQ confirms to Consumerist that there is no such coupon and this is a fake.
Two Maryland sheriff's deputies were fatally shot and a suspect killed on Wednesday in a shootout at a Baltimore-area Panera restaurant.	**[Verdict]: True** **[Evidence]:** Two Maryland sheriff's deputies were fatally shot and a suspect killed Wednesday in a shootout at a Baltimore-area Panera restaurant filled with lunchtime customers. (Reuters) Authorities found a semiautomatic handgun in Evans's vehicle, which he might have been living in.
A dying child was made an honorary fireman by the Phoenix Fire Department.	**[Verdict]: True** **[Evidence]:** We'll make him an honorary Fireman for the day. He can come down to the fire station, eat with us, go out on all the fire calls, the whole nine yards! The Fire Chief decided that the Phoenix Fire Department should make sure the dying boy had an experience truly befitting a fireman.
A declared-dead jockey returned to the track and shocked the grandstand crowd.	**[Verdict]: True** **[Evidence]:** When the crowd realized that the shirtless, bloodied, toe-tagged man who was staggering across the grandstand area was the jockey who had been declared dead about a half hour earlier, the crowd and the race officials rushed towards Neves, as shock turned to celebration.

Table 3.7: Snapshot of claims with assessment from Credibility Classifier, and manually annotated snippets as evidence.

3.7 Error Analysis and Discussion

Poor performance on detecting false claims: As we see from the results, the system accuracy for detecting false claims is low compared to that for the true claims. While performing an error analysis of the results, we observed that many of the well-written articles from reputed web-sources refer to the false claims in the *negated* form such as "...the company's spokesperson *denied* that...". Our model does not capture these finer linguistic aspects like implicit or explicit negation, and, therefore, commits mistakes. In future, we would like to propose features which capture these finer semantics of the article text to have a more accurate system.

Marginal contribution of web-source reliability: Results also indicate that the performance of the full model configuration (LG+SR) achieves only slight improvement over the configuration LG. This can be attributed to the fact that these rank measures (PageRank and AlexaRank) capture the authority and popularity of the web-sources, but not their reliability from the credibility point of view. For example, the PageRank of the satirical news website *The Onion* is very high (7 out of 10). However, this does not indicate anything about its reliability. Hence, as future work, it would be interesting to design an algorithm which automatically captures the ranking of web-sources based on their credibility.

Understanding the credibility assessment output: While performing error analysis, we observed that the probability scores do not help in understanding the output. This is also true for related truth-finding approaches. It would thus be nice to have interpretable evidence as an additional output of the system which can explain the credibility assessment. Table 3.7 gives a snapshot of claims with the credibility assessment given by our system, along with manual annotation of snippets that can be used as evidence. As future work, we want to automate this process of generating evidence.

3.8 Conclusions

In this chapter, we proposed a generic framework for credibility analysis of unstructured textual claims in an open-domain setting. Our approach for credibility analysis makes use of the language style and source reliability of evidence articles reporting the claim to assess its credibility.

Experiments on analyzing the credibility of real-world claims, from the fact-checking website *Snopes*, and on hoaxes and fictitious persons listed on *Wikipedia*, demonstrate the effectiveness of our approach. As future work, we want to estimate source reliability from the credibility perspective, investigate the role of refined linguistic aspects like negation, and understanding the article's perspective about the claim. We further enhance our approach for credibility analysis in Chapter 4.

4
Feature-Based Credibility Assessment

Contents

4.1 Introduction

Our model for automated credibility assessment, proposed in the last chapter, requires that sources of evidence or counter-evidence are easily retrieved from the web. It disregards the crucial cues for assessing the credibility: the stance of the article towards the claim and the reliability of the underlying web-sources. Moreover, it can not cope with newly emerging claims, and it does not provide user-interpretable explanations for its verdict on the claim’s credibility.

In this chapter, we enhance our approach proposed in Chapter 3 and overcome these limitations by automatically assessing the credibility of emerging claims, with sparse presence in web-sources, and generating suitable explanations from judiciously selected sources. To this end, we retrieve diverse evidence articles about the claim and model the mutual interaction between the stance (i.e., support or refute) and the language style of the evidence articles, the reliability of the sources, and the claim's temporal footprint on the web. Extensive experiments demonstrate the viability of our method and its superiority over prior works. We show that our methods work well for early detection of emerging claims, as well as for claims with a limited presence on the web and social media.

State of the Art and its Limitations: As described in Chapter 2, within prior work on credibility analysis (e.g., Dong et al. [2015]; Li et al. [2011b, 2012, 2015c]), the important aspect of providing explanations for credibility assessments has not been addressed. In most works, the analysis focuses on structured statements and exhibits major limitations: (i) claims take the form of subject-predicate-object triples [Nakashole and Mitchell, 2014] (e.g., `Obama_BornIn_Kenya`), (ii) questionable values for the object are easy to identify [Li et al., 2011b, 2012] (e.g., `Kenya`), (iii) conflicts and alternative values are easy to determine [Yin et al., 2008] (e.g., `Kenya` vs. `USA`) and/or (iv) domain-specific metadata is available (e.g., user metadata in online communities such as who-replied-to-whom) [Mukherjee et al., 2014; Kumar et al., 2016].

In our own prior work [Popat et al., 2016] (in Chapter 3), we addressed some of these limitations by assessing the credibility of *textual claims*: arbitrary statements made in natural language in arbitrary kinds of online communities or other web-sources. Based on automatically found evidence from the web, our method could assess the credibility of a claim. However, like all other prior works, we restricted ourselves to computing a binary verdict (true or false) without providing explanations. Moreover, we assumed that we could easily retrieve ample evidence or counter-evidence from a (static) snapshot of the web, disregarding the dynamics of how claims emerge, spread, and are supported or refuted (i.e., the stance of a web-source towards the claim).

Overview of our approach: In this chapter, we overcome the limitations of these prior works (including our own [Popat et al., 2016]; in Chapter 3). We assess the credibility of newly emerging and "long-tail" claims with a sparse presence on the web by determining the *stance, reliability, and trend* of retrieved sources of evidence or counter-evidence, and by providing user interpretable *explanations* for the credibility verdict.

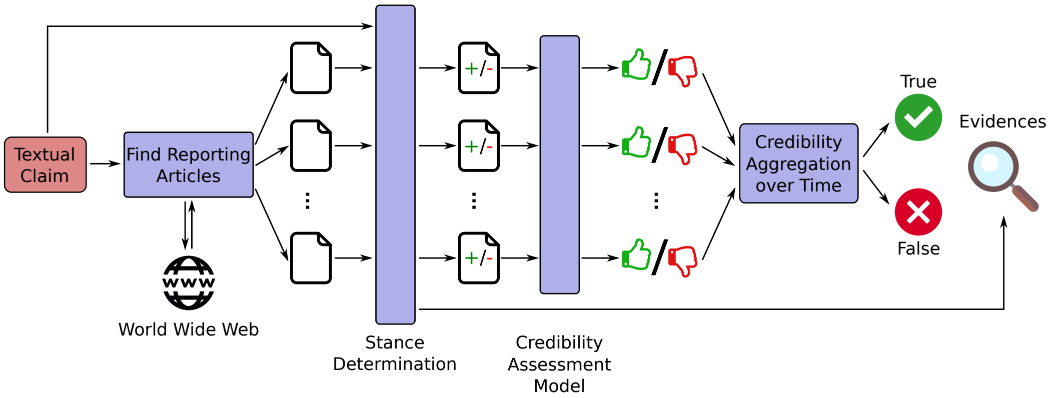


Figure 4.1: System framework for credibility assessment (+/- labels for articles indicate the stance i.e support/refute towards the claim).

Claim: Facebook soon plans to charge monthly subscription fees to users of the social network.
Assessment: False
Explanation: The rumor that Facebook will suddenly start charging users to access the site has become one of the social media eras perennial chain letters. (*cnn.com*)

Table 4.1: A sample claim with assessment and explanation.

Table 4.1 shows an example of the input and output of our method. For the given example, our model assesses its credibility as *false* and provides user-interpretable explanations in the form of informative snippets automatically extracted from an article published by a reliable web-source refuting this claim — exploiting the interplay between multiple factors to show the explanation.

Our method works as follows. Given a newly emerging claim in the form of a sentence at time t, we first use a search engine to identify documents from diverse web-sources referring to the claim. We refer to these documents as *reporting articles*. For assessing the credibility of the emerging claim, our model captures the interplay between several factors: the *language* of the reporting articles (e.g., bias, subjectivity, etc.), the *reliability* of the web-sources generating the articles, and the *stance* of the article towards the claim (i.e., whether it supports or refutes the claim). We propose two inference methods for the model: *Distant Supervision* and joint inference with a *Conditional Random Field* (CRF). The former approach learns all the factors sequentially, whereas the latter treats them jointly.

To tackle emerging claims and consider the temporal aspect, we harness the temporal footprint of the claim on the web, i.e., the dynamic trend in the timestamps of reporting articles that support or refute a claim. Finally, a joint method combines the content- and trend-aware models.

As evidence, our model extracts informative snippets from relevant reporting articles for the claim published by reliable sources, along with the stance (supporting or refuting) of the source towards the claim. Figure 4.1 gives a pictorial overview of the overall model. Extensive experiments with claims from the fact-checking website *snopes.com* and *wikipedia.com* demonstrate the strengths of our content-aware and trend-aware models by achieving significant improvements over various baselines. By combining them, we achieve the best performance for assessing the credibility of newly emerging claims. We show that our model can detect emerging *false* or *true* claims with a macro-averaged accuracy of 80% within 5 days of its origin on the web, with as low as 6 reporting articles per-claim.

Novel contributions of this chapter can be summarized as:

- Exploring the interplay between factors like language, reliability, stance, and trend of sources of evidence and counter-evidence for credibility assessment of textual claims (see Section 4.3).
- Probabilistic models for joint inference over the above factors that give user-interpretable explanations (see Section 4.4).
- Experiments with real-world emerging and long-tail claims on the web and social media (see Section 4.5).

4.2 Model and Notation

Our approaches based on distant supervision and CRF exploit the rich interaction taking place between various factors like source reliability and stance over time, article objectivity, and claim credibility for the assessment of claims. Figure 4.2 depicts this interaction. Consider a set of textual claims $\langle C \rangle$ in the form of sentences or short paragraphs, and a set of web-sources $\langle WS \rangle$ containing articles $\langle A^t \rangle$ that report on the claims at time t.

The following edges between the variables, and their labels, capture their interplay:

- Each claim $c_i \in C$ is connected to its reporting article $a^t_{ij} \in A^t$ published at time t.
- Each reporting article a^t_{ij} is connected to its web-source $ws_j \in WS$.
- For the joint CRF model, each claim c_i is also connected to the web-source ws_j that published an article a^t_{ij} on it at time t.

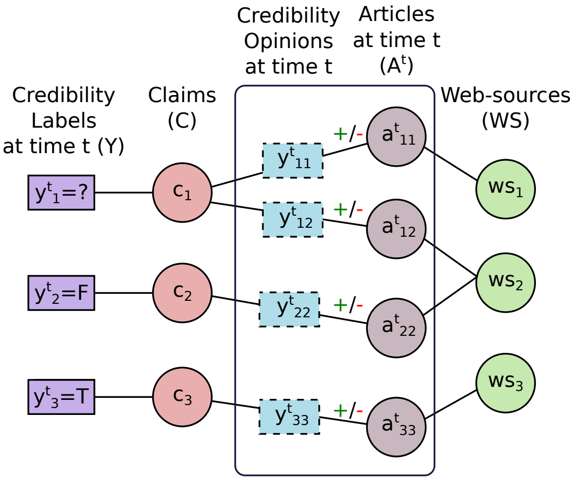


Figure 4.2: Factors for credibility analysis (+/- labels for edges indicate the article's stance i.e support/refute for the claim).

- Each article a_{ij}^t is associated with a random variable y_{ij}^t that depicts the *credibility opinion* (*True* or *False*) of the article a_{ij}^t (from ws_j) regarding c_i at time t — considering both the *stance* and *language* of the article.

- Each claim c_i is associated with a binary random variable y_i^t that depicts its *credibility label* at time t, where $y_i^t \in \{T, F\}$ (T stands for *True*, whereas F stands for *False*). y_i^t aggregates the individual credibility assessment y_{ij}^t of the articles a_{ij}^t for c_i at time t taking into account the reliability of their web-sources.

Problem statement: Given the labels of a subset of the claims (e.g., y_2^t for c_2, and y_3^t for c_3), our objective is to predict the credibility label of the newly emerging claim (e.g., y_1^t for c_1 at each time point t). The article set $\langle A^t \rangle$ and its predicted credibility label y^t for the newly emerging claim changes with time t as the evidence evolves.

4.3 Credibility Assessment Factors

We consider various factors for assessing the credibility of a textual claim. The following sections explain these factors.

4.3.1 Language Stylistic Features

The credibility of textual claims heavily depends on the style in which it is reported. A true claim is assumed to be reported in an objective and unbiased language. On the other hand, highly subjective or sensationalized style of writing diminishes the credibility of a claim [Nakashole and Mitchell, 2014]. We use the same language features (F^L) (e.g., a set of assertive and factive verbs, hedges, report verbs, subjective and biased words, etc.) as our prior work [Popat et al., 2016] (see Section 3.4.1) to capture the linguistic style of the reporting articles:

- *Assertive and factive verbs* (e.g., "claim", "indicate") capture the degree of certainty to which a proposition holds.
- *Hedges* are the mitigating words (e.g., "may") which soften the degree of commitment to a proposition.
- *Implicative words* (e.g., "preclude") trigger presupposition in an utterance.
- *Report verbs* (e.g., "deny") emphasize the attitude towards the source of the information.
- *Discourse markers* (e.g., "could", "maybe") capture the degree of confidence, perspective, and certainty in the statements.
- Lastly, a lexicon of *subjectivity and bias* capture the attitude and emotions of the writer while writing an article.

4.3.2 Finding Stance and Evidence

In order to assess the credibility of a claim, it is important to understand whether the evidence articles reporting the claim are supporting it or not. For example, an article from a reliable source like *truthorfiction.com* refuting the claim will make the claim less credible.

In order to understand the stance of an article, we divide the article into a set of snippets, and extract the snippets that are strongly related to the claim. This set of snippets helps in determining the overall score with which the article refutes or supports the claim. We compute both the support and refute scores, and use them as two separate features in our model.

The method for stance determination is outlined in Algorithm 1. Step 3 of the algorithm ensures that the snippets we consider are related to the claim. It removes snippets having overlap less than a threshold (η), where we consider all unigrams and bigrams for the overlap measure. In case all the snippets are removed in Step 3, we ignore the article. We varied η from 20% to 80% on withheld tuning data, and found $\eta = 40\%$ to give the optimal performance.

In Step 4, we use a *Stance Classifier* (described in the next section) to determine whether a snippet $s \in S \setminus S^{'}$ supports or refutes the claim. Let p_s^+ and p_s^- denote the corresponding support or refute probability of a snippet s coming from the

Algorithm 1 Stance Determination Method

```
Input: Claim $c_i$ and a corresponding reporting article $a^t_{ij}$ at time $t$
Output: Stance scores (support & refute) of $a^t_{ij}$ about $c_i$
1: Given $a^t_{ij}$, generate all possible snippets $\langle S \rangle$ of up to four consecutive sentences
2: Compute unigram & bigram overlap $\langle O \rangle$ of $c_i$ with each snippet in $\langle S \rangle$
3: Remove snippets $\langle S' \rangle$ with percentage overlap $o_s$ with $c_i < \eta$
4: For each remaining snippet $s \in S \setminus S'$, calculate its stance (support or refute)
   using a stance classifier
5: For each such snippet $s$, compute a combined score as the product of its stance
   probability and overlap score
6: Select top-k snippets $\langle S_{topK} \rangle$ based on the combined score
7: Return the average of stance support & refute scores of snippets in $\langle S_{topK} \rangle$
```

classifier. We combine the stance probability of each snippet s with its overlap score o_s with the target claim: $\langle p_s^+ \times o_s, p_s^- \times o_s \rangle$. Then, we sort the snippets based on $\max(p_s^+ \times o_s, p_s^- \times o_s)$ and retrieve the top-k snippets S_{topK}. In our experiments (in Section 4.5), we set k to five. The idea is to capture the snippets which are highly related to the claim, and also have a strong refute or support probability.

Evidence: In the later stage, these snippets in $\langle S_{topK} \rangle$ are used as evidence supporting the result of our credibility classifier.

Feature vector construction: For each article a^t_{ij}, we average the two stance probabilities (for support and for refute) over the top-k snippets $s \in S_{topK}$ as two separate features: $F^{St}(a^t_{ij}) = \langle avg(\langle p_s^+ \rangle), avg(\langle p_s^- \rangle) \rangle$.

4.3.2.1 Stance Classifier

Goal: Given a piece of text, the stance classifier should give the probability of how likely the text refutes or supports a claim based on the linguistic features.

Data: Hoax debunking websites like *snopes.com*, *truthorfiction.com*, and *politifact.com* compile articles about contentious claims along with a manual analysis of the origin of the claim and its corresponding credibility label. We extract these analysis sections from such sources along with their manually assigned credibility labels (*true* or *false*). The *Stance Classifier* used in Step 4 of Algorithm 1 is trained using this dataset (withheld from the test cases later used in experiments). The articles confirming a claim are used as positive instances for the *"support"* class, whereas the articles debunking a claim are used as negative instances for the *"refute"* class.

Features: We consider all the unigrams and bigrams present in the training data as features, *ignoring all the named entities* (with part-of-speech tags NNP and NNPS). This is to prevent overfitting the model with popular entities (like "obama", "trump", "iphone", etc.) which frequently appear in hoax articles.

Model: We use the L_2 regularized Logistic Regression (primal formulation) from the LibLinear package [Fan et al., 2008].

4.3.2.2 Training with Data Imbalance

Hoax debunking websites, by nature, mostly contain articles that *refute* rumors and urban legends. As a result, the training data for the stance classifier is imbalanced towards negative training instances from the "refute" class. For example, in *snopes.com*, this data imbalance is 2.8 to 1. In order to learn a balanced classifier, we adjust the classifier's loss function by placing a large penalty[1] for misclassifying instances from the positive or "support" class which boosts certain features from that class. The overall effect is that the classifier makes fewer mistakes for positive instances, leading to a more balanced classification.

4.3.3 Credibility-driven Source Reliability

Our prior work [Popat et al., 2016] used the PageRank and AlexaRank of web sources as a proxy for their reliability (see Section 3.4.2). However, these measures only capture the authority and popularity of the web-sources, and not their reliability from the credibility perspective. For instance, the satirical news website *The Onion* has a very high PageRank score (7 out of 10). Hence, we propose a new approach for measuring the source reliability that takes the authenticity of its articles into account.

For each web-source, we determine the stance of its articles (regarding the respective claims) using the *Stance Classifier* explained above. A web-source is considered *reliable* if it contains articles that *refute false claims* and *support true claims*. Given a web-source ws_j with articles $\langle a^t_{ij} \rangle$ for claims $\langle c_i \rangle$ with corresponding credibility labels $\langle y^t_i \rangle$, we compute its reliability as:

$$reliability(ws_j) = \frac{\sum_{a^t_{ij}} \mathbf{1}\{St_{a^t_{ij}} = \text{`+'}, y^t_i = T\} + \sum_{a^t_{ij}} \mathbf{1}\{St_{a^t_{ij}} = \text{`-'}, y^t_i = F\}}{\text{cardinality}(\langle a^t_{ij} \rangle)}$$

where $\mathbf{1}\{.\}$ is an indicator function which takes the value 1 if its argument is true, and 0 otherwise; $\{St_{a^t_{ij}} = \text{`+'}\}$ and $\{St_{a^t_{ij}} = \text{`-'}\}$ indicate that the article a^t_{ij} is

[1] We set the weight parameter in the LibLinear classifier to attribute a large penalty in the loss function for the class with less number of training instances.

supporting or refuting the claim, respectively. Thus, the first term in the numerator in the above equation counts the number of articles where a source *supports a true claim*, whereas the second term counts the number of articles where it *refutes a false claim*. Later, we use this reliability score of a source to weigh the credibility score of articles from a given source.

4.4 Credibility Assessment Models

We describe our different approaches for credibility assessment in the following sections.

4.4.1 Content-aware Assessment

Since the content-aware models are agnostic of time, we drop the superscripts t for all the variables in this section for notational brevity and better readability.

4.4.1.1 Model Based on Distant Supervision

As credibility labels are available per-claim, and not per-reporting-article, our first approach extends the distant supervision based approach used in our prior work [Popat et al., 2016] (see Section 3.4.3) by incorporating stance and improved source reliabilities. We attach the (observed) label y_i of each claim c_i to each article a_{ij} reporting the claim (i.e., setting labels $y_{ij} = y_i$). Using these $\langle y_{ij} \rangle$ as the corresponding training labels for $\langle a_{ij} \rangle$ with the corresponding feature vectors $\langle F^L(a_{ij}) \cup F^{St}(a_{ij}) \rangle$, we train an L_1-regularized logistic regression model on the training data along with the guard against data imbalance (see Section 4.3.2.2).

For any *test* claim c_i whose credibility label is unknown, and its corresponding reporting articles $\langle a_{ij} \rangle$, we use this *Credibility Classifier* to obtain the corresponding credibility labels $\langle y_{ij} \rangle$ of the articles. We determine the overall credibility label y_i of c_i by considering a weighted contribution of its *per-article* credibility probabilities, using the corresponding source reliability values as weights.

$$y_i = \underset{l \in \{T,F\}}{\arg\max} \sum_{a_{ij}} \big[reliability(ws_j) * Pr(y_{ij} = l) \big]$$

4.4.1.2 Joint Model Based on CRF

The model described in the previous section learns the parameters for article stance, source reliability and claim credibility separately. A potentially more powerful approach is to capture the mutual interaction among these aspects in a probabilistic graphical model with joint inference, specifically a Conditional Random Field (CRF).

Consider all the web-sources $\langle WS \rangle$, evidence articles $\langle A \rangle$, claims $\langle C \rangle$ and claim credibility labels $\langle Y \rangle$ to be nodes in a graph (see Figure 4.2). Let $\langle A_i \rangle$ be the set of all articles related to claim c_i. Each claim $c_i \in C$ is associated with a binary random variable $y_i \in Y$, where $y_i \in \{0, 1\}$ indicates whether the claim is *false* or *true*, respectively. We denote the *reliability* of web-source ws_j with α_j.

The CRF operates on the cliques of this graph. A clique, in our setting, is formed amongst a claim $c_i \in C$, a source $ws_j \in WS$ and an article $a_{ij} \in A$ about c_i found in ws_j. Different cliques are connected via common sources and claims. There are as many cliques in the graph as the number of reporting articles. Let $\phi_{a_{ij}}(y_i, c_i, ws_j, a_{ij})$ be a potential function for the clique corresponding to a_{ij}. Each clique has a set of associated feature functions $F^{a_{ij}}$ with a weight vector θ. We denote the individual features and their weights as $f_k^{a_{ij}}$ and θ_k. The features are constituted by the stylistic, stance, and reliability features (see Sections 4.3.1, 4.3.2 & 4.3.3): $F^{a_{ij}} = \{\alpha_j\} \cup F^L(a_{ij}) \cup F^{St}(a_{ij})$.

We estimate the conditional distribution:

$$Pr(y_i | c_i, \langle ws_j \rangle, \langle a_{ij} \rangle; \theta) \propto \prod_{a_{ij}=1}^{|A_i|} \phi_{a_{ij}}(y_i, c_i, ws_j, a_{ij}; \theta)$$

The contribution of the potential of every clique $\phi_{a_{ij}}$ towards a claim c_i is weighed by the reliability of the source that takes its stance into account. Consider $\psi_{a_{ij}}(ws_j; \alpha_j, \theta_0)$ to be the potential for this reliability-stance factor. Therefore,

$$Pr(y_i | c_i, \langle ws_j \rangle, \langle a_{ij} \rangle; \theta) = \frac{1}{Z_i} \prod_{a_{ij}=1}^{|A_i|} \Big[\psi_{a_{ij}}(ws_j; \alpha_j, \theta_0) \times \phi_{a_{ij}}(y_i, c_i, ws_j, a_{ij}; \theta) \Big]$$

where, $Z_i = \sum_{y_i \in \{0,1\}} \prod_{a_{ij}=1}^{|A_i|} \Big[\psi_{a_{ij}}(ws_j; \alpha_j, \theta_0) \times \phi_{a_{ij}}(y_i, c_i, ws_j, a_{ij}; \theta) \Big]$ is the normalization factor. Assuming each factor takes the exponential family form, with features and weights made explicit:

$$\begin{aligned} Pr(y_i | c_i, \langle ws_j \rangle, \langle a_{ij} \rangle; \theta) &= \frac{1}{Z_i} \prod_{a_{ij}=1}^{|A_i|} \Big[\exp(\theta_0 \times \alpha_j) \times \exp(\sum_{k=1}^{K} \theta_k \times f_k^{a_{ij}}(y_i, c_i, ws_j, a_{ij})) \Big] \\ &= \frac{1}{Z_i} \exp(\theta_0 \times \sum_{a_{ij}=1}^{|A_i|} \alpha_j + \sum_{a_{ij}=1}^{|A_i|} \sum_{k=1}^{K} \theta_k \times f_k^{a_{ij}}(y_i, c_i, ws_j, a_{ij})) \\ &= \frac{1}{Z_i} \exp(\theta^T \cdot F^i) \end{aligned}$$

where, $F^i = \Big[\sum_{a_{ij}=1}^{|A_i|} \alpha_j \; \sum_{a_{ij}=1}^{|A_i|} f_1^{a_{ij}} \; \sum_{a_{ij}=1}^{|A_i|} f_2^{a_{ij}} \cdots \sum_{a_{ij}=1}^{|A_i|} f_K^{a_{ij}} \Big]$ and $\theta = [\theta_0 \; \theta_1 \; \theta_2 \; \cdots \theta_K]$.

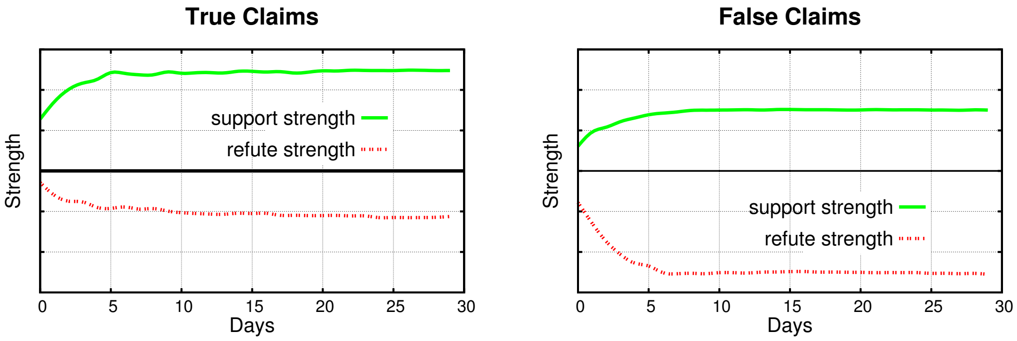


Figure 4.3: Trend of stance for *True* and *False* Claims.

We maximize the conditional log-likelihood of the data:

$$LL(\theta) = \sum_{i=1}^{|C|} \left[\theta^T \cdot F^i \ - \ \log \sum_{y_i} \exp(\theta^T \cdot F^i) \right] - \sigma ||\theta||_1$$

The L_1 regularization on the feature weights enforces the model to learn sparse features. The optimization for $\theta^* = \text{argmax}_\theta LL(\theta)$ is the same as that of logistic regression, with the *transformed* feature space. We use code from LibLinear [Fan et al., 2008] for optimization that implements trust region Newton method for large-scale logistic regression, with guard against data imbalance (see Section 4.3.2.2).

4.4.2 Trend-aware Assessment

Our hypothesis for this model is that the trend of evidence articles supporting *true* claims increases much faster than the trend of refuting them over time; whereas, for *false* claims, there is a trend of refuting them over time, rather than supporting them. To validate our hypothesis, we plot the cumulative number of supporting and refuting articles for each claim — aggregated over all the claims in our dataset — till each day $t \in [1 - 30]$ after the origin of a claim. As we can see from Figure 4.3, the cumulative support strength increases faster than the refute strength for true claims and vice versa for false claims. We want to exploit this insight of evolving trends for credibility assessment of newly emerging claims. Thus, we revise our credibility assessment each day with new incoming evidence (i.e., articles discussing the claim) based on the trend of support and refute.

In this approach, the credibility $Cr_{\text{trend}}(c_i, t)$ of a claim c_i at each day t is influenced by two components: (i) the *strength* of support and refute till time t (denoted by $q^+_{i,t}$ and $q^-_{i,t}$, respectively), and (ii) the *slope* of the trendline of support and refute (denoted by $r^+_{i,t}$ and $r^-_{i,t}$, respectively) till time t for the claim. Let $\langle A^+_{i,t} \rangle$ and $\langle A^-_{i,t} \rangle$ denote the *cumulative* number of supporting and refuting articles for claim

c_i till day t. The cumulative support and refute strength for the claim c_i till each day t is given by the mean of the stance scores, i.e., support and refute, denoted by p^+ and p^- (see Section 4.3.2), respectively — of all the articles reporting on the claim till that day, weighed by the reliability of their sources:

$$q^+_{i,t} = \frac{\sum_{a^t_{ij} \in A^+_{i,t}} p^+(a^t_{ij}) \times reliability(ws_j)}{|A^+_{i,t}|}$$

$$q^-_{i,t} = \frac{\sum_{a^t_{ij} \in A^-_{i,t}} p^-(a^t_{ij}) \times reliability(ws_j)}{|A^-_{i,t}|}$$

The slope of the trendline for the support and refute strength for the claim c_i till each day t is given by:

$$r^+_{i,t} = \frac{t \cdot \sum_{t'=1}^{t} (q^+_{i,t'} \cdot t') - \sum_{t'=1}^{t} q^+_{i,t'} \cdot \sum_{t'=1}^{t} t'}{t \cdot \sum_{t'=1}^{t} t'^2 - (\sum_{t'=1}^{t} t')^2}$$

$$r^-_{i,t} = \frac{t \cdot \sum_{t'=1}^{t} (q^-_{i,t'} \cdot t') - \sum_{t'=1}^{t} q^-_{i,t'} \cdot \sum_{t'=1}^{t} t'}{t \cdot \sum_{t'=1}^{t} t'^2 - (\sum_{t'=1}^{t} t')^2}$$

The trend-based credibility score of claim c_i at time t aggregates the strength and slope of the trendline for support and refute as:

$$Cr_{\text{trend}}(c_i, t) = [q^+_{i,t} \cdot (1 + r^+_{i,t})] - [q^-_{i,t} \cdot (1 + r^-_{i,t})]$$

4.4.3 Content and Trend-aware Assessments

The content-aware approach analyzes the *language* of reporting articles from various sources. Whereas, the trend-aware approach captures the *temporal footprint* of the claim on the web for credibility assessment taking into account the trend of how various web-sources support or refute a claim over time. Hence, to take advantage of both the approaches, we combine their assessments for any claim c_i at time t as follows:

$$Cr_{comb}(c_i, t) = \alpha \cdot Cr_{\text{content}}(c_i, t) + (1 - \alpha) \cdot Cr_{\text{trend}}(c_i, t) \tag{4.1}$$

where, $Cr_{\text{content}}(c_i, t) = [Pr(y_i = \text{true})]$ (see Section 4.4.1) and $Cr_{\text{trend}}(c_i, t)$ are the credibility scores provided by the content-aware approach and trend-aware approach, respectively. $\alpha \in [0 - 1]$ denotes the combination weight.

Total Claims	4856
True claims	1277 (26.3%)
False claims	3579 (73.7%)
Web articles	133272
Relevant articles	80421
Relevant web-sources	23260

Table 4.2: *Snopes* data statistics.

4.5 Experiments

4.5.1 Datasets

For assessing the performance of our approaches, we performed case studies on two real-world datasets: (i) Snopes (*snopes.com*) and (ii) Wikipedia (*wikipedia.com*), which are made available online[2].

Snopes

Snopes is a well-known fact-checking website that validates Internet rumors, e-mail forwards, hoaxes, urban legends, and other stories of unknown or questionable origin receiving around 300,000 visits a day[3]. They typically collect these rumors and claims from *social media*, news websites, e-mails by users, etc. Each website article verifies a single claim, e.g., *"Clown masks have been banned in the United States, and wearing one can result in a $50,000 fine."*. The credibility of such claims are *manually* analyzed by Snopes' editors and labeled as *True* or *False*. For more details about the dataset, please refer to Popat et al. [2016] (see Section 3.5.1).

We collected these fact-checking articles from *Snopes* that are published until February 2016. For each claim c_i, we fired the *claim text* as a *query* to the Google search engine[4] and extracted the first three result pages (i.e., 30 articles) as a set of evidence articles $\langle a_{ij} \rangle$. We then crawled all these articles (using jsoup[5]) from their corresponding web-sources $\langle ws_j \rangle$. We removed search results from the *snopes.com* domain to avoid any kind of bias.

Statistics of the data crawled from *snopes.com* is given in Table 4.2. "Relevant" articles denote articles containing *at least* one snippet maintaining a stance (support or refute) about the target claim, as determined by our *Stance Classifier*. Similarly,

[2] http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/impact/web-credibility-analysis/ (accessed 8 July, 2019)

[3] http://www.nytimes.com/2010/07/15/technology/personaltech/15pogue-email.html (accessed 8 July, 2019)

[4] Our system has no dependency on Google. Other search engines or other means of evidence gathering could easily be used.

[5] https://jsoup.org/ (accessed 8 July, 2019)

	Hoaxes	**Fictitious People**
Total Claims	100	57
Web articles	2813	1552
Relevant articles	2092	1136
Relevant web-sources	1250	705

Table 4.3: Wikipedia data statistics.

relevant web-sources denote sources with at least one relevant article for any of the claims in our dataset.

Wikipedia

Wikipedia contains a list of proven hoaxes[6] and fictitious people[7] (like fictional characters from novels). We used the same dataset as our prior work [Popat et al., 2016] (see Section 3.5.2) of 100 hoaxes and 57 fictitious people. The ground-truth label for all of these claims is *False*. The statistics of the dataset is reported in Table 4.3. As described earlier, we used a search engine to get a set of reporting articles for these claims by firing queries like "*<ENTITY> exists*" and "*<ENTITY> is genuine*". Similar to the previous case, we removed results from the *wikipedia.org* domain.

Time-series Dataset

As new claims emerge on the web, they are gradually picked up for reporting by various web-sources. To assess the performance of our trend-aware and combined approach for *emerging* claims, we require time-series data which mimics the behavior of emerging evidence (i.e., reporting articles) for newly emerged claims. Most of the prior works on rumor propagation dealt with online social networks (e.g., Twitter) [Kwon et al., 2013; Zubiaga et al., 2016] where it is easy to trace the information diffusion. It is quite difficult to get such time-series data for the open web. In absence of any readily available dataset, we use a search engine to crawl the results.

Many of the Snopes articles contain the origin date of the claims. We were able to obtain 439 claims (54 *True* and 385 *False*) along with their date of origin on the web from Snopes. Now, to mimic the time-series behavior, we hit the Google search engine (using date restriction feature) and retrieved relevant reporting articles on a claim (first page of search results) on each day, starting from its day of origin to the next 30 days. We obtained 6000 *relevant* articles overall — as determined by

[6] https://en.wikipedia.org/wiki/List_of_hoaxes#Proven_hoaxes (accessed 8 July, 2019)

[7] https://en.wikipedia.org/wiki/List_of_fictitious_people (accessed 8 July, 2019)

Refute Class	**Support Class**
rumor, hoax, fake, false, satirical, fake news, spoof, fiction, circulate, not true, fictitious, not real, fabricate, reveal, can not, humor, mis- information, mock, unclear ...	review, editorial, accurate, speech, honor, display, marital, history, coverage, coverage story, read, now live, story, say, additional information, anticipate, examine ...

Table 4.4: Top contributing features for determining stance.

Reliable	**Non Reliable**
wikipedia.org, *thatsfake.com*, *ibtimes.co.in*, *huffingtonpost.com*, *nydailynews.com*, *cnn.com*, *aljazeera.com ...*	*americannews.com*, *theonion.com*, *fox6now.com*, *huzlers.com*, *weeklyworldnews.com*, *dailycurrant.com ...*

Table 4.5: Top-ranked reliable and non-reliable sources.

our Stance Classifier. Using this time series dataset, the system's goal is to assess the credibility of a claim as soon as possible from its date of origin, given the set of reporting articles available in those initial days.

4.5.2 Stance and Source Reliability Assessment

To determine the stance of an article towards the claim, we trained our *Stance Classifier* (Section 4.3.2) using the *Snopes* data. The articles confirming (i.e., supporting) claims were taken as positive instances, whereas those debunking (i.e., refuting) claims were considered as negative instances. This trained model was used for determining the stance in both *Snopes* and *Wikipedia* datasets. We obtained **76.69%** accuracy with 10-fold cross-validation on labeled *Snopes* data for stance classification. Top contributing features for both classes are shown in Table 4.4.

As described in Section 4.3.3, we used the outcome of the stance determination algorithm to learn the reliability of various web-sources. The most reliable and most unreliable sources, as determined by our method, are given in Table 4.5.

4.5.3 Content-aware Assessment on Snopes

We perform 10-fold cross-validation on the claims by using 9-folds of the data for training, and the remaining fold for testing. The algorithm learned the *Credibility Classifier* and web-source reliabilities from the reporting articles and their corresponding sources present only in the training split. In case of a new web-source in test data, not encountered in the training data, its reliability score was set to

0.5 (i.e., equally probable of being reliable or not). We ignored all *Snopes*-specific references from the data and the search engine results in order to remove any training bias. For addressing the data imbalance issue (see Section 4.3.2.2), we set the penalty for the true class to 2.8 — given by the ratio of the number of *false* claims to *true* claims in the *Snopes* data.

4.5.3.1 Evaluation Measures

We report the overall accuracy of the model, Area-under-Curve (AUC) values of the ROC (Receiver Operating Characteristic) curve, precision, recall and F1 scores for the *False* claim class. *Snopes*, primarily being a hoax debunking website, is biased towards reporting *False* claims — the data imbalance being 2.8 : 1. Hence, we also report the *per-class accuracy* and the *macro-averaged accuracy* which is the average of *per-class* accuracy — giving equal weight to both classes irrespective of the data imbalance.

4.5.3.2 Baselines

We compare our approach with the following baselines implemented based on their respective proposed methods:

- **ZeroR:** A trivial baseline that always labels a claim as the class with the largest proportion in the dataset, i.e., *false* in our case.
- **Fact-finder Approaches:** Approaches based on: (i) Generalized Sum [Pasternack and Roth, 2011], (ii) Average-Log [Pasternack and Roth, 2011], (iii) TruthFinder [Yin et al., 2008] and (iv) Generalized Investment [Pasternack and Roth, 2010] and (v) Pooled Investment [Pasternack and Roth, 2010]; implemented following the same method as suggested in [Samadi, 2015].
- **Truth Assessment:** Recent work on truth checking [Nakashole and Mitchell, 2014] utilizes the objectivity score of the reporting articles to find the truth. "Objectivity Detector" was constructed using the code[8] of Mukherjee and Weikum [2015]. A claim was labeled *true* if the sum of the objectivity scores of its reporting articles was higher than the sum of the subjective scores, and *false* otherwise.
- **Our Prior Work (Lang. & Auth.):** We also use our prior approach proposed in Chapter 3 [Popat et al., 2016] which considers only the language of the reporting articles, and PageRank and AlexaRank based features for source authority to assess the credibility of claims.

[8] Code: http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/impact/credibilityanalysis/ (accessed 8 July, 2019)

Configuration	**Macro-averaged Accuracy (%)**
ZeroR	50.00
Generalized Investment [Pasternack and Roth, 2010]	54.33
Truth Assessment [Nakashole and Mitchell, 2014]	56.06
TruthFinder [Yin et al., 2008]	56.91
Generalized Sum [Pasternack and Roth, 2011]	62.82
Pooled Investment [Pasternack and Roth, 2010]	63.09
Average-Log [Pasternack and Roth, 2011]	65.89
Lang. & Auth. [Popat et al., 2016]	73.10
Our Approach: CRF	**80.00**
Our Approach: Distant Supervision	**82.00**

Table 4.6: Performance comparison of our model vs. related baselines with 10-fold cross-validation on Snopes dataset.

4.5.3.3 Model Configurations

Along with the above baselines, we also report the results of our model with different feature configurations for linguistic style, stance, and credibility-driven web-source reliability:

- Models using only *language* (LG) features, only *stance* (ST) features, and their combination (LG + ST). These configurations use simple averaging of *per-article* credibility scores to determine the overall credibility of the target claim.

- The aggregation over articles is refined by considering the reliability of the web-source who published the article, considering *language and source reliability* (LG + SR), and *stance and source reliability* (ST + SR).

- Finally, all the aspects *language, stance* and *source reliability* (LG + ST + SR) are considered together.

4.5.3.4 Results

Table 4.6 shows the 10-fold cross-validation macro-averaged accuracy of our model against various baselines. As we can see from the table, our methods outperform all the baselines by a large margin. Table 4.7 shows the performance comparison of the different configurations, with the ROC curves plotted in Figure 4.4. We can observe that using only language stylistic features (LG) is not sufficient; it is important to understand the stance (ST) of the article as well. Considering stance along with the language boosts the *Macro-averaged Accuracy* by $\sim 5\%$ points.

	Configuration	Overall Accuracy (%)	*True* Claims Accuracy (%)	*False* Claims Accuracy (%)	Macro-averaged Accuracy (%)	AUC	*False* Claims Precision	*False* Claims Recall	*False* Claims F1-Score
	CRF	**84.02**	71.26	**88.74**	80.00	0.86	0.89	**0.89**	**0.89**
Distant Supervision	**LG + ST + SR**	81.39	**83.21**	80.78	**82.00**	**0.88**	**0.93**	0.81	0.87
	ST + SR	79.43	80.12	79.22	79.67	0.86	0.92	0.79	0.85
	LG + ST	71.98	77.47	70.04	73.76	0.81	0.89	0.70	0.78
	Lang. + Auth.	71.96	75.43	70.77	73.10	0.80	0.89	0.71	0.79
	LG + SR	69.78	74.55	68.13	71.34	0.77	0.88	0.68	0.77
	ST	67.15	72.77	65.17	68.97	0.76	0.87	0.65	0.74
	LG	66.65	74.12	64.02	69.07	0.75	0.87	0.64	0.74

Table 4.7: Credibility classification results on Snopes dataset with different feature configurations (LG: language stylistic, ST: stance, SR: web-source reliability).

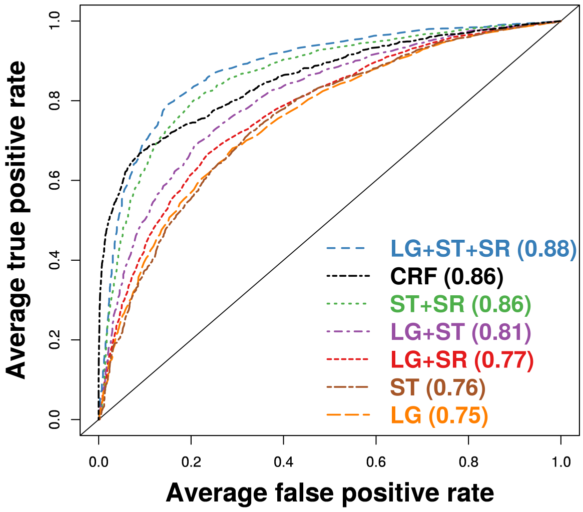


Figure 4.4: ROC curves for different model configurations for Snopes dataset.

The full model configuration, i.e., source reliability along with language style and stance features (LG + ST + SR), significantly boosts *Macro-averaged Accuracy* by ∼ 10% points. High precision, recall and F1 scores for the *False* claim class show the strength of our model in detecting *False* claims. It also outperforms our prior work by a big margin which highlights the contribution of the *stance* and credibility-driven *source reliability* features.

We can observe from Table 4.7 that even though the overall accuracy of our CRF method is highest, it has comparatively a low performance on the true-claims class. Unlike the approach using Distant Supervision, the objective function in CRF is geared towards maximizing the *overall* accuracy, and therefore biased towards the false claims due to data imbalance. This persists even after adjusting the loss function during training to favor the positive class.

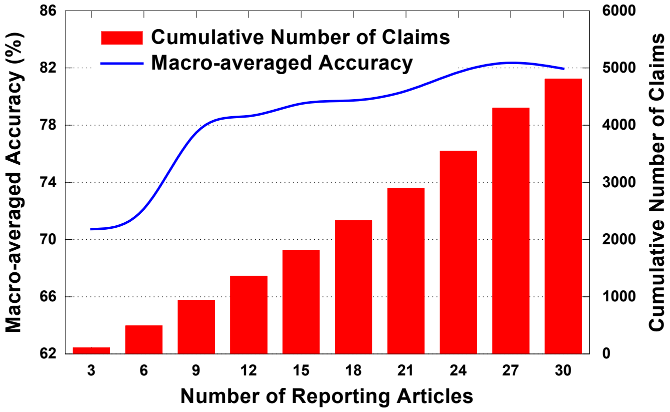


Figure 4.5: Performance on "long-tail" claims from Snopes dataset.

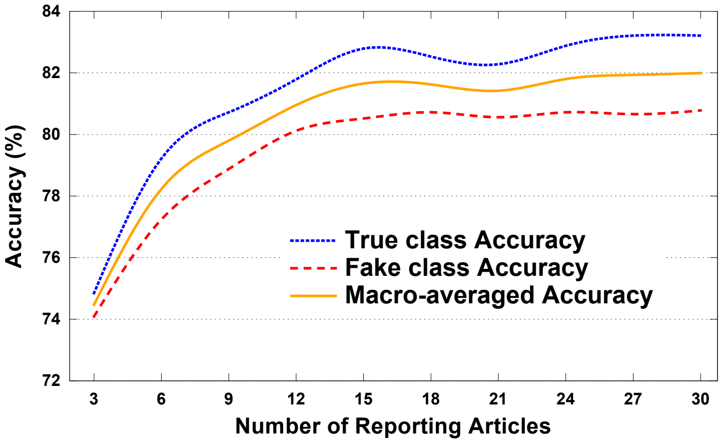


Figure 4.6: Performance by varying number of reporting articles per claim.

4.5.4 Handling "Long-tail" Claims

In this experiment, we test the performance of our content-aware approach on "long-tail" claims from the *Snopes* dataset that have only few reporting articles. We dissected the overall 10-fold cross-validation performance of our model based on the number of reporting articles of the claims. While calculating the performance, we considered *only* those claims which have $\leq k$ reporting articles, where $k \in \{3, 6, 9, \cdots 30\}$. Figure 4.5 shows the change in the *Macro-averaged Accuracy* for claims having different number of reporting articles. The Y-axis on the right hand side depicts the cumulative number of selected claims. The right-most bar in Figure 4.5 shows the performance of the LG + ST + SR configuration reported in Table 4.7. From the graph, we observe that our content-aware approach performs well even for "long-tail" claims having as few as 3 or 6 reporting articles.

	Social Media	**Web**
Total claims	1566	1566
True claims	416	416
Fake claims	1150	1150
Relevant Web articles	6615	32668

Table 4.8: Data statistics: *Social Media* as a source of evidence.

Configuration	**Overall Accuracy (%)**	***True* Claims Accuracy (%)**	***False* Claims Accuracy (%)**	**Macro-averaged Accuracy (%)**
Social Media	76.12	77.34	75.66	76.50
Web	84.23	86.01	83.56	84.78

Table 4.9: Performance of credibility classification with different sources of evidence.

4.5.5 Varying Number of Reporting Articles

For the *Snopes* dataset, we also studied how the overall model performance changes with the number of reporting articles being considered. Here, we considered only the first k reporting articles per claim to train the model, and performed 10-fold cross-validation, where $k \in \{3, 6, 9, \cdots 30\}$. Figure 4.6 shows the change in the *per-class Accuracy* and *Macro-averaged Accuracy* with varying number of reporting articles. As the number of reporting articles increases, the performance of the model also increases in a linear fashion — getting stabilized after about 15 reporting articles.

4.5.6 Social Media as a Source of Evidence

Generally, social media is considered to be very noisy [Baldwin et al., 2013]. To test the reliability of social media in providing credibility verdicts for claims, we performed an additional experiment. We considered the following social media sites as potential sources of evidence: *Facebook*, *Twitter*, *Quora*, *Reddit*, *Wordpress*, *Blogspot*, *Tumblr*, *Pinterest*, *Wikia*. We selected the set of claims from the *Snopes* dataset (statistics are reported in Table 4.8) that had at least 3 reporting articles from the above-mentioned sources. In the first configuration – *Social Media* – we used reporting articles only from these sources for credibility classification. In the second configuration – *Web* – we considered reporting articles from all sources on the web, including the social media sources. 10-fold cross-validation results for this task are reported in Table 4.9.

Test Data	#Claims	Lang.+Auth. Accuracy (%)	LG+ST+SR Accuracy (%)
WikiHoaxes	100	84.00	88.00
WikiFictitious People	57	66.07	82.14

Table 4.10: Accuracy of credibility classification on *Wikipedia* dataset (LG: language stylistic, ST: stance, SR: web-source reliability).

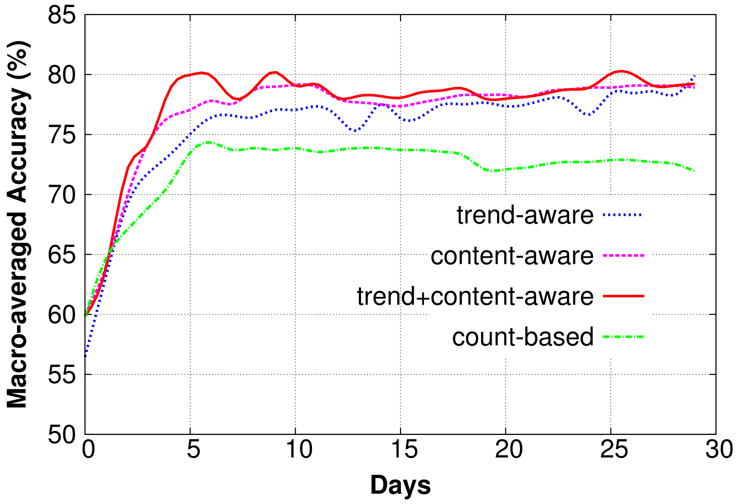


Figure 4.7: Comparison of macro-averaged accuracy for assessing the credibility of newly emerging claims.

As we can observe from the results, relying *only* on social media results in a big drop in accuracy. Our system still performs decently. However, the system performance is greatly improved (∼ 8% points) by adding other sources of evidence from the web.

4.5.7 Content-aware Assessment on Wikipedia

To evaluate the generality of our content-aware approach, we train our model on the Snopes dataset and test it on the *Wikipedia* dataset of hoaxes and fictitious people. The results in Table 4.10 demonstrate significant performance improvements over our prior work, *Lang.+Auth.* [Popat et al., 2016] (refer to Chapter 3), and the effectiveness of the *stance* and credibility-driven *source reliability* features in our model. Similar to the *Snopes* setting, we removed all references to Wikipedia from the data and search engine results. As we can see from the results, our system is able to detect hoaxes and fictitious people with high accuracy, although the claim descriptions here are stylistically quite different from those of *Snopes*.

4.5.8 Credibility Assessment of Emerging Claims

The goal of this experiment is to evaluate the performance of our approach with respect to the early assessment of newly emerging claims having a sparse presence on the web. Using the time-series dataset (see Section 4.5.1), we assess the credibility of the emerging claims on each day t starting from their date of origin by considering the evidence (i.e., reporting articles) *only till day* t. We compare the macro-accuracy of the following approaches on each day t:

- *count-based approach:* In this approach, on each day t, we compare the cumulative number of supporting and refuting articles for a claim *till that day*. The stance is obtained using Algorithm 1 in Section 4.3.2. If the number of supporting articles is higher than the number of refuting ones, the claim is labeled *true*, and *false* otherwise.

- *trend-aware approach:* As described in Section 4.4.2, this analyzes the trend till day t to assess the credibility.

- *content-aware approach:* As described in Section 4.4.1, our model analyzes the content of relevant articles till day t and predicts the credibility of the claim.

- *content & trend-aware approach:* This combined approach considers credibility scores from both the models: content-aware and trend-aware (see Section 4.4.3). We varied the combination weight $\alpha \in [0-1]$ in steps of 0.1 on withheld development set, and found $\alpha = 0.4$ to give the optimal performance.

Results: Figure 4.7 shows the comparison of our approach with the baselines. As we observe in the figure, the count-based (baseline) approach performs the worst — thereby, ascertaining that simply counting the number of supporting / refuting articles is not enough. The best performance is achieved by the combined *content & trend-aware* approach. During the early days after a new claim has emerged, it leverages the trend to achieve the best performance. The results also highlight that we achieve *early detection* of emerging claims within 4−5 days of its day of origin on the web with high macro-averaged accuracy (ca. 80%). At the end of a month, after the claim has emerged, all the approaches (except count-based) converge to similar results. The improvements in macro-accuracy for all of the respective approaches are statistically significant with p-value $< 2e-16$ using paired sample t-test.

4.5.9 Evidence for Credibility Classification

Given a claim, our *Stance Classifier* extracts top-ranked snippets from the reporting articles along with their stance (*support* or *refute* probabilities). Combined with

Claim	Verdict & Evidence
Titanium rings can be removed from swollen fingers only through amputation.	**[Verdict]:** **False** **[Evidence]:** A rumor regarding titanium rings maintains that ... This is completely untrue. In fact, you can use a variety of removal techniques to safely and effectively remove a titanium ring.
The use of solar panels drains the sun of energy.	**[Verdict]:** **False** **[Evidence]:** Solar panels do not suck up the Sun's rays of photons. Just like wind farms do not deplete our planet of wind. These renewable sources of energy are not finite like fossil fuels. Wind turbines and solar panels are not vacuums, nor do they divert this energy from other systems.
Facebook soon plans to charge monthly subscription fees to users of the social network.	**[Verdict]:False** **[Evidence]:** The rumor that Facebook will suddenly start charging users to access the site has become one of the social media eras perennial chain letters.
Soviet Premier Nikita Khrushchev was denied permission to visit Disneyland during a state visit to the U.S. in 1959.	**[Verdict]:** **True** **[Evidence]:** Soviet Premier Nikita Khrushchev's good-will tour of the United States in September 1959. While some may have heard of Khrushchev's failed attempt to visit Disneyland, many do not realize that this was just one of a hundred things that went wrong on this trip.
Between 1988 and 2006, a man lived at a Paris airport.	**[Verdict]:** **True** **[Evidence]:** Mehran Karimi Nasseri (born 1942) is an Iranian refugee who lived in the departure lounge of Terminal One in Charles de Gaulle Airport from 26 August 1988 until July 2006, when he was hospitalized for an unspecified ailment. His autobiography has been published as a book (The Terminal Man) and was the basis for the 2004 Tom Hanks movie The Terminal.

Table 4.11: Example claims with credibility verdict and automatically generated evidence from the Stance Classifier.

the verdict (*true* or *false*) from the *Credibility Classifier*, this yields evidence for the verdict. Table 4.11 shows examples of our model's output for some claims, along with the verdict and evidence. In contrast to all previous approaches, the assessment of our model can be easily interpreted by the user.

4.6 Conclusions

In this chapter, we propose enhanced approaches leveraging the stance, reliability, and trend of sources of evidence and counter-evidence for credibility assessment of textual claims. Our experiments demonstrate that our system performs well on assessing the credibility of newly emerging claims within 4 to 5 days of its day of origin on the web with 80% accuracy; as well as for "long-tail" claims having as few as three reporting articles. Despite the fact that *social media* is very noisy, we show that our system can effectively harness evidence from such sources to validate or falsify a claim.

In contrast to prior approaches, we provide explanations for our credibility verdict in the form of informative snippets from articles published by reliable sources that can be easily interpreted by the users. Experiments with data from the real-world fact-checking website *snopes.com* and reported cases of hoaxes and fictitious persons in Wikipedia demonstrate the superiority of our approaches over prior works.

5

Neural-Network-Based Credibility Assessment

Contents

5.1 Introduction

Prior approaches for automated fact-checking either assume claims to be in a structured form or do not consider external evidence apart from labeled training instances. Our models proposed in Chapter 3 & Chapter 4 counter this deficit by considering external evidence articles related to a claim. However, these methods require substantial feature modeling and a rich set of lexicons.

In this chapter, we overcome the limitations of prior work (including our own) with an end-to-end model for evidence-aware credibility assessment of arbitrary textual claims, without any human intervention. It presents a neural network model, that judiciously aggregates signals from external evidence articles, the language of these articles and the trustworthiness of their sources. It also derives informative features for generating user-comprehensible explanations that make the neural network predictions transparent to the end-user. Experiments with four datasets and ablation studies show the strength of our method.

State of the Art and Limitations: Prior work on "truth discovery" (see Li et al. [2015b] for a survey)[1] largely focused on structured facts, typically in the form of subject-predicate-object triples, or on social media platforms like Twitter, Sina Weibo, etc. Recently, methods have been proposed to assess the credibility of claims in natural language form [Popat et al., 2017; Rashkin et al., 2017; Wang, 2017], such as news headlines, quotes from speeches, blog posts, etc.

The methods geared for general text input address the problem in different ways. On the one hand, methods like Rashkin et al. [2017]; Wang [2017] train neural networks on labeled claims from sites like *PolitiFact.com*, providing credibility assessments without any explicit feature modeling. However, they use only the text of questionable claims and no external evidence or interactions that provide limited context for credibility analysis. These approaches also do not offer any explanation of their verdicts. On the other hand, our prior approach in Popat et al. [2017] (refer to Chapter 4) considers external evidence in the form of other articles (retrieved from the Web) that confirm or refute a claim, and jointly assesses the language style (using subjectivity lexicons), the trustworthiness of the sources, and the credibility of the claim. This is achieved via a pipeline of supervised classifiers. On the upside, this method generates user-interpretable explanations by pointing to informative snippets of evidence articles. On the downside, it requires substantial feature modeling and rich lexicons to detect bias and subjectivity in the language style.

Approach and Contribution: To overcome the limitations of the prior works, we present *DeClarE* [2], an end-to-end neural network model for assessing and explaining the credibility of arbitrary claims in natural-language text form. Our approach combines the best of both families of prior methods. Similar to Popat et al. [2017] (refer to Chapter 4), DeClarE incorporates external evidence or counter-evidence from the Web as well as signals from the language style and the trustworthiness of the underlying sources. However, our method does not require any feature engineering, lexicons, or other manual intervention. Rashkin et al. [2017]; Wang [2017] also develop an end-to-end model, but DeClarE goes far beyond in terms

[1] As fully objective and unarguable truth is often elusive or ill-defined, we use the term *credibility* rather than "truth".

[2] Debunking Claims with Interpretable Evidence

of considering external evidence and joint interactions between several factors, and also in its ability to generate user-interpretable explanations in addition to highly accurate assessments. For example, given the natural-language input claim *"the gun epidemic is the leading cause of death of young African-American men, more than the next nine causes put together"* by Hillary Clinton, DeClarE draws on evidence from the Web to arrive at its verdict *credible*, and returns annotated snippets like the one in Table 5.6 as explanation. These snippets, which contain evidence in the form of statistics and assertions, are automatically extracted from web articles from sources of varying credibility.

Given an input claim, DeClarE searches for web articles related to the claim. It considers the *context* of the claim via word embeddings and the (language of) web articles captured via a bidirectional LSTM (biLSTM), while using an *attention* mechanism to focus on parts of the articles according to their relevance to the claim. DeClarE then aggregates all the information about claim source, web article contexts, attention weights, and trustworthiness of the underlying sources to assess the claim. It also derives informative features for interpretability, like source embeddings that capture trustworthiness and salient words captured via attention.

Key contributions of this chapter are:

- **Model:** An end-to-end neural network model which automatically assesses the credibility of natural-language claims, without any hand-crafted features or lexicons.
- **Interpretability:** An *attention* mechanism in our model that generates user-comprehensible explanations, making credibility verdicts transparent and interpretable.
- **Experiments:** Extensive experiments on four datasets and ablation studies, demonstrating the effectiveness of our method over state-of-the-art baselines.

5.2 End-to-end Framework for Credibility Assessment

Consider a set of N claims $\langle C_n \rangle$ from the respective origins/sources $\langle CS_n \rangle$, where $n \in [1, N]$. Each claim C_n is reported by a set of M articles $\langle A_{m,n} \rangle$ along with their respective sources $\langle AS_{m,n} \rangle$, where $m \in [1, M]$. Each corresponding tuple of the claim and its origin, reporting articles and article sources – $\langle C_n, CS_n, A_{m,n}, AS_{m,n} \rangle$ forms a training instance in our setting, along with the credibility label of the claim used as ground-truth during network training. Figure 5.1 gives a pictorial overview of our model. In the following sections, we provide a detailed description of our approach.

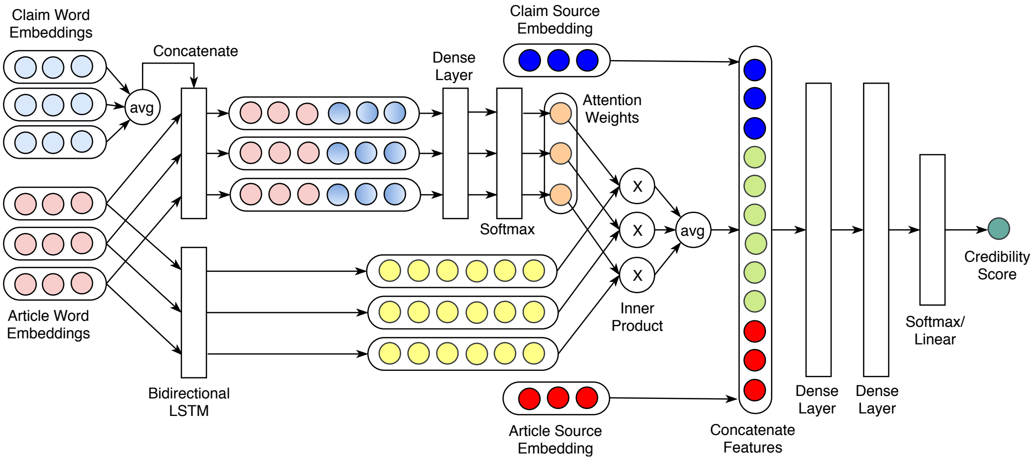


Figure 5.1: Framework for credibility assessment. Upper part of the pipeline combines the article and claim embeddings to get the claim specific attention weights. Lower part of the pipeline captures the article representation through biLSTM. Attention focused article representation along with the source embeddings are passed through dense layers to predict the credibility score of the claim.

5.2.1 Input Representations

The input claim C_n of length l is represented as $[c_1, c_2, ..., c_l]$ where $c_l \in \Re^d$ is the d-dimensional word embedding of the l-th word in the input claim. The source/origin of the claim CS_n is represented by a d_s-dimensional embedding vector $cs_n \in \Re^{d_s}$.

A reporting article $A_{m,n}$ consisting of k tokens is represented by $[a_{m,n,1}, a_{m,n,2}, ..., a_{m,n,k}]$, where $a_{m,n,k} \in \Re^d$ is the d-dimensional word embedding vector for the k-th word in the reporting article $A_{m,n}$. The claim and article word embeddings have shared parameters. The source of the reporting article $AS_{m,n}$ is represented as a d_s-dimensional vector, $as_{m,n} \in \Re^{d_s}$. For the sake of brevity, we drop the notation subscripts n and m in the following sections by considering only a single training instance – the input claim C_n from source CS_n, the corresponding article $A_{m,n}$ and its sources $AS_{m,n}$ given by: $\langle C, CS, A, AS \rangle$.

5.2.2 Article Representation

To create a representation of an article, which may capture task-specific features such as whether it contains objective language, we use a bidirectional Long Short-Term Memory (LSTM) network as proposed by Graves et al. [2005]. A basic LSTM cell consists of various gates to control the flow of information through timesteps in a sequence, making LSTMs suitable for capturing long and short-range dependencies in the text that may be difficult to capture with standard recurrent neural networks (RNNs). Given an input word embedding of tokens $\langle a_k \rangle$, an LSTM cell performs various non-linear transformations to generate a hidden vector state h_k for each token at each timestep k.

We use bidirectional LSTMs in place of standard LSTMs. Bidirectional LSTMs capture both the previous timesteps (past features) and the future timesteps (future features) via forward and backward states respectively. Correspondingly, there are two hidden states that capture past and future information that are concatenated to form the final output as: $h_k = [\overrightarrow{h_k}, \overleftarrow{h_k}]$.

5.2.3 Claim Specific Attention

As we previously discussed, it is important to consider the relevance of an article with respect to the claim; specifically, focusing or *attending* to parts of the article that discuss the claim. This is in contrast to prior works [Popat et al., 2017; Rashkin et al., 2017; Wang, 2017] that ignore either the article or the claim, and therefore miss out on this important interaction.

We propose an attention mechanism to help our model focus on salient words in the article with respect to the claim. To this end, we compute the importance of each term in an article with respect to an overall representation of the corresponding claim. Additionally, incorporating attention helps in making our model transparent and interpretable, because it provides a way to generate the most salient words in an article as evidence of our model's verdict.

Following Wieting et al. [2015], the overall representation of an input claim is generated by taking an average of the word embeddings of all the words therein:

$$\bar{c} = \frac{1}{l} \sum_{l} c_l$$

We combine this overall representation of the claim with each article term:

$$\hat{a}_k = a_k \oplus \bar{c}$$

where $\hat{a}_k \in \Re^{d+d}$ and $\oplus$ denotes the concatenate operation. We then perform a transformation to obtain claim-specific representations of each article term:

$$a'_k = \mathbf{f}(W_a \hat{a}_k + b_a)$$

where W_a and b_a are the corresponding weight matrix and bias terms, and $\mathbf{f}$ is an activation function[3], such as *ReLU*, *tanh*, or the identity function. Following this, we use a softmax activation to calculate an attention score α_k for each word in the article capturing its relevance to the claim context:

$$\alpha_k = \frac{\exp(a'_k)}{\sum_k \exp(a'_k)} \tag{5.1}$$

[3] In our model, the *tanh* activation function gives best results.

5.2.4 Per-Article Credibility Score of Claim

Now that we have article term representations given by $\langle h_k \rangle$ and their relevance to the claim given by $\langle \alpha_k \rangle$, we need to combine them to predict the claim's credibility. In order to create an attention-focused representation of the article considering both the claim and the article's language, we calculate a weighted average of the hidden state representations for all article tokens based on their corresponding attention scores:

$$g = \frac{1}{k} \sum_k \alpha_k \cdot h_k \tag{5.2}$$

We then combine all the different feature representations: the claim source embedding (cs), the attention-focused article representation (g), and the article source embedding (as). In order to merge the different representations and capture their joint interactions, we process them with two fully connected layers with non-linear activations.

$$\begin{aligned} d_1 &= relu(W_c(g \oplus cs \oplus as) + b_c) \\ d_2 &= relu(W_d d_1 + b_d) \end{aligned}$$

where W and b are the corresponding weight matrix and bias terms.

Finally, to generate the overall credibility label of the article for classification tasks, or credibility score for regression tasks, we process the final representation with a final fully connected layer:

$$\text{Classification:} \quad s = sigmoid(d_2) \tag{5.3}$$

$$\text{Regression:} \quad s = linear(d_2) \tag{5.4}$$

5.2.5 Credibility Aggregation

The credibility score in the above step is obtained considering a single reporting article. As previously discussed, we have M reporting articles per claim. Therefore, once we have the per-article credibility scores from our model, we take an average of these scores to generate the overall credibility score for the claim:

$$cred(C) = \frac{1}{M} \sum_m s_m \tag{5.5}$$

This aggregation is done after the model is trained.

5.3 Experiments

5.3.1 Datasets

We evaluate our approach and demonstrate its generality by performing experiments on four different datasets: a general fact-checking website, a political fact-checking website, a news review community, and a SemEval Twitter rumor dataset.

Snopes

Snopes (www.snopes.com) is a general fact-checking website where editors manually investigate various kinds of rumors reported on the Internet. We used the Snopes dataset provided by Popat et al. [2017] (see Section 4.5.1). This dataset consists of rumors analyzed on the Snopes website along with their credibility labels (*true* or *false*), sets of reporting articles, and their respective web sources.

PolitiFact

PolitiFact is a political fact-checking website (www.politifact.com) in which editors rate the credibility of claims made by various political figures in US politics. We extract all articles from PolitiFact published before December 2017. Each article includes a claim, the speaker (political figure) who made the claim, and the claim's credibility rating provided by the editors.

PolitiFact assigns each claim to one of six possible ratings: *true, mostly true, half true, mostly false, false* and *pants-on-fire*. Following Rashkin et al. [2017], we combine *true, mostly true* and *half true* ratings into the class label *true* and the rest as *false* – hence considering only binary credibility labels. To retrieve the reporting articles for each claim (similar to Popat et al. [2017]), we issue each claim as a query to a search engine[4] and retrieve the top 30 search results with their respective web sources.

NewsTrust

NewsTrust is a news review community in which members review the credibility of news articles. We use the NewsTrust dataset made available by Mukherjee and Weikum [2015]. This dataset contains NewsTrust stories from May 2006 to May 2014. Each story consists of a news article along with its source, and a set of reviews and ratings by community members. NewsTrust aggregates these ratings and assigns an overall credibility score (on a scale of 1 to 5) to the posted article. We map the attributes in this data to the inputs expected by DeClarE as follows: the title and

[4] We use the Bing search API.

Dataset	**SN**	**PF**	**NT**	**SE**
Total claims	4341	3568	5344	272
True claims	1164	1867	-	127
False claims	3177	1701	-	50
Unverified claims	-	-	-	95
Claim sources	-	95	161	10
Articles	29242	29556	25128	3717
Article sources	336	336	251	89

Table 5.1: Data statistics (SN: Snopes, PF: PolitiFact, NT: NewsTrust, SE: SemEval).

the web source of the posted (news) article are mapped to the input claim and claim source, respectively. Reviews and their corresponding user identities are mapped to reporting articles and article sources, respectively. We use this dataset for the regression task of predicting the credibility score of the posted article.

SemEval-2017 Task 8

As the fourth dataset, we consider the benchmark dataset released by SemEval-2017 for the task of determining credibility and stance of social media content (Twitter) Derczynski et al. [2017]. The objective of this task is to predict the credibility of a questionable tweet (*true*, *false* or *unverified*) along with a confidence score from the model. It has two sub-tasks: (i) a *closed* variant in which models only consider the questionable tweet, and (ii) an *open* variant in which models consider both the questionable tweet and additional context consisting of snapshots of relevant sources retrieved immediately before the rumor was reported, a snapshot of an associated Wikipedia article, news articles from digital news outlets, and preceding tweets about the same event. Testing and development datasets provided by organizers have 28 tweets (1021 reply tweets) and 25 tweets (256 reply tweets), respectively.

Data Processing

In order to have minimum support for training, claim sources with less than 5 claims in the dataset are grouped into a single dummy claim source, and article sources with less than 10 articles are grouped similarly (5 articles for SemEval as it is a smaller dataset).

For Snopes and PolitiFact, we need to extract relevant snippets from the reporting articles for a claim. Therefore, we extract snippets of 100 words from each reporting article having the maximum relevance score: $sim = sim_{\text{bow}} \times sim_{\text{semantic}}$, where

Parameter	SN	PF	NT	SE
Word embedding length	100	100	300	100
Claim source embedding length	-	4	8	4
Article source embedding length	8	4	8	4
LSTM size (for each pass)	64	64	64	16
Size of fully connected layers	32	32	64	8
Dropout	0.5	0.5	0.3	0.3

Table 5.2: Model parameters used for each dataset (SN: Snopes, PF: PolitiFact, NT: NewsTrust, SE: SemEval).

sim_{bow} is the fraction of claim words that are present in the snippet, and sim_{semantic} represents the cosine similarity between the average of claim word embeddings and snippet word embeddings. We also enforce a constraint that the *sim* score is at least δ. We varied δ from 0.2 to 0.8 and found 0.5 to give the optimal performance on a withheld dataset. We discard all articles related to Snopes and PolitiFact websites from our datasets to have an unbiased model. Statistics of the datasets after pre-processing is provided in Table 5.1. All the datasets are made publicly available at `https://www.mpi-inf.mpg.de/dl-cred-analysis/`.

5.3.2 Experimental Setup

When using the Snopes, PolitiFact and NewsTrust datasets, we reserve 10% of the data as validation data for parameter tuning. We report 10-fold cross validation results on the remaining 90% of the data; the model is trained on 9-folds and the remaining fold is used as test data. When using the SemEval dataset, we use the data splits provided by the task's organizers. The objective for Snopes, PolitiFact and SemEval experiments is binary (credibility) classification, while for NewsTrust the objective is to predict the credibility score of the input claim on a scale of 1 to 5 (i.e., credibility regression). We represent terms using pre-trained GloVe Wikipedia 6B word embeddings [Pennington et al., 2014]. Since our training datasets are not very large, we do not tune the word embeddings during training. The remaining model parameters are tuned on the validation data; the parameters chosen are reported in Table 5.2. We use Keras with a Tensorflow backend to implement our system. All the models are trained using Adam optimizer [Kingma and Ba, 2014] (learning rate: 0.002) with categorical cross-entropy loss for classification and mean squared error loss for regression task. We use L2-regularizers with the fully connected layers as well as dropout. For all the datasets, the model is trained using each claim-article pair as a separate training instance.

Dataset	Configuration	*True* Claims Accuracy (%)	*False* Claims Accuracy (%)	Macro F1-Score	AUC
Snopes	LSTM-text	64.65	64.21	0.66	0.70
	CNN-text	67.15	63.14	0.66	0.72
	Distant Supervision	**83.21**	**80.78**	**0.82**	**0.88**
	DeClarE (Plain)	74.37	78.57	0.78	0.83
	DeClarE (Plain+Attn)	78.34	78.91	0.79	0.85
	DeClarE (Plain+SrEmb)	77.43	79.80	0.79	0.85
	DeClarE (Full)	78.96	78.32	0.79	0.86
PolitiFact	LSTM-text	63.19	61.96	0.63	0.66
	CNN-text	63.67	63.31	0.64	0.67
	Distant Supervision	62.53	62.08	0.62	0.68
	DeClarE (Plain)	62.67	69.05	0.66	0.70
	DeClarE (Plain+Attn)	65.53	68.49	0.66	0.72
	DeClarE (Plain+SrEmb)	66.71	69.28	0.67	0.74
	DeClarE (Full)	**67.32**	**69.62**	**0.68**	**0.75**

Table 5.3: Comparison of various approaches for credibility classification on Snopes and PolitiFact datasets.

To evaluate and compare the performance of DeClarE with other state-of-the-art methods, we report the following measures:

- Credibility Classification (Snopes, PolitiFact, and SemEval): accuracy of the models in classifying *true* and *false* claims separately, macro F1-score and Area-Under-Curve (AUC) for the ROC (Receiver Operating Characteristic) curve.

- Credibility Regression (NewsTrust): Mean Square Error (MSE) between the predicted and true credibility scores.

5.3.3 Results: Snopes and Politifact

We compare our approach with the following state-of-the-art models: (i) LSTM-text, a recent approach proposed by Rashkin et al. [2017]. (ii) CNN-text: a CNN based approach proposed by Wang [2017]. (iii) Distant Supervision: state-of-the-art distant supervision based approach proposed by Popat et al. [2017] (refer to Chapter 4). (iv) DeClare (Plain): our approach with only biLSTM (no attention and source embeddings). (v) DeClarE (Plain+Attn): our approach with only biLSTM and attention (no source embeddings). (vi) DeClarE (Plain+SrEmb): our approach with only biLSTM and source embeddings (no attention). (vii) DeClarE (Full): end-to-end system with biLSTM, attention and source embeddings.

The results when performing credibility classification on the Snopes and PolitiFact datasets are shown in Table 5.3. DeClarE outperforms LSTM-text and

Configuration	MSE
CNN-text	0.53
CCRF+SVR	0.36
LSTM-text	0.35
DistantSup	0.35
DeClarE (Plain)	0.34
DeClarE (Full)	**0.29**

Table 5.4: Comparison of various approaches for credibility regression on NewsTrust dataset.

CNN-text models by a large margin on both datasets. On the other hand, for the Snopes dataset, the performance of DeClarE (Full) is slightly lower than the Distant Supervision configuration (p-value of 0.04 with a pairwise t-test). However, the advantage of DeClarE over Distant Supervision approach is that it does not rely on handcrafted features and lexicons, and can generalize well to arbitrary domains without requiring any seed vocabulary. It is also to be noted that both of these approaches use external evidence in the form of reporting articles discussing the claim, which are not available to the LSTM-text and CNN-text baselines. This demonstrates the value of external evidence for credibility assessment.

On the PolitiFact dataset, DeClarE outperforms all the baseline models by a margin of 7-9% AUC (p-value of 9.12e$-$05 with a pairwise t-test) with similar improvements in terms of Macro F1. Performance comparison of DeClarE's various configurations indicates the contribution of each component of our model, i.e., biLSTM capturing article representations, attention mechanism, and source embeddings. The additions of both the attention mechanism and source embeddings improve performance over the plain configuration in all cases when measured by Macro F1 or AUC.

5.3.4 Results: NewsTrust

When performing credibility regression on the NewsTrust dataset, we evaluate the models in terms of mean squared error (MSE; lower is better) for credibility rating prediction. We use the first three models described in Section 5.3.3 as baselines. For CNN-text and LSTM-text, we add a linear fully connected layer as the final layer of the model to support regression. Additionally, we also consider the state-of-the-art CCRF+SVR model based on Continuous Conditional Random Field (CCRF) and Support Vector Regression (SVR) proposed by Mukherjee and Weikum [2015]. The results are shown in Table 5.4. We observe that DeClarE (Full) outperforms all four baselines, with a 17% decrease in MSE compared to the best-performing baselines (i.e., LSTM-text and Distant Supervision). The DeClarE (Plain) model

Configuration	**Macro Accuracy**	**RMSE**
IITP (Open)	0.39	0.746
NileTMRG (Close)	0.54	0.673
DeClarE (Plain)	0.46	0.687
DeClarE (Full)	**0.57**	**0.604**

Table 5.5: Comparison of various approaches for credibility classification on SemEval dataset.

performs substantially worse than the full model, illustrating the value of including attention and source embeddings. CNN-text performs substantially worse than the other baselines.

5.3.5 Results: SemEval

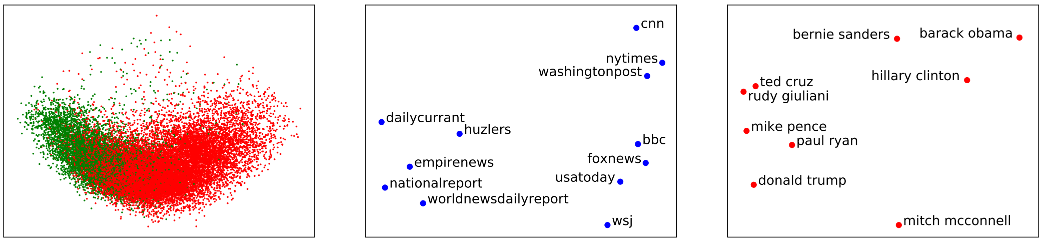
On the SemEval dataset, the objective is to perform credibility classification of a tweet while also producing a classification confidence score. We compare the following approaches and consider both variants of the SemEval task: (i) *NileTMRG* [Enayet and El-Beltagy, 2017]: the best performing approach for the *close* variant of the task, (ii) *IITP* [Singh et al., 2017]: the best performing approach for the *open* variant of the task, (iii) DeClare (Plain): our approach with only biLSTM (no attention and source embeddings), and (iv) DeClarE (Full): our end-to-end system with biLSTM, attention and source embeddings.

We use the evaluation measure proposed by the task's organizers: macro F1-score for the overall classification and Root-Mean-Square Error (RMSE) over the confidence scores. Results are shown in Table 5.5. We observe that DeClarE (Full) outperforms all the other approaches — thereby, re-affirming its power in harnessing external evidence.

5.4 Discussion

5.4.1 Analyzing Article Representations

In order to assess how our model separates articles reporting false claims from those reporting true ones, we employ dimensionality reduction using Principal Component Analysis (PCA) to project the article representations (g in Equation 5.2) from a high dimensional space to a 2d plane. The projections are shown in Figure 5.2a. We observe that DeClarE obtains clear separability between credible versus non-credible articles in Snopes dataset.



(a) Projections of article representations using PCA; DeClarE obtains clear separation between representations of non-credible articles (*red*) vs. true ones (*green*).

(b) Projections of article source representations using PCA; DeClarE clearly separates fake news sources from authentic ones.

(c) Projections of claim source representations using PCA; DeClarE clusters politicians of similar ideologies close to each other in the embedding space.

Figure 5.2: Dissecting the article, article source, and claim source representations learned by DeClarE.

5.4.2 Analyzing Source Embeddings

Similar to the treatment of article representations, we perform an analysis with the claim and article source embeddings by employing PCA and plotting the projections. We sample a few popular news sources from Snopes and claim sources from PolitiFact. These news sources and claim sources are displayed in Figure 5.2b and Figure 5.2c, respectively. From Figure 5.2b we observe that DeClarE clearly separates fake news sources like *nationalreport*, *empirenews*, *huzlers*, etc. from mainstream news sources like *nytimes*, *cnn*, *wsj*, *foxnews*, *washingtonpost*, etc. Similarly, from Figure 5.2c we observe that DeClarE locates politicians with similar ideologies and opinions close to each other in the embedding space.

5.4.3 Analyzing Attention Weights

Attention weights help understand what DeClarE focuses on during learning and how it affects its decisions – thereby, making our model transparent to the end-users. Table 5.6 illustrates some interesting claims and salient words (highlighted) that DeClarE focused on during learning. Darker shades indicate higher weights given to the corresponding words. As illustrated in the table, DeClarE gives more attention to important words in the reporting article that are relevant to the claim and also play a major role in deciding the corresponding claim's credibility. In the first example on Table 5.6, highlighted words such as "*..barely true...*" and "*..sketchy evidence...*" help our system to identify the claim as *not credible.* On the other hand, highlighted words in the last example, like, "*..reveal...*" and "*..documenting reports...*" help our system to assess the claim as *credible.*

[False] Barbara Boxer: "Fiorina's plan would mean slashing Social Security and Medicare." ***Article Source:*** *nytimes.com* *least of glimmer of truth while ignoring critical facts that would give a different impression mr adair cited a couple examples of barely true claims including this one in california democratic sen barbara boxer claimed that republican challenger carly fiorina s plan would mean slashing social security and medicare but we found there was sketchy evidence to support that fiorina hasn t said much about her ideas on social security and medicare and what she has said doesn t provide much proof of slashing and then there s this one in pennsylvania in the pennsylvania senate race republican pat toomey*
[True] Hillary Clinton: "The gun epidemic is the leading cause of death of young African-American men, more than the next nine causes put together." ***Article Source:*** *thetrace.org* *away the leading cause of death by francesca mirabile september 27 2016 during the first presidential debate monday night democratic nominee hillary clinton offered a chilling statistic on firearm homicides and the victimization of black males the gun epidemic is the leading cause of death of young african american men more than the next nine causes put together she said data from the centers for disease control and prevention confirms her assertion of all black males between the ages of 15 and 24 that died in 2014 a majority 54 percent were killed with a gun nearly nine in 10*
[False] : Coca-Cola's original diet cola drink, TaB, took its name from an acronym for "totally artificial beverage." ***Article Source:*** *foxnews.com* *the first diet colas being the first in 1952 cocacola execs at that time were hesitant to affix the term diet to cocacola so the name tab was chosen as a tribute to those who were keeping tab of their weight according to cola legend the drink was actually dubbed tab as an acronym for totally artificial beverage a great story which unfortunately cocacola says is completely untrue the name was actually chosen by computer and market research the saccharin scandal in the 70s did its damage and the introduction of diet coke in the early 1980s pushed tab even*
[True] : Household paper shredders can pose a danger to children and pets. ***Article Source:*** *byegoff.com* *packages while still protecting any private information that may be contained in the papers in theory the personal home paper shredder makes much sense personal or pet injuries from paper shredders a growing number of reported injuries reveal that home shredders pose a danger to any user and are especially dangerous to children and pets in fact the federal consumer product safety commission issued a paper shredder safety alert documenting reports of incidents involving finger amputations lacerations and other finger injuries directly connected to the use of home shredders*

Table 5.6: Interpretation via attention (weights) ([*True*]/[*False*] indicates the verdict from DeClarE).

5.5 Conclusion

In this chapter, we propose a completely automated end-to-end neural network model, DeClarE, for evidence-aware credibility assessment of natural language claims without requiring hand-crafted features or lexicons. DeClarE captures signals from external evidence articles and models joint interactions between various factors like the context of a claim, the language of reporting articles, and trustworthiness of their sources. Extensive experiments on real-world datasets demonstrate our effectiveness over state-of-the-art baselines.

6

Web Interface for Credibility Analysis

Contents

6.1 Introduction

After proposing methods for automated credibility assessment in the previous chapters, we present CredEye in this chapter. CredEye is a web interface for automatic credibility assessment which takes a natural language claim as input from the user and automatically analyzes its credibility based on the feature-based model proposed in Chapter 4. Additionally, it automatically extracts supporting evidence in the form of enriched snippets, which makes the verdicts of CredEye transparent and interpretable. Two recent systems along similar lines are ClaimBuster [Hassan et al., 2017] and ClaimVerif [Zhi et al., 2017]. However, neither of these consider the language style of the articles that serve as evidence or counter-evidence. Also, neither provides feature-level explanations of their assessment scores; rather they merely list online articles related to the claim. The unique point of CredEye is that it considers language style as a key component of its assessments, and also provides explanations in terms of automatically extracted snippets from supporting and refuting articles enriched with language features.

Given an input claim in arbitrary textual form on an arbitrary topic, CredEye automatically retrieves relevant articles from the Web, using a search engine. It

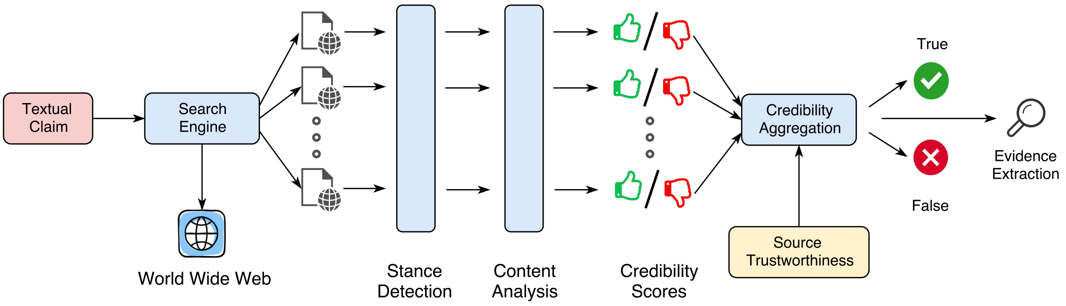


Figure 6.1: Credibility analysis pipeline of CredEye.

analyzes the credibility of each text by language features, the stance of the text, and the trustworthiness of the source, aggregating all these into an overall verdict. The UI of CredEye (see Figure 6.2) enables users to dissect and drill down into the assessment by browsing through judiciously and automatically selected snippets with the markup of indicative words. The latter capture linguistic features that express bias and subjectivity (decreasing credibility) or neutral and objective language (increasing credibility). Details of the analysis are shown in the form of per-article and per-source scores. CredEye is available at https://gate.d5.mpi-inf.mpg.de/credeye/.

6.2 Credibility Analysis Pipeline

CredEye takes a *natural language claim* as input from the user and computes its credibility assessment along with enriched evidence as output. Its core is the analysis of the credibility of the claim, based on the overall evidence or counter-evidence from a set of automatically retrieved Web articles. We have developed three methods to this end: a pipeline of classifiers and scoring models, a joint-inference model in the form of a Conditional Random Field, and a deep-learning neural network based on a bidirectional LSTM. In our experiments (see below) – with limited training data – the pipeline architecture performed best. Hence, we focus on this configuration. Note that the scarceness of training samples is typical in coping with misinformation, not just a limitation of our experiments.

Figure 6.1 gives an overview of the system architecture. The pipeline consists of the following stages: (i) *Retrieval* of articles from diverse Web sources by sending the claim text to a search engine, (ii) *Stance Detection* to understand the stance of each article, (iii) *Content Analysis* to understand the credibility of each article by utilizing the language style and stance-related features, (iv) *Credibility Aggregation* to merge these per-article assessments to compute the overall scoring of the claim being *true* or *false*, and (v) *Evidence Extraction* to extract supporting evidence in the form of informative snippets from the relevant web articles.

Method	True-Claims Accuracy (%)	False-Claims Accuracy (%)	Macro-Avg. Accuracy (%)
Pipeline	83.20	80.78	82.00
CRF	71.26	88.74	80.00
LSTM	77.90	78.27	78.09

Table 6.1: Different configurations of CredEye.

The classifiers are trained by distant supervision using data from snopes.com (see Section 4.4.1.1), a popular fact-checking website that *manually* validates Internet rumors, hoaxes, urban legends, and other stories of unknown or questionable origin. We used 5,000 claims from Snopes, each labeled true or false, and retrieved 30 relevant Web articles for each of them. By assuming that the unlabeled Web articles should predominantly inherit the claim's label (hence *distant* supervision), we could train logistic-regression classifiers for per-article stance and per-article credibility. Table 6.1 shows accuracy results for the Snopes data, using 10-fold cross-validation.[1]

6.2.1 Querying the Web

To extract web articles relevant to the input claim, we use the Bing search API, which allows us to restrict results to specific types (e.g., the entire web, only news, only social media, etc.) and geo-locations. Our system supports five such configurations for selecting articles from: (i) the entire web (no restrictions), (ii) all news websites, (iii) popular US news websites, (iv) popular UK news websites, and (v) social media websites (like Quora, Twitter, Facebook, blogs, etc.). For this demo, we focus on English language articles without further restrictions.

Knowledge Base Lookup: Before moving to the next stage of the pipeline, we determine if the credibility of an input claim can be easily assessed by a Knowledge Base (KB) lookup. To this end, we first check if a representative <*subject, verb, object*> triplet could be extracted from the input claim. If yes, we query for the corresponding "subject+verb" and "object+verb", and check if the claim can be assessed from the retrieved instant answer. For instance, given the claim *"Obama was born in Kenya"*, the system queries for "obama+born" in Bing and assesses the claim as false based on the retrieved instant answer. Instead of relying on Bing's internal KB, it is also possible to use any other KB for this lookup.

[1] Data available at https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/impact/web-credibility-analysis/

6.2.2 Stance Detection

False claims are refuted by articles from trusted Web sources. Therefore, it is necessary to understand an article's stance towards the claim. To this end, we divide each retrieved article into a set of overlapping snippets and extract snippets that are strongly related to the claim in terms of unigram and bigram overlap. We use the qualifying snippets to compute support and refute scores, using logistic regression classifiers trained on claims and evidence articles from Snopes. The scores are fed as features into the subsequent content analysis.

6.2.3 Content Analysis

The content analysis of the articles is the core part and distinguishing characteristic of CredEye. It assesses the credibility of each article based on a suite of linguistic features (see Popat et al. [2017] for more details - Chapter 4).

Features: Our hypothesis is that true and thus credible claims are reported in an objective and unbiased language. On the other hand, subjective or sensational style of reporting a claim decreases its credibility. To capture the language style of the article, we derive features from a predefined set of lexicons (e.g., assertive and factive verbs, hedges, report verbs, subjective and biased words, etc.). In addition, the support and refute scores from the stance detection step are used as features.

Classifier: The credibility assessment model is a logistic regression classifier with L1-regularization, distantly trained on Snopes samples.

6.2.4 Credibility Aggregation

Not all Web sources are trustworthy. Hence, to aggregate per-article credibility scores, it is essential to determine the trustworthiness of each article's source.

Source Trustworthiness: Computing the trustworthiness of a source hinges on the following hypothesis: a Web source is trustworthy if it *refutes* non-credible claims and *supports* credible ones. We calculate the trustworthiness $tw(s)$ of source s as :

$$tw(s) = \frac{\#articles\_support\_true + \#articles\_refute\_false}{\#total\_articles} \tag{6.1}$$

where $\#articles\_support\_true$ is the number of articles from s that support credible claims, $\#articles\_refute\_false$ represents the number of articles from s that refute non-credible claims, and $\#total\_articles$ is the total number of articles from s. We use the Snopes training data to pre-compute these trustworthiness scores for a wide variety of sources, including news sites, online communities, Wikipedia, and more. When we encounter a new source which is not present in our training data, we assign a default trustworthiness score of 0.1 (as used in our experiments).

Claim Credibility: Given a claim c and a set of relevant articles $\{a_i\}$ from sources $\{s_i\}$, we aggregate the per-article credibility scores as:

$$P(c = credible) = \frac{\sum_i tw(s_i) * p_{a_i}(c = credible)}{\sum_i tw(s_i)} \tag{6.2}$$

Here, $P(c = credible)$ denotes the aggregated score for the claim being credible, $p_{a_i}(c = credible)$ is the credibility score of a_i, and $tw(s_i)$ is the trustworthiness of s_i. This aggregation penalizes the credibility scores from non-trustworthy sources.

6.2.5 Evidence Extraction

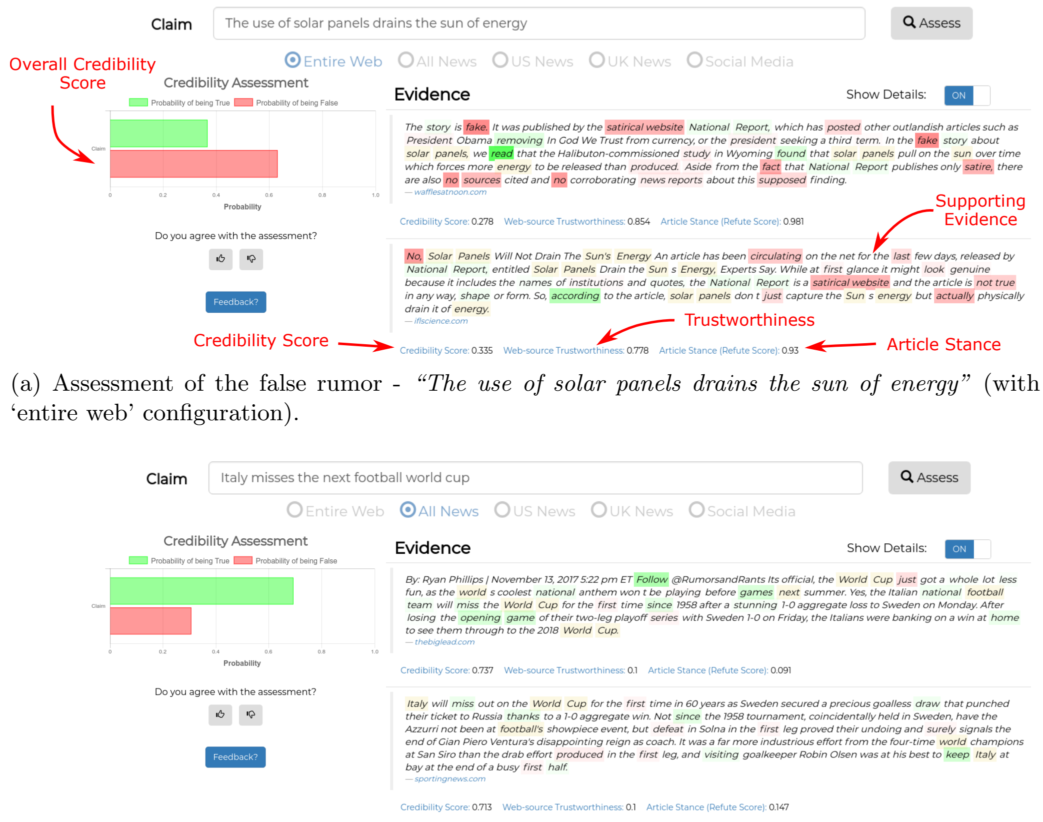
To present users with comprehensible evidence for credibility verdicts, we utilize the snippets of articles extracted in the stance detection step. From each article, CredEye selects the snippet that is most related to the claim and has a support or refute score that is above a threshold and agrees with the overall verdict.

In addition, CredEye enriches the presented snippets by highlighting salient words and bigrams. Words that are also present in the claim are highlighted in *yellow*. Words which contribute most towards the aggregated credibility score are highlighted in different shades of *green* (signaling credibility) and *red* (signaling non-credibility). The intensity of colors reflects the words' importance for the assessment (based on feature weights from the classifier). The highlighted words and bigrams are judiciously selected from the features of the stance detection step, and also from various lexicons of subjective and emotional language (e.g., OpinionFinder MPQA).

6.3 Web Interface

CredEye can be accessed at https://gate.d5.mpi-inf.mpg.de/credeye/ (a recorded screencast available at https://youtu.be/t0SKDjovJiU). Here, we consider two scenarios: (i) a false rumor *"The use of solar panels drains the sun of energy"* with 'entire web' configuration (see Figure 6.2a) and (ii) a true statement *"Italy misses the next football world cup"* with 'all news' configuration (see Figure 6.2b).

As shown in Figure 6.2, the *input area* of CredEye contains a text box where the user can enter any natural language text as an input claim for assessment along with a specific configuration to restrict the article sources. Upon submitting the claim, the back-end server of CredEye carries out its analysis and returns its verdict along



Claim
The use of solar panels drains the sun of energy
Assess
● Entire Web ○ All News ○ US News ○ UK News ○ Social Media
Credibility Assessment
🟩 Probability of being True 🟥 Probability of being False
Claim
0 0.2 0.4 0.6 0.8 1.0
Probability
Do you agree with the assessment?
Feedback?
Evidence
Show Details: ON

Overall Credibility Score

The story is fake. It was published by the satirical website National Report, which has posted other outlandish articles such as President Obama removing In God We Trust from currency, or the president seeking a third term. In the fake story about solar panels, we read that the Halibuton-commissioned study in Wyoming found that solar panels pull on the sun over time which forces more energy to be released than produced. Aside from the fact that National Report publishes only satire, there are also no sources cited and no corroborating news reports about this supposed finding.
— wafflesatnoon.com

Credibility Score: 0.278 Web-source Trustworthiness: 0.854 Article Stance (Refute Score): 0.981

Supporting Evidence

No, Solar Panels Will Not Drain The Sun's Energy An article has been circulating on the net for the last few days, released by National Report, entitled Solar Panels Drain the Sun s Energy, Experts Say. While at first glance it might look genuine because it includes the names of institutions and quotes, the National Report is a satirical website and the article is not true in any way, shape or form. So, according to the article, solar panels don t just capture the Sun s energy but actually physically drain it of energy.
— iflscience.com

Trustworthiness

Credibility Score Credibility Score: 0.335 Web-source Trustworthiness: 0.778 Article Stance (Refute Score): 0.93 Article Stance

(a) Assessment of the false rumor - *"The use of solar panels drains the sun of energy"* (with 'entire web' configuration).

Claim
Italy misses the next football world cup
Assess
○ Entire Web ● All News ○ US News ○ UK News ○ Social Media
Credibility Assessment
🟩 Probability of being True 🟥 Probability of being False
Claim
0 0.2 0.4 0.6 0.8 1.0
Probability
Do you agree with the assessment?
Feedback?
Evidence
Show Details: ON

By: Ryan Phillips | November 13, 2017 5:22 pm ET Follow @RumorsandRants Its official, the World Cup just got a whole lot less fun, as the world s coolest national anthem won t be playing before games next summer. Yes, the Italian national football team will miss the World Cup for the first time since 1958 after a stunning 1-0 aggregate loss to Sweden on Monday. After losing the opening game of their two-leg playoff series with Sweden 1-0 on Friday, the Italians were banking on a win at home to see them through to the 2018 World Cup.
— thebiglead.com

Credibility Score: 0.737 Web-source Trustworthiness: 0.1 Article Stance (Refute Score): 0.091

Italy will miss out on the World Cup for the first time in 60 years as Sweden secured a precious goalless draw that punched their ticket to Russia thanks to a 1-0 aggregate win. Not since the 1958 tournament, coincidentally held in Sweden, have the Azzurri not been at football's showpiece event, but defeat in Solna in the first leg proved their undoing and surely signals the end of Gian Piero Ventura's disappointing reign as coach. It was a far more industrious effort from the four-time world champions at San Siro than the drab effort produced in the first leg, and visiting goalkeeper Robin Olsen was at his best to keep Italy at bay at the end of a busy first half.
— sportingnews.com

Credibility Score: 0.713 Web-source Trustworthiness: 0.1 Article Stance (Refute Score): 0.147

(b) Assessment of the true statement - *"Italy misses the next football world cup"* (with 'all news' configuration).

Figure 6.2: CredEye interface.

with evidence snippets, displayed in the *output area*. The output includes the overall assessment, displayed in the form of green (true) and red (false) bars. There are also buttons for providing feedback.

The most interesting part of the output is the explanation of the assessment, in the form of enriched text snippets from the Web articles that were retrieved during the analysis. As shown in Figure 6.2, salient words in the snippets are highlighted in different colors (see Section 6.2.5). Phrases present in the articles like *"fake"*, *"satirical website"*, *"supposed"*, etc. in Figure 6.2a reduce the credibility of the claim which helps our credibility assessment pipeline to classify it as false. On the other hand, absence of biased and subjective words (decreasing credibility) in addition to objective words like *"follow"*, *"keep"*, *"games"*, etc. in Figure 6.2b increase the credibility of the claim. Hence, our pipeline assesses this factual statement as credible. In addition, CredEye shows the sub-scores from the various stages of its pipeline: the per-article credibility score, the refute score from the stance detection, and the trustworthiness of the source.

6.4 Conclusion

In this chapter, we present a web interface for automatic credibility assessment, CredEye. The CredEye system is a step towards coping with misinformation. One of its limitations is the lack of in-depth understanding of the exact scope and finer tone of claims. For instance, in a claim like "the US Civil War ended slavery world-wide" – it is challenging for the system to understand its finer scope 'world-wide'. Retrieving sufficient evidence or counter-evidence is another bottleneck where we hinge on search-engine results.

7
Determining Stance

Contents

7.1 Introduction

People express their perspectives about controversial claims through various channels like editorials, blog posts, social media, and discussion forums. A better understanding of such controversial claims requires analyzing them from different perspectives. Stance classification is a necessary step for inferring these perspectives in terms of supporting or opposing the claim. Moreover, recent research [FNC-1, 2016; Baly et al., 2018; Chen et al., 2019] has shown stance classification to be a critical step for information credibility and automated fact-checking. In this chapter, we present a neural network model for stance classification leveraging BERT representations and augmenting them with a novel consistency constraint. Experiments on the *Perspectrum* dataset, consisting of claims and users' perspectives from various debate websites, demonstrate the effectiveness of our approach over state-of-the-art baselines.

Prior Work and Limitations: Prior approaches for stance classification proposed in Somasundaran and Wiebe [2010]; Anand et al. [2011]; Walker et al. [2012]; Hasan and Ng [2013, 2014]; Sridhar et al. [2015]; Sun et al. [2018] rely on various linguistic features, e.g., n-grams, dependency parse tree, opinion lexicons, and sentiment to determine the stance of perspectives regarding controversial topics. Ferreira and Vlachos [2016] further incorporate natural language claims and propose a logistic regression model using the lexical and semantic features of claims and perspectives. SemEval tasks [Mohammad et al., 2016; Kochkina et al., 2017] and other approaches [Chen and Ku, 2016; Lukasik et al., 2016; Sobhani et al., 2017] have focused on determining stance only in Tweets.

Bar-Haim et al. [2017] propose classifiers based on hand-crafted lexicons to identify important phrases in perspectives and their consistency with the claim to predict the stance. However, their model critically relies on manual lexicons and assumes that the important phrases in claims are already identified.

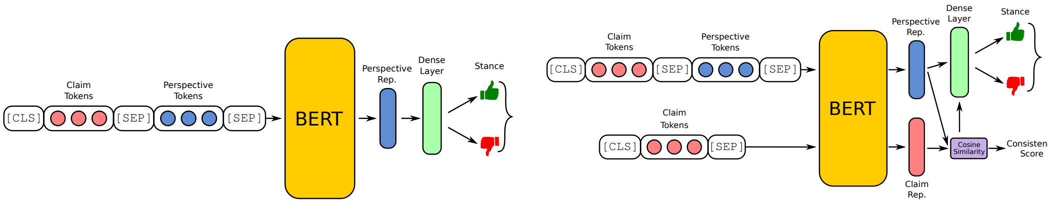
Neural-network-based approaches for stance classification learn the claim and perspective representations separately and later combine them with conditional LSTM encoding [Augenstein et al., 2016], attention mechanisms [Du et al., 2017] or memory networks [Mohtarami et al., 2018]. Some neural network models also incorporate lexical features [Riedel et al., 2017; Hanselowski et al., 2018]. None of these approaches leverage knowledge acquired from massive external corpora.

Approach and Contributions: To overcome the limitations of prior works, we present STANCY, a neural network model for stance classification. Given an input pair of a claim and a user's perspective, our model predicts whether the perspective is *supporting* or *opposing* the claim. For example, the claim *"You have nothing to worry about surveillance, if you have done nothing wrong"* is supported by the user perspective *"Information gathered through surveillance could be used to fight terrorism"* and opposed by another user perspective *"With surveillance, the user privacy will go away!"*.

Our model for stance classification leverages representations from the BERT (Bidirectional Encoder Representations from Transformers) neural network model [Devlin et al., 2019]. BERT is trained on huge text corpora and serves as background knowledge. We fine-tune BERT for our task which also allows us to jointly model claims and perspectives. Furthermore, we enhance our model by augmenting it with a novel consistency constraint to capture agreement between the claim and perspective. Key contributions of this chapter are:

- **Model**: A neural network model for stance classification leveraging BERT representations learned over massive external corpora and a novel consistency constraint to jointly model claims and perspectives.

- **Interpretability**: A simple approach to interpret the contribution of perspective tokens in deciding their stance towards the claim.



(a) $\text{BERT}_{\text{BASE}}$: Fine-tuning BERT for stance classification.

(b) $\text{BERT}_{\text{CONS}}$: Enhancing BERT using the joint loss ($loss_{ce}$ for stance classification and $loss_{cos}$ for consistency).

Figure 7.1: BERT-based methods for determining the stance of the perspective with respect to the claim.

- **Experiments**: Experiments on a recent dataset, *Perspectrum*, highlighting the effectiveness of our approach with error analysis.

7.2 BERT-based Approaches

In this section, first we describe the base model, $\text{BERT}_{\text{BASE}}$, that is adapted for the stance classification [Chen et al., 2019]. Thereafter, we present our consistency-aware model, $\text{BERT}_{\text{CONS}}$.

7.2.1 Adapting BERT for Stance Classification

The goal of the stance classification task is to determine the stance of the user *Perspective (P)* with respect to the *Claim (C)*. Since this task involves a pair of sentences (C and P), we follow the approach for sentence pair classification task as proposed in Devlin et al. [2019]; Chen et al. [2019].

In order to obtain the representation $X^{P|C}$ of P with respect to C, this sentence pair is fused into a single input sequence by using a special classification token (`[CLS]`) and a separator token (`[SEP]`): `[CLS]` C_{toks} `[SEP]` P_{toks} `[SEP]`. The input sequences are tokenized using WordPiece tokenization. The final hidden state representation corresponding to the `[CLS]` token is used as $X^{P|C} \in R^H$. The classification probability is given by passing this representation through the softmax layer:

$$\hat{y} = softmax(X^{P|C}W^T) \tag{7.1}$$

where softmax layer weights $W \in R^{H \times K}$ and K is the number of stance (classification) labels. All the parameters of BERT and W are fine-tuned jointly by minimizing the cross-entropy loss ($loss_{ce}$). The architecture of this model, $\text{BERT}_{\text{BASE}}$, is shown in Figure 7.1a.

7.2.2 Consistency-aware Stance Classification

In this setting, we want to incorporate the consistency between the claim (C) and perspective (P) representations. We hypothesize that the latent representations of claim and perspective should be *dissimilar* if the perspective *opposes* the claim, whereas their representations should be *similar* if the claim is *supported* by the perspective. We capture this with the following components.

Claim Representation: To capture the latent representation of the claim, we use only the claim text as the input sequence to BERT, i.e., `[CLS]` C `[SEP]`. The final hidden state of the first input token (`[CLS]`) is used as the claim's representation $X^C \in R^H$.

Perspective Representation: Latent representation of the perspective (with respect to the claim) is captured by fusing the two sequences as described in Section 7.2.1. We pack the claim and perspective pair as a single input sequence and use the final hidden state of the first input token as the perspective representation $X^{P|C} \in R^H$.

Capturing Consistency: To incorporate the consistency between claim and perspective representations, we use the *cosine embedding loss*:

$$loss_{cos} = \begin{cases} 1 - \cos(X^C, X^{P|C}) & y_{sim} = 1 \\ max(0, \cos(X^C, X^{P|C})) & y_{sim} = -1 \end{cases}$$

where cos(.) is the cosine similarity function. y_{sim} is equal to 1 if the perspective is *supporting* the claim (*similar* representations), and -1 if the claim is *opposed* by the perspective (*dissimilar* representations).

Joint Loss: The classification probabilities are determined by concatenating $X^{P|C}$ and $\cos(X^C, X^{P|C})$ and passing it through a softmax layer. However, unlike the BERT$_{\text{BASE}}$ configuration, parameters of the consistency-aware model are learned by optimizing the joint loss function: $loss = loss_{ce} + loss_{cos}$. With this joint loss function, we enforce consistency between latent representations of the claim and perspective. The architecture of this consistency-aware model, BERT$_{\text{CONS}}$, is shown in Figure 7.1b.

Split	Supporting Pairs	Opposing Pairs	Total Pairs
train	3603	3404	7007
dev	1051	1045	2096
test	1471	1302	2773
Total	6125	5751	11876

Table 7.1: Perspectrum data statistics.

7.3 Experimental Setup

For our experiments, we consider the base version of BERT[1] with 12 layers, 768 hidden size, and 12 attention heads. We fine-tune BERT-based models using the Adam optimizer with learning rates $\{1, 3, 5\} \times 10^{-5}$ and training batch sizes $\{24, 28, 32\}$. We choose the best parameters based on the development split of the dataset. For measuring the performance, we use per-class and macro-averaged Precision/Recall/F1.

7.3.1 Dataset

We evaluate our approach on the *Perspectrum* dataset [Chen et al., 2019]. *Perspectrum* contains claims and users' perspectives from various online debate websites like `idebate.com`, `debatewise.org`, and `procon.org`. Each claim has different perspectives along with the stance (*supporting* or *opposing* the claim). We use the same train/dev/test split as provided in the released dataset. Statistics of the dataset is shown in Table 7.1.

7.3.2 Baselines

We use the following baselines:

- **LSTM**: A long short-term memory (LSTM) model, in which we pass the claim and perspective word representations (using GloVE-6B word embeddings of size 300) through a bidirectional LSTM. Then we concatenate the final hidden states of the claim and perspective, and pass it through dense layers with ReLU activations.

- **ESIM**: An enhanced sequential inference model (ESIM) for natural language inference proposed in Chen et al. [2017].

[1] GitHub implementation: `https://github.com/huggingface/pytorch-transformers` (accessed 15 July, 2019)

Approach	Supporting			Opposing			Overall (Macro)		
	Prec.	**Recall**	**F1**	**Prec.**	**Recall**	**F1**	**Prec.**	**Recall**	**F1**
LSTM	63.42	58.80	61.02	56.99	61.67	59.24	60.20	60.24	60.13
ESIM	64.38	61.32	62.81	58.53	61.67	60.06	61.46	61.50	61.44
MLP	64.53	60.98	62.71	58.50	62.14	60.26	61.51	61.56	61.48
WordAttn	64.43	63.43	63.93	59.40	60.45	59.92	62.07	62.03	62.04
LangFeat	63.74	75.05	68.94	64.75	51.77	57.53	64.24	63.41	63.23
BERT$_{\text{BASE}}$	78.43	80.08	79.25	76.95	**75.12**	76.02	77.69	77.60	77.63
BERT$_{\text{CONS}}$	**79.05**	**84.64**	**81.75**	**81.14**	74.65	**77.76**	**80.09**	**79.65**	**79.95**
Human	-	-	-	-	-	-	91.3	90.6	90.9

Table 7.2: Comparison of our approach BERT$_{\text{CONS}}$ with different baseline models for stance classification.

- **MLP**: Multi-layer perceptron (MLP) based model using lexical and similarity-based features – presented as a *simple but tough-to-beat baseline* for stance detection in Riedel et al. [2017].

- **WordAttn**: Our implementation of word-by-word attention-based model using long short-term memory networks [Rocktäschel et al., 2016].

- **LangFeat**: A random forest classifier using linguistic lexicons like NRC lexicon[2] [Mohammad and Turney, 2010], hedges (e.g., *possibly*, *might*, etc.), positive/negative sentiment words[3] [Hu and Liu, 2004], MPQA subjective lexicon[4] [Wilson et al., 2005] and bias lexicon [Recasens et al., 2013] along with sentiment scores as features.

- **BERT$_{\text{BASE}}$**: Approach proposed in Chen et al. [2019] (as described in Section 7.2.1).

- **Human**: Human performance on this task as reported in Chen et al. [2019].

7.4 Results and Discussion

Stance classification performance of our model and the baselines on the *test* split of the Perspectrum dataset are presented in Table 7.2. Our consistency-aware model BERT$_{\text{CONS}}$ outperforms all the other baselines. It achieves a performance improvement of about 2 points in F1-score over the strong baseline corresponding to the BERT$_{\text{BASE}}$ model (p-value of $4.985e{-}4$ as per the McNemar test). This highlights the value addition achieved by incorporating consistency cues. Since

[2] https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm (accessed 15 July, 2019)
[3] http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar (accessed 15 July, 2019)
[4] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/ (accessed 15 July, 2019)

Opposing Class	Supporting Class
unauthorized, falsely, even though, unlike, cannot, not everyone, could strike, could further weaken, jeopardize, impacts, may not provide, ...	enabling, ensuring, prevail, positive discrimination, gains, help reduce, would improve, right, would allow, encourage, more effective, ...

Table 7.3: Top phrases for determining stance.

the BERT-based models incorporate the knowledge acquired from massive external corpora, our model, BERT$_{\text{CONS}}$, captures better semantics and outperforms the other baselines.

7.4.1 Interpreting Token-level Contribution

Due to the massive structure of BERT with a complex attention mechanism, it is difficult to interpret the significance of different lexical units in the text. Therefore, we propose a simple technique to interpret the contribution of each token in the text in determining the stance.

Given the claim (C) and perspective (P) pair, we tokenize P into *phrases*. We record the change in stance classification probabilities by adding one perspective phrase at a time to the input:

$$\Delta_i = |BERT_{CONS}(C, P_i) - BERT_{CONS}(C, P_{i-1})|$$

where P_i is the prefix of P up to the i^{th} phrase. This helps us in understanding the contribution of each perspective phrase towards determining the stance – the larger the change in the classification probabilities, the larger the contribution. For this analysis, we consider unigrams and chunks from a shallow parser as phrases. The top contributing phrases for the *supporting* and *opposing* classes are shown in Table 7.3.

7.4.2 Error Analysis

In this section, we analyze why the task of stance classification is challenging and why the performance of the best model configuration is far from human performance as observed by the performance gap in Table 7.2.

Negations: One of the major challenges in solving this task is understanding negations and their scope. For example, given the claim *"College education is worth it"*, the perspective *"Many college graduates are employed in jobs that* ***do not*** *require college degrees"* is *opposing* the claim. However, our model is not able to capture that the negation phrase '*do not require*' opposes the claim. On the other hand, the presence of negation in the perspective does not necessarily imply that it is *opposing* the claim. Contrast this with the claim *"Chess must be at the Olympics"* and perspective *"Chess is currently **not** an Olympic sport, but it should be"* – where the negation is merely a part of the statement and the stance is given by the discourse segment following '*but*'.

Commonsense: Determining the stance may require commonsense knowledge. For example, the claim *"Chess must be at the Olympics"* is *opposed* by the perspective *"Olympic sports are supposed to be **physical**"*. To understand this, the model should have the background knowledge that chess is not a physical sport.

Semantics: Understanding the stance also involves a deeper understanding of semantics. For example, given the claim *"Make all museums free of charge"* is *opposed* by the perspective *"State funding should be used **elsewhere**"*. Here, the word *'elsewhere'* is the key cue which determines the stance. However, the presence of the word *'elsewhere'* does not necessarily imply that the perspective is opposing the claim. For instance, the perspective *"We could spend the money **elsewhere**"* is *supporting* the claim *"The EU should significantly reduce the amount it spends on agricultural production subsidies"*. Hence, the polarity of the word *'elsewhere'* is determined by the context and semantics of the statement.

7.5 Conclusion

In this chapter, we propose a consistency-aware neural network model for stance classification. Our model leverages representations from the BERT model trained over massive external corpora and a novel consistency constraint to jointly model claims and perspectives. Our experiments on a recent benchmark highlight the advantages of our approach. We also study the gap in human performance and the performance of the best model for stance classification.

8 Conclusions and Outlook

This thesis investigates the problem of automated credibility assessment of textual information. We propose a general framework for assessing the credibility of textual claims, in an open-domain setting, without any assumptions about the structure of the claim, or characteristics of the community where it is made. This framework lays the foundation for our feature-based and neural-network-based models for credibility analysis.

Our credibility assessment models, together with the extensive experiments on real-world datasets, highlight the significance of considering external evidence for automated credibility assessment. Specifically, our feature-based model emphasizes the importance of capturing the interplay between the language style of the evidence articles, their stance towards the claim, and the reliability of the underlying web sources. Our neural-network-based model further eliminates the dependency on feature engineering and external lexicons. Most importantly, unlike prior methods, both our models have the ability to explain why a certain statement is credible or not by extracting interpretable evidence from judiciously selected web-sources. Our Web interface, CredEye, enables users to assess and dissect the credibility of a textual claim. In addition, the stance classification model we propose highlights the effectiveness of capturing semantic consistency for predicting whether the user perspective is supporting or refuting the controversial claim.

Going forward, there are several interesting challenges that need to be addressed for solving this extremely complex and challenging task of automated credibility assessment. Some of these extensions are as follows:

- **Incorporating temporal information:** The spread of misinformation is a dynamic process. Hence, capturing the temporal footprint of information also plays a critical role in understanding how misinformation propagates. As revealed by our results, incorporating the dynamic trend in which various web articles support or refute the claim helps in assessing its credibility. However, capturing temporal footprints of all the web content is a daunting task which requires continuous monitoring of the entire web. More work is required to fully integrate temporal information with credibility assessment models.

- **Multifaceted credibility and stance:** Our models for automated credibility assessment assign a credibility score indicating how likely the information is true or false. However, information credibility is complex and multifaceted. For instance, the given piece of information could be only partially credible within a specific context. Similarly, while determining the stance of a user comment towards the controversial claim, we predict how likely the comment supports or opposes the claim. However, the perspective could be only partially supporting the claim, or it may have no opinion towards the claim. In order to better understand this multifaceted nature of information credibility and stance, further research is required.

- **Automatic detection of "check-worthy" claims:** Credibility assessment methods proposed in this dissertation expect to get the controversial claim as input. Our models can be further extended by adding a pre-processing module which monitors web articles, social media, news, etc., and automatically detects the claims that require credibility assessment.

- **Credibility assessment of multimedia content:** With the rise of technology, multimedia content has also been plagued by misinformation in the form of *deepfake*, manipulated photos, etc. Hence, another interesting extension to this thesis could be to develop models which assess the credibility of the multimedia content.

Bibliography

Aseel Addawood and Masooda Bashir. “what is your evidence?” a study of controversial topics on social media. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 1–11, 2016. (Cited on page 20.)

B. Thomas Adler and Luca de Alfaro. A content-driven reputation system for the wikipedia. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 261–270, 2007. (Cited on page 19.)

Sadia Afroz, Michael Brennan, and Rachel Greenstadt. Detecting hoaxes, frauds, and deception in writing style online. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, SP '12, pages 461–475, 2012. (Cited on pages 4 and 17.)

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 30–38, 2011. (Cited on page 17.)

Rajesh K. Aggarwal and Guojun Wu. Stock market manipulations. *The Journal of Business*, 79(4):1915–1953, 2006. (Cited on page 1.)

T. Ahmad, H. Akhtar, A. Chopra, and M. W. Akhtar. Satire detection from web documents using machine learning methods. In *2014 International Conference on Soft Computing and Machine Intelligence*, pages 102–105, 2014. (Cited on page 17.)

Leman Akoglu, Rishi Chandy, and Christos Faloutsos. Opinion fraud detection in online reviews by network effects. In *International AAAI Conference on Web and Social Media*, 2013. (Cited on page 16.)

L. Alvisi, A. Clement, A. Epasto, S. Lattanzi, and A. Panconesi. Sok: The evolution of sybil defense via social networks. In *2013 IEEE Symposium on Security and Privacy*, pages 382–396, 2013. (Cited on page 14.)

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E. Fox Tree, Robeson Bowmani, and Michael Minor. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 1–9, 2011. (Cited on pages 17 and 81.)

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, 2016. (Cited on pages 18 and 82.)

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *J. Mach. Learn. Res.*, 11:1803–1831, 2010. (Cited on page 20.)

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364, 2013. (Cited on page 54.)

Meital Balmas. When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism. *Communication Research*, 41(3):430–454, 2014. (Cited on page 1.)

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 21–27, 2018. (Cited on pages 17 and 81.)

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, 2017. (Cited on pages 18 and 82.)

P. Bellot, A. Doucet, S. Geva, S. Gurajada, J. Kamps, G. Kazai, M. Koolen, A. Mishra, V. Moriceau, J. Mothe, M. Preminger, E. SanJuan, R. Schenkel, X. Tannier, M. Theobald, M. Trappett, A. Trotman, M. Sanderson, F. Scholer, and Q. Wang. Report on inex 2013. *SIGIR Forum*, 47(2):21–32, 2013. (Cited on page 20.)

Alessandro Bessi and Emilio Ferrara. Social bots distort the 2016 u.s. presidential election online discussion. *First Monday*, 21(11), 2016. (Cited on page 14.)

Alex Beutel, Wanhong Xu, Venkatesan Guruswami, Christopher Palow, and Christos Faloutsos. Copycatch: Stopping group attacks by spotting lockstep behavior in social networks. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 119–130, 2013. (Cited on page 14.)

Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011. (Cited on page 1.)

Paul R. Brewer, Dannagal Goldthwaite Young, and Michelle Morreale. The Impact of Real News about "Fake News": Intertextual Processes and Political Satire. *International Journal of Public Opinion Research*, 25(3):323–343, 2013. (Cited on page 1.)

Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30(1-7):107–117, 1998. (Cited on pages 6 and 18.)

Clint Burfoot and Timothy Baldwin. Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 161–164, 2009. (Cited on page 17.)

Marc-Allen Cartright, Henry A. Feild, and James Allan. Evidence finding using a collection of books. In *Proceedings of the 4th ACM Workshop on Online Books, Complementary Social Media and Crowdsourcing*, BooksOnline '11, pages 11–18, 2011. (Cited on page 20.)

Carlos Castillo, Debora Donato, Aristides Gionis, Vanessa Murdock, and Fabrizio Silvestri. Know your neighbors: Web spam detection using the web topology. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 423–430, 2007. (Cited on page 19.)

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 675–684, 2011. (Cited on page 13.)

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, 2017. (Cited on page 85.)

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. Seeing things from a different angle: Discovering diverse perspectives about claims. 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)*, pages 542–557, 2019. (Cited on pages 9, 17, 18, 81, 83, 85, and 86.)

Wei-Fan Chen and Lun-Wei Ku. UTCNN: a deep learning model of stance classification on social media text. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1635–1645, 2016. (Cited on pages 18 and 82.)

Yimin Chen, Niall J. Conroy, and Victoria L. Rubin. Misleading online content: Recognizing clickbait as “false news”. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, WMDD ’15, pages 15–19, 2015. (Cited on page 4.)

Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW ’17, pages 1217–1230, 2017. (Cited on page 16.)

Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, 9(6):811–824, 2012. (Cited on page 14.)

Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web*, WWW ’03, pages 519–528, 2003. (Cited on page 17.)

Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW ’16 Companion, pages 273–274, 2016. (Cited on page 14.)

Luca De Alfaro, Ashutosh Kulshreshtha, Ian Pye, and B. Thomas Adler. Reputation systems for open collaboration. *Commun. ACM*, 54(8):81–87, 2011. (Cited on page 19.)

Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113 (3):554–559, 2016. (Cited on page 15.)

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEvalACL 2017, Vancouver, Canada, August 3-4, 2017*, pages 69–76, 2017. (Cited on pages 18 and 66.)

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, 2019. (Cited on pages 3, 9, 18, 82, and 83.)

John P. Dickerson, Vadim Kagan, and V. S. Subrahmanian. Using sentiment to detect bots on twitter: Are humans more opinionated than bots? In *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '14, pages 620–627, 2014. (Cited on page 14.)

Xin Luna Dong and Divesh Srivastava. Compact explanation of data fusion decisions. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 379–390, 2013. (Cited on page 11.)

Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. Integrating conflicting data: The role of source dependence. *Proc. VLDB Endow.*, 2(1):550–561, 2009. (Cited on pages 11 and 12.)

Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. Knowledge-based trust: Estimating the trustworthiness of web sources. *Proc. VLDB Endow.*, 8(9):938–949, 2015. (Cited on pages 19 and 36.)

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. Stance classification with target-specific neural attention networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI '17, 2017. (Cited on pages 18 and 82.)

Javid Ebrahimi, Dejing Dou, and Daniel Lowd. A joint sentiment-target-stance model for stance classification in tweets. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2656–2665, 2016. (Cited on page 18.)

Omar Enayet and Samhaa R. El-Beltagy. Niletmrg at semeval-2017 task 8: Determining rumour and veracity support for rumours on twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEvalACL 2017, Vancouver, Canada, August 3-4, 2017*, pages 470–474, 2017. (Cited on page 70.)

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, 2008. (Cited on pages 42 and 45.)

Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Commun. ACM*, 59(7):96–104, 2016. (Cited on page 14.)

William Ferreira and Andreas Vlachos. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168, 2016. (Cited on pages 18 and 82.)

Andrew J. Flanagin and Miriam J. Metzger. Digital media and youth: Unparalleled opportunity and unprecedented responsibility. *The John D. and Catherine T. MacArthur FoundationSeries on Digital Media and Learning*, pages 5–28, 2008. (Cited on page 6.)

FNC-1. Fake news challenge stage 1 (fnc-1): Stance detection, 2016. (Cited on pages 17 and 81.)

Adrien Friggeri, Lada Adamic, Dean Eckles, and Justin Cheng. Rumor cascades. In *International AAAI Conference on Web and Social Media*, 2014. (Cited on page 14.)

Alban Galland, Serge Abiteboul, Amélie Marian, and Pierre Senellart. Corroborating information from disagreeing views. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 131–140, 2010. (Cited on pages 11 and 12.)

Jing Gao, Qi Li, Bo Zhao, Wei Fan, and Jiawei Han. Truth discovery and crowdsourcing aggregation: A unified perspective. *Proc. VLDB Endow.*, 8(12): 2048–2049, 2015. (Cited on page 11.)

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 913–922, 2018. (Cited on page 15.)

Zafar Gilani, Liang Wang, Jon Crowcroft, Mario Almeida, and Reza Farahbakhsh. Stweeler: A framework for twitter bot analysis. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, pages 37–38, 2016. (Cited on page 14.)

Ben Goertzel and Cassio Pennachin. *Artificial General Intelligence (Cognitive Technologies)*. 2007. (Cited on page 3.)

Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. Bidirectional lstm networks for improved phoneme classification and recognition. In *Proceedings of the 15th International Conference on Artificial Neural Networks: Formal Models and Their Applications - Volume Part II*, ICANN'05, pages 799–804, 2005. (Cited on page 62.)

Aditi Gupta and Ponnurangam Kumaraguru. Credibility ranking of tweets during high impact events. In *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, PSOSM '12, pages 2:2–2:8, 2012. (Cited on page 13.)

Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *WWW Companion 2013*, WWW '13 Companion, pages 729–736, 2013. (Cited on pages 1 and 13.)

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, 2003. (Cited on page 19.)

Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. Combating web spam with trustrank. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, VLDB '04, pages 576–587, 2004. (Cited on page 19.)

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, 2018. (Cited on pages 18 and 82.)

Christopher Harris. Detecting deceptive opinion spam using human computation. In *AAAI Workshops*, 2012. (Cited on page 16.)

Kazi Saidul Hasan and Vincent Ng. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1348–1356, 2013. (Cited on pages 17 and 81.)

Kazi Saidul Hasan and Vincent Ng. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762, 2014. (Cited on page 81.)

Naeemul Hassan, Bill Adair, James Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. The quest to automate fact-checking. *Proceedings of the 2015 Computation + Journalism Symposium*, 2015. (Cited on page 2.)

Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. Claimbuster: The first-ever end-to-end fact-checking system. *Proc. VLDB Endow.*, 10(12):1945–1948, 2017. (Cited on page 73.)

Bryan Hooi, Neil Shah, Alex Beutel, Stephan Günnemann, Leman Akoglu, Mohit Kumar, Disha Makhija, and Christos Faloutsos. BIRDNEST: bayesian inference for ratings-fraud detection. In *Proceedings of the 2016 SIAM International Conference on Data Mining, Miami, Florida, USA, May 5-7, 2016*, pages 495–503, 2016. (Cited on page 16.)

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, 2004. (Cited on page 86.)

Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2972–2978, 2016. (Cited on page 13.)

Nitin Jindal and Bing Liu. Analyzing and detecting review spam. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, pages 547–552, 2007. (Cited on page 16.)

Nitin Jindal and Bing Liu. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 219–230, 2008. (Cited on page 16.)

Sepandar D. Kamvar, Mario T. Schlosser, and Hector Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pages 640–651, 2003. (Cited on page 18.)

Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 324–332, 2018. (Cited on page 14.)

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. (Cited on page 67.)

Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46 (5):604–632, 1999. (Cited on pages 6 and 18.)

Elena Kochkina, Maria Liakata, and Isabelle Augenstein. Turing at SemEval-2017 task 8: Sequential approach to rumour stance classification with branch-LSTM. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 475–480, 2017. (Cited on pages 18 and 82.)

Srijan Kumar and Neil Shah. False information on web and social media: A survey. *ArXiv*, abs/1804.08559, 2018. (Cited on pages 1 and 13.)

Srijan Kumar, Robert West, and Jure Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 591–602, 2016. (Cited on pages 3, 15, 22, and 36.)

Srijan Kumar, Justin Cheng, Jure Leskovec, and V.S. Subrahmanian. An army of me: Sockpuppets in online discussion communities. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 857–866, 2017. (Cited on pages 15 and 16.)

Srijan Kumar, Bryan Hooi, Disha Makhija, Mohit Kumar, Christos Faloutsos, and V.S. Subrahmanian. Rev2: Fraudulent user prediction in rating platforms. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 333–341, 2018. (Cited on page 16.)

Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. Prominent features of rumor propagation in online social media. In *2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, December 7-10, 2013*, pages 1103–1108, 2013. (Cited on pages 14 and 48.)

Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1675–1684, 2016. (Cited on page 19.)

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Interpretable & explorable approximations of black box models. *CoRR*, abs/1707.01154, 2017. (Cited on page 19.)

Thomas Lavergne, Tanguy Urvoy, and François Yvon. Detecting fake content with relative entropy scoring. In *PAN 2008*, 2008. (Cited on page 13.)

Kyumin Lee, James Caverlee, and Steve Webb. Uncovering social spammers: Social honeypots + machine learning. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 435–442, 2010. (Cited on page 14.)

Kyumin Lee, Brian David Eoff, and James Caverlee. Seven months with the devils: a long-term study of content polluters on twitter. In *In AAAI International Conference on Weblogs and Social Media (ICWSM)*, 2011. (Cited on page 14.)

Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Ann. Appl. Stat.*, 9(3):1350–1371, 2015. (Cited on page 19.)

Fangtao Li, Minlie Huang, Yi Yang, and Xiaoyan Zhu. Learning to identify review spam. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, pages 2488–2493, 2011a. (Cited on page 16.)

Huayi Li, Zhiyuan Chen, Arjun Mukherjee, Bing Liu, and Jidong Shao. Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns. In *International AAAI Conference on Web and Social Media*, 2015a. (Cited on page 16.)

Huayi Li, Geli Fei, Shuai Wang, Bing Liu, Weixiang Shao, Arjun Mukherjee, and Jidong Shao. Bimodal distribution and co-bursting in review spam detection. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1063–1072, 2017. (Cited on page 16.)

Jiwei Li, Myle Ott, Claire Cardie, and Eduard Hovy. Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1566–1576, 2014a. (Cited on page 16.)

Qi Li, Yaliang Li, Jing Gao, Lu Su, Bo Zhao, Murat Demirbas, Wei Fan, and Jiawei Han. A confidence-aware approach for truth discovery on long-tail data. *Proc. VLDB Endow.*, 8(4):425–436, 2014b. (Cited on pages 11 and 12.)

Qi Li, Yaliang Li, Jing Gao, Bo Zhao, Wei Fan, and Jiawei Han. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 1187–1198, 2014c. (Cited on pages 11, 12, and 19.)

Xian Li, Weiyi Meng, and Clement Yu. T-verifier: Verifying truthfulness of fact statements. In *Proceedings of the 2011 IEEE 27th International Conference on Data Engineering*, ICDE '11, pages 63–74, 2011b. (Cited on pages 12, 22, and 36.)

Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, and Divesh Srivastava. Truth finding on the deep web: Is the problem solved? *Proc. VLDB Endow.*, 6(2):97–108, 2012. (Cited on pages 11, 12, 22, and 36.)

Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. A survey on truth discovery. *SIGKDD Explorations*, 17(2):1–16, 2015b. (Cited on pages 6, 11, 22, and 60.)

Yaliang Li, Qi Li, Jing Gao, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. On the discovery of evolving truth. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 675–684, 2015c. (Cited on pages 11, 12, 22, and 36.)

Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Wirawan Lauw. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 939–948, 2010. (Cited on page 16.)

Chenghua Lin, Yulan He, and Richard Everson. Sentence subjectivity detection with weakly-supervised learning. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1153–1161, 2011. (Cited on page 17.)

Tal Linzen, Grzegorz Chrupała, and Afra Alishahi. Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP. 2018. (Cited on page 19.)

Zachary C. Lipton. The mythos of model interpretability. *Queue*, 16(3):30:31–30:57, 2018. (Cited on page 19.)

Bing Liu. *Sentiment Analysis and Opinion Mining.* 2012. (Cited on page 17.)

Elizabeth F. Loftus. Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory*, 12(4):361–366, 2005. (Cited on page 2.)

Michal Lukasik, P. K. Srijith, Duy Vu, Kalina Bontcheva, Arkaitz Zubiaga, and Trevor Cohn. Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 393–398, 2016. (Cited on pages 18 and 82.)

Shanshan Lyu, Wentao Ouyang, Huawei Shen, and Xueqi Cheng. Truth discovery by claim and source embedding. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 2183–2186, 2017. (Cited on page 11.)

Fenglong Ma, Yaliang Li, Qi Li, Minghui Qiu, Jing Gao, Shi Zhi, Lu Su, Bo Zhao, Heng Ji, and Jiawei Han. Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 745–754, 2015. (Cited on pages 11 and 12.)

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 3818–3824, 2016. (Cited on page 13.)

Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis: Can we trust what we rt? In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 71–79, 2010. (Cited on pages 1 and 14.)

Tsvetomila Mihaylova, Georgi Karadzhov, Pepa Atanasova, Ramy Baly, Mitra Mohtarami, and Preslav Nakov. SemEval-2019 task 8: Fact checking in community question answering forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 860–869, 2019. (Cited on page 15.)

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *CoRR*, abs/1706.07269, 2017. (Cited on page 19.)

Tanushree Mitra and Eric Gilbert. CREDBANK: A large-scale social media corpus with associated credibility annotations. In *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015*, pages 258–267, 2015. (Cited on page 13.)

Tanushree Mitra, Graham P. Wright, and Eric Gilbert. A parsimonious language model of social media credibility across disparate events. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 126–145, 2017. (Cited on page 13.)

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, 2016. (Cited on pages 18 and 82.)

Saif M. Mohammad and Peter D. Turney. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, 2010. (Cited on page 86.)

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. Stance and sentiment in tweets. *ACM Trans. Internet Technol.*, 17(3):26:1–26:23, 2017. (Cited on page 18.)

Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Lluís Màrquez, and Alessandro Moschitti. Automatic stance detection using end-to-end memory networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 767–776, 2018. (Cited on pages 18 and 82.)

C.A. Morgan, Steven Southwick, George Steffian, Gary A. Hazlett, and Elizabeth F. Loftus. Misinformation can influence memory for recently experienced, highly stressful events. *International Journal of Law and Psychiatry*, 36(1):11–17, 2013. (Cited on page 2.)

L. Mui, M. Mohtashemi, and A. Halberstadt. A computational model of trust and reputation. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences*, pages 2431–2439, 2002. (Cited on page 19.)

Arjun Mukherjee, Bing Liu, and Natalie Glance. Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 191–200, 2012. (Cited on page 16.)

Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Natalie Glance. What yelp fake review filter might be doing? In *International AAAI Conference on Web and Social Media*, 2013. (Cited on page 16.)

Subhabrata Mukherjee. *Probabilistic graphical models for credibility analysis in evolving online communities.* PhD thesis, Ph. D. Thesis, Computer Science Department, Saarland University, Saarbrücken, Germany, 2017. (Cited on page 15.)

Subhabrata Mukherjee and Gerhard Weikum. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 353–362, 2015. (Cited on pages 15, 24, 25, 30, 50, 65, and 69.)

Subhabrata Mukherjee, Gerhard Weikum, and Cristian Danescu-Niculescu-Mizil. People on drugs: Credibility of user statements in health communities. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 65–74, 2014. (Cited on pages 15, 22, 24, and 36.)

Subhabrata Mukherjee, Sourav Dutta, and Gerhard Weikum. Credible review detection with limited information using consistency features. In Paolo Frasconi, Niels Landwehr, Giuseppe Manco, and Jilles Vreeken, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 195–213, 2016. (Cited on page 15.)

Subhabrata Mukherjee, Kashyap Popat, and Gerhard Weikum. Exploring latent semantic factors to find useful product reviews. In *Proceedings of the Seventeenth SIAM International Conference on Data Mining (SDM 2017)*, pages 480–488, 2017. (Cited on page 15.)

Ndapandula Nakashole and Tom M. Mitchell. Language-aware truth assessment of fact candidates. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 1009–1019, 2014. (Cited on pages 12, 22, 25, 30, 36, 40, 50, and 51.)

Preslav Nakov, Tsvetomila Mihaylova, Lluís Màrquez, Yashkumar Shiroya, and Ivan Koychev. Do not trust the trolls: Predicting credibility in community question answering forums. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 551–560, 2017. (Cited on page 15.)

Nam P. Nguyen, Guanhua Yan, My T. Thai, and Stephan Eidenbenz. Containment of misinformation spread in online social networks. In *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, pages 213–222, 2012. (Cited on page 14.)

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 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*, pages 309–319, 2011. (Cited on page 16.)

Myle Ott, Claire Cardie, and Jeffrey T. Hancock. 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, 2013. (Cited on page 16.)

Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, 2010. (Cited on page 17.)

Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, 2004. (Cited on page 17.)

Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(12):1–135, 2008. doi: 10.1561/1500000011. (Cited on page 17.)

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, 2002. (Cited on page 17.)

Jeff Pasternack and Dan Roth. Knowing what to believe (when you already know something). In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 877–885, 2010. (Cited on pages 11, 12, 50, and 51.)

Jeff Pasternack and Dan Roth. Making better informed trust decisions with generalized fact-finding. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, pages 2324–2329, 2011. (Cited on pages 11, 12, 50, and 51.)

Jeff Pasternack and Dan Roth. Latent credibility analysis. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 1009–1020, 2013. (Cited on pages 11 and 12.)

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. (Cited on page 67.)

María del Pilar Salas-Zárate, Mario Andrés Paredes-Valverde, Miguel Ángel Rodriguez-García, Rafael Valencia-García, and Giner Alor-Hernández. Automatic detection of satire in twitter: A psycholinguistic-based approach. *Knowledge-Based Systems*, 128:20–33, 2017. (Cited on page 17.)

Kashyap Popat. Assessing the credibility of claims on the web. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 735–739, 2017. (Cited on page 10.)

Kashyap Popat. "Interpretable Credibility Assessment of Web Claims" by Kashyap Popat with Prateek Jain As Coordinator. *SIGWEB Newsl.*, (Summer):3:1–3:2, 2018. (Cited on page 10.)

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, 2016. (Cited on pages 7, 8, 36, 40, 42, 43, 47, 48, 50, 51, and 55.)

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 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, 2017. (Cited on pages 8, 17, 60, 63, 65, 68, and 76.)

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. Credeye: A credibility lens for analyzing and explaining misinformation. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, 2018a. (Cited on page 9.)

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, 2018b. (Cited on page 8.)

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. STANCY: Stance classification based on consistency cues. 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)*, EMNLP '19, 2019. (Cited on page 9.)

Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1589–1599, 2011. (Cited on page 13.)

Walter Quattrociocchi, Antonio Scala, and C Sunstein. Echo chambers on facebook. *SSRN Electronic Journal*, 2016. (Cited on page 15.)

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 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, 2017. (Cited on pages 4, 12, 60, 63, 65, and 68.)

J. Ratkiewicz, M. D. Conover, M. Meiss, A. Flammini, and F. Menczer. Detecting and tracking political abuse in social media. In *In Proceedings of the 5th AAAI International Conference on Weblogs and Social Media (ICWSM'11)*, 2011a. (Cited on page 15.)

Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. Truthy: Mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 249–252, 2011b. (Cited on page 15.)

Kumar Ravi and Vadlamani Ravi. A novel automatic satire and irony detection using ensembled feature selection and data mining. *Knowledge-Based Systems*, 120:15–33, 2017. (Cited on page 17.)

Shebuti Rayana and Leman Akoglu. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 985–994, 2015. (Cited on page 16.)

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, 2013. (Cited on pages 17, 24, and 86.)

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, 2016. (Cited on page 20.)

Benjamin Riedel, Isabelle Augenstein, Georgios P. Spithourakis, and Sebastian Riedel. A simple but tough-to-beat baseline for the fake news challenge stance detection task. *CoRR*, abs/1707.03264, 2017. (Cited on pages 18, 82, and 86.)

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, 2015. (Cited on page 20.)

Marko Robnik-Šikonja and Igor Kononenko. Explaining classifications for individual instances. *IEEE Trans. on Knowl. and Data Eng.*, 20(5):589–600, 2008. (Cited on page 19.)

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomas Kocisky, and Phil Blunsom. Reasoning about entailment with neural attention. In *International Conference on Learning Representations (ICLR)*, 2016. (Cited on page 86.)

Victoria L. Rubin, Yimin Chen, and Niall J. Conroy. Deception detection for news: Three types of fakes. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, ASIST '15, pages 83:1–83:4, 2015. (Cited on page 17.)

Natali Ruchansky, Sungyong Seo, and Yan Liu. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 797–806, 2017. (Cited on page 13.)

Mehdi Samadi. *Facts and Reasons: Anytime Web Information Querying to Support Agents and Human Decision Making.* PhD thesis, Ph. D. Thesis, Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, USA, 2015. (Cited on page 50.)

Mehdi Samadi, Partha Talukdar, Manuela Veloso, and Manuel Blum. Claimeval: Integrated and flexible framework for claim evaluation using credibility of sources. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 222–228, 2016. (Cited on page 12.)

Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *CoRR*, abs/1708.08296, 2017. (Cited on page 20.)

Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. The spread of low-credibility content by social bots. In *Nature Communications*, 2018. (Cited on page 14.)

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36, 2017. (Cited on pages 1, 6, and 13.)

Aameek Singh and Ling Liu. Trustme: Anonymous management of trust relationships in decentralized p2p systems. In *Proceedings of the 3rd International Conference on Peer-to-Peer Computing*, P2P '03, pages 142–, 2003. (Cited on page 18.)

Vikram Singh, Sunny Narayan, Md. Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. IITP at semeval-2017 task 8 : A supervised approach for rumour evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEvalACL 2017, Vancouver, Canada, August 3-4, 2017*, pages 497–501, 2017. (Cited on page 70.)

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557, 2017. (Cited on pages 18 and 82.)

Swapna Somasundaran and Janyce Wiebe. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 226–234, 2009. (Cited on page 17.)

Swapna Somasundaran and Janyce Wiebe. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124, 2010. (Cited on pages 17 and 81.)

Dhanya Sridhar, Lise Getoor, and Marilyn Walker. Collective stance classification of posts in online debate forums. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, pages 109–117, 2014. (Cited on page 17.)

Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. Joint models of disagreement and stance in online debate. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 116–125, 2015. (Cited on pages 17 and 81.)

Kate Starbird. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *International AAAI Conference on Web and Social Media*, 2017. (Cited on page 14.)

Tao Stein, Erdong Chen, and Karan Mangla. Facebook immune system. In *Proceedings of the 4th Workshop on Social Network Systems*, SNS '11, pages 8:1–8:8, 2011. (Cited on page 14.)

V. S. Subrahmanian, Amos Azaria, Skylar Durst, Vadim Kagan, Aram Galstyan, Kristina Lerman, Linhong Zhu, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. The darpa twitter bot challenge. *Computer*, 49(6):38–46, 2016. (Cited on page 14.)

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 697–706, 2007. (Cited on page 3.)

Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. Stance detection with hierarchical attention network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2399–2409, 2018. (Cited on page 81.)

Behzad Tabibian, Isabel Valera, Mehrdad Farajtabar, Le Song, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. Distilling information reliability and source trustworthiness from digital traces. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 847–855, 2017. (Cited on page 19.)

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Comput. Linguist.*, 37(2):267–307, 2011. (Cited on page 17.)

Rudra M. Tripathy, Amitabha Bagchi, and Sameep Mehta. A study of rumor control strategies on social networks. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1817–1820, 2010. (Cited on page 14.)

Peter D. Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, 2002. (Cited on page 17.)

Onur Varol, Emilio Ferrara, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017.*, pages 280–289, 2017. (Cited on page 14.)

Michela Del Vicario, Gianna Vivaldo, Alessandro Bessi, Fabiana Zollo, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Echo chambers: Emotional contagion and group polarization on facebook. *CoRR*, abs/1607.01032, 2016. (Cited on page 15.)

V. G. Vinod Vydiswaran, Chengxiang Zhai, and Dan Roth. Gauging the internet doctor: Ranking medical claims based on community knowledge. In *Workshop on Data Mining for Medicine and HealthCare, DMMH'11 - Held with the KDD Conference, the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD-2011*, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 42–51, 2011. (Cited on page 12.)

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 647–653, 2017. (Cited on page 13.)

Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, 2014. (Cited on page 3.)

V. G.Vinod Vydiswaran, Chengxiang Zhai, Dan Roth, and Peter Pirolli. Biastrust: Teaching biased users about controversial topics. In *CIKM 2012 - Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 1905–1909, 2012. (Cited on page 12.)

V.G. Vinod Vydiswaran, ChengXiang Zhai, and Dan Roth. Content-driven trust propagation framework. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 974–982, 2011. (Cited on page 19.)

Marilyn A. Walker, Pranav Anand, Robert Abbott, and Ricky Grant. Stance classification using dialogic properties of persuasion. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 592–596, 2012. (Cited on pages 17 and 81.)

Fulton Wang and Cynthia Rudin. Falling rule lists. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12, 2015*, 2015. (Cited on page 19.)

Guan Wang, Sihong Xie, Bing Liu, and Philip S. Yu. Review graph based online store review spammer detection. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*, ICDM '11, pages 1242–1247, 2011. (Cited on page 16.)

William Yang Wang. “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, 2017. (Cited on pages 12, 60, 63, and 68.)

Yao Wang and Julita Vassileva. Trust and reputation model in peer-to-peer networks. In *Proceedings of the 3rd International Conference on Peer-to-Peer Computing*, P2P ’03, pages 150–, 2003. (Cited on page 19.)

Janyce Wiebe and Ellen Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing’05, pages 486–497, 2005. (Cited on page 17.)

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. Learning subjective language. *Comput. Linguist.*, 30(3):277–308, 2004. (Cited on page 17.)

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. Towards universal paraphrastic sentence embeddings. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. (Cited on page 63.)

A. G. Wilson, J. Yosinski, P. Simard, R. Caruana, and W. Herlands. Proceedings of NIPS 2017 Symposium on Interpretable Machine Learning. *arXiv e-prints*, 2017. (Cited on page 19.)

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT ’05, pages 347–354, 2005. (Cited on page 86.)

K. Wu, S. Yang, and K. Q. Zhu. False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st International Conference on Data Engineering*, pages 651–662, 2015. (Cited on page 13.)

Sihong Xie, Guan Wang, Shuyang Lin, and Philip S. Yu. Review spam detection via temporal pattern discovery. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’12, pages 823–831, 2012. (Cited on page 16.)

Qiongkai Xu and Hai Zhao. Using deep linguistic features for finding deceptive opinion spam. In *Proceedings of COLING 2012: Posters*, pages 1341–1350, 2012. (Cited on page 16.)

Fan Yang, Yang Liu, Xiaohui Yu, and Min Yang. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, MDS '12, pages 13:1–13:7, 2012. (Cited on page 13.)

Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 412–420, 1997. (Cited on page 19.)

Zhi Yang, Christo Wilson, Xiao Wang, Tingting Gao, Ben Y. Zhao, and Yafei Dai. Uncovering social network sybils in the wild. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, IMC '11, pages 259–268, 2011. (Cited on page 16.)

Andrew Yates, Nazli Goharian, and Ophir Frieder. Extracting adverse drug reactions from social media. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2460–2467, 2015. (Cited on page 13.)

Xiaoxin Yin, Jiawei Han, and Philip S. Yu. Truth discovery with multiple conflicting information providers on the web. *IEEE Trans. on Knowl. and Data Eng.*, 20(6): 796–808, 2008. (Cited on pages 11, 12, 22, 36, 50, and 51.)

Qiang Zhang, Emine Yilmaz, and Shangsong Liang. Ranking-based method for news stance detection. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 41–42, 2018. (Cited on page 18.)

Bo Zhao, Benjamin I. P. Rubinstein, Jim Gemmell, and Jiawei Han. A bayesian approach to discovering truth from conflicting sources for data integration. *Proc. VLDB Endow.*, 5(6):550–561, 2012. (Cited on pages 11 and 12.)

Zhe Zhao, Paul Resnick, and Qiaozhu Mei. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 1395–1405, 2015. (Cited on page 13.)

Shi Zhi, Bo Zhao, Wenzhu Tong, Jing Gao, Dian Yu, Heng Ji, and Jiawei Han. Modeling truth existence in truth discovery. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1543–1552, 2015. (Cited on page 11.)

Shi Zhi, Yicheng Sun, Jiayi Liu, Chao Zhang, and Jiawei Han. Claimverif: A real-time claim verification system using the web and fact databases. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 2555–2558, 2017. (Cited on page 73.)

Arkaitz Zubiaga, Geraldine Wong Sak Hoi, Maria Liakata, Rob Procter, and Peter Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS ONE*, 11(3), 2016. (Cited on page 48.)

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.*, 51(2):32:1–32:36, 2018. (Cited on page 6.)