KTH ROYAL INSTITUTE
OF TECHNOLOGY

Doctoral Thesis in Information and Communication Technology

Toward automated veracity assessment of data from open sources using features and indicators

MARIANELA GARCIA LOZANO

Stockholm, Sweden 2024

Toward automated veracity assessment of data from open sources using features and indicators

MARIANELA GARCIA LOZANO

Academic Dissertation which, with due permission of the KTH Royal Institute of Technology, is submitted for public defence for the Degree of Doctor of Philosophy on Monday the 3rd of June 2024, at 1:30 p.m. in sal C, Kistagången 16.

Doctoral Thesis in Information and Communication Technology
KTH Royal Institute of Technology
Stockholm, Sweden 2024

ISBN 978-91-8040-927-8
TRITA-EECS-AVL-2024:47

Printed by: Universitetsservice US-AB, Sweden 2024

To all cats suffering from chronic curiosity, Ada, Miranda, and Edward.

Abstract

This dissertation hypothesizes that the key to automated veracity assessment of data from open sources is the careful estimation and extraction of relevant features and indicators. These features and indicators provide added value to a quantifiable veracity assessment, either directly or indirectly. The importance and usefulness of a veracity assessment largely depend on the specific situation and reason for which it is being conducted. Factors such as the recipient of the veracity assessment, the scope of the assessment, and the metrics used to measure accuracy and performance, all play a role in determining the value and perceived quality of the assessment.

Five peer-reviewed publications; two journal articles, two conference articles, and one workshop article, are included in this compilation thesis.

The main contributions of the work presented in this dissertation are: i) a compilation of challenges with manual methods of veracity assessment, ii) a road map for addressing the identified challenges, iii) identification of the state-of-the-art and gap analysis of veracity assessment of open-source data, iv) exploration of indicators such as topic geo-location tracking over time and stance classification, and v) evaluation of various feature types, model transferability, and style obfuscation attacks and the impact on accuracy for automated veracity assessment of a type of deception: fake reviews.

Sammanfattning

Denna avhandling har som hypotes att nyckeln till automatiserad trovärdighetsbedömning av data från öppna källor ligger i det noggranna urvalet och estimeringen av relevanta särdrag och indikatorer. Dessa särdrag och indikatorer ger ett direkt eller indirekt mervärde till en kvantifierbar trovärdighetsbedömning. Betydelsen och användbarheten av en trovärdighetsbedömning beror till stor del på den specifika kontexten och anledningen till att den genomförs. Faktorer som mottagaren av trovärdighetsbedömningen, omfattningen av bedömningen och de mått som används för att mäta noggrannhet och prestanda, spelar alla in för att bestämma värdet och den upplevda kvalitén på bedömningen.

Fem referentgranskade publikationer ingår i denna sammanläggningsavhandling; två tidskriftsartiklar, två konferensartiklar och en workshopartikel.

De huvudsakliga bidragen från arbetet som presenteras i denna avhandling är: i) en sammanställning av utmaningar relaterade till manuella metoder för trovärdighetsbedömning, ii) en plan för att ta itu med de identifierade utmaningarna, iii) identifiering av forskningsfronten och en gapanalys av trovärdighetsbedömning av data från öppna källor, iv) studie av indikatorer såsom geolokalisering av ämnen och spårning av dem över tid samt klassificering av individers reaktioner i inlägg på sociala medier, och v) en utvärdering av särdragstyper som påverkar noggrannheten för automatisk trovärdighetsbedömning applicerat på en typ av bedrägeri: falska recensioner.

Acknowledgements

This Odyssean journey would not have been possible without my fantastic navigators, co-captains, and shipmates. Over the years, you have patiently supported and steered me right. I will forever be deeply grateful that you kept cheering me on and did not allow me to lose hope, even when I felt the final destination was a long way off. I raise my glass to you!

Vladimir Vlassov, Anne Håkansson, Joel Brynielsson, Magnus Rosell, Ulrik Franke, Edward Tjörnhammar, Stefan Varga, Johan Fernquist, Hanna Lilja, Maja Karasalo, Jonah Schreiber, Rassul Ayani, Christian Mårtensson, and Pontus Svenson

I want to express my appreciation for my awesome colleagues, past and present, at FOI and KTH. You always manage to provide wise words of encouragement, much-needed distractions, and laughter, making my life brighter and easier. You are the ones who make everything worthwhile. I wish you all marvelous adventures in your lives.

Farshad Moradi, Farzad Kamrani, Tove Gustavi, Johan Sabel, and Katharina Rasch

My amazing and caring family and friends have shown nothing but the greatest love and support, and for that, I will always remember and cherish the kindness that you have shown me. You are forever in my heart and have my eternal gratitude. May love, happiness, and inspiration be your constant companions!

– ***Marianela***

Acronyms

AI	Artificial intelligence
AMT	Amazon mechanical turk
BERT	Bidirectional encoder representations from transformers
Bi-LSTM	Bidirectional long short-term memory
BoW	Bag of words
CNN	Convolutional neural network
CRF	Conditional random fields
CSI	CLiPS stylometry investigation corpus
DAL	Dictionary of affect in language
DL	Deep learning
GAN	Generative adversarial networks
GDPR	General data protection regulation
GloVe	Global vectors for word representation
GPT	Generative pre-trained transformer
GRNN	General regression neural network
GRU	Gated recurrent unit
LDA	Latent Dirichlet allocation
LIWC	Linguistic inquiry and word count
LLM	Large language model
LSTM	Long short-term memory
ML	Machine learning
MM	Multimodal model
NATO	North atlantic treaty organization
NLP	Natural language processing
NLU	Natural language understanding
PCA	Principal component analysis
PCFG	Probabilistic context-free grammar
POS	Part of speech
RDF	Resource description framework
RNN	Recurrent neural network
SDQC	Support, deny, query, comment
SLDA	Streaming latent Dirichlet allocation
STANAG	Standardization agreements (NATO)
STO	NATO's science and technology organization
SVM	Support vector machine
TF – IDF	Term frequency — Inverse document frequency
UNESCO	United nations educational, scientific and cultural organization

Contents

Part I

Overview

Chapter 1

Introduction

> ... if a Lie be believ'd only for an Hour, it has done its Work, and there is no farther occasion for it. Falsehood flies, and the Truth comes limping after it ...
>
> *Jonathan Swift (1710)*

Today, internet-based platforms, such as social media and digital news, are a primary source of information for many people [82, 193]. When the Coronavirus pandemic started in 2019, online news consumption increased, and, e.g., in 2020, on average, 35% used Facebook[1] to find, discuss, and share information about the virus [155]. However, in the report by Newman *et al.* from 2023 [154] more than half, i.e., 56%, of over 90,000 online news consumers,[2] said they were concerned about what is real and what is fake in the internet when it comes to news.

The challenge of using *open-source data*, i.e., all data that is publicly available, is that it often comes with the price of *noise*. This noise results from much data being irrelevant, ambiguous, contradicting, biased, or plain wrong [21], and it leads to uncertainty in the data's veracity. As described by García Lozano *et al.*, veracity is a term that can often be used interchangeably with words such as truth, trustworthiness, uncertainty, credibility, and quality [68]. The noise can be the effect of unintentional typos and mistakes or due to deliberate introductions of fabricated, subjective, or false data. Another challenge with this noisy data is that false data, such as false news, often generates a higher degree of engagement than real news [197]. In addition, technologies that have become highly accessible, such

[1] In the UK, US, Germany, Argentina, South Korea, and Spain.

[2] Total sample of respondents in Africa = 6063, North America = 4231, Latin America = 12,149, Asia-Pacific = 22,477, and Europe = 48,975 [154].

as deep learning[3] and, in particular, generative adversarial networks,[4] transformers,[5] and attention,[6] make it much easier to fabricate media such as text, pictures, videos, and sound [104, 204, 222].

Despite the presence of noise, leveraging the vast amount of data available from open sources can be highly beneficial [22]. For example, collecting open-source data is generally less expensive and less risky than obtaining it from other sources. Additionally, it provides insights into developments that may not be a priority for other systems or may not be accessible through other approaches. These developments may include innovative applications of new technologies, shifts in popular attitudes, emergence of new ideological movements, trends, disillusionment with leadership, discontent, and more. In order to recognize, quantify, and handle noisy data, one needs to assess its veracity.

In the case of data from open sources, manual, i.e., done by a person, veracity assessment is virtually unfeasible considering the staggering amounts of data continuously being generated by numerous and unknown sources. Hence, there is a need for *automated veracity assessment* support in the form of approaches, methods, algorithms, and tools.

1.1 Thesis statement

The statement of this dissertation is as follows:

> *The careful extraction, estimation, and use of relevant features and indicators are key to automating and quantifying veracity assessment of open-source data.*

[3] Deep learning (DL) is a branch of machine learning (ML) where the methods are based on neural networks and feature learning (representational learning). Neural networks, inspired by the brain's function [138], use interconnected nodes (neurons) in a layered structure to process data. A neuron is a computational unit with one or more weighted inputs, a function that combines the inputs, and an output. Feature learning algorithms automatically analyze the raw data and learn significant features or representations for a specific task, replacing the more traditional manual feature engineering task.

[4] Generative adversarial networks (GAN) [80] consists of two competing neural networks. The first (generative) model captures the data distribution of a training set and is trained to generate data that cannot be distinguished from the original dataset. While the second (discriminator) model is trained to differentiate between generated data and original data by estimating the probability that a data sample is generated.

[5] Transformers are available in a large number of variants but are in their most basic form a sequence-to-sequence [144, 204] neural network which often consists of an encoder and a decoder. The encoder transforms input data, such as text or sound, into numerical representations and processes the representations iteratively, one layer after another, mixing information from other input representations via some attention method. The decoder consists of decoding layers that iteratively process the encoder's output with a cross-attention to obtain contextualized input representations, as well as an attention function to focus on the decoder's own output so far.

[6] The attention method [222] is a technique to determine the importance (weight) of different tokens in a sentence. This is done by calculating weights for each word's numerical representation (embedding) within a specific section of a sentence, known as the context window.

The term feature is related to the term indicator. A feature is often used in the sense of structure, form, or characteristic found in the data to be analyzed, e.g., the account name spreading a statement, the time of publication, publication language, or included URLs. Using machine learning terminology, a feature is an individually measurable property or characteristic of an observed phenomenon, preferably independent of other features found in the data [23].

An indicator refers to variables that are observed or induced from the data and their relationship to the outside world. Indicators are an indirect way of obtaining information and can be used to assess veracity in some context. For example, suppose that a majority of social media users challenge and question a statement spread on social media. That questioning can indicate that the statement's truthfulness is in doubt, and the veracity assessment becomes low.

The term *assessment* is used for making a judgment or appraisal about something. When it is used in conjunction with veracity, the combined term *veracity assessment* means a judgment of an aspect of interest regarding the data, its source, or its context.

Assessing veracity is a multifaceted task that requires varying input, such as data, facts, indicators, and features. Facts are irrefutable statements that are verifiable by experiments, measurements, and events that have happened or can be found in standard reference sources. A quantifiable veracity assessment is a measurable value with a degree of confidence.

The extracted and estimated features and indicators directly or indirectly convey added value to a quantifiable veracity assessment. The value and perceived quality, including accuracy and performance, for the recipient of a veracity assessment, depends on the context and purpose with which veracity is assessed.

1.2 Research questions

The thesis statement argues that the key to automated veracity assessment of open-source data lies in extracting, estimating, and using relevant features and indicators. The value and quality of automated veracity assessment depend on the context and purpose of the assessment, which guides selected assessment approaches, methods, algorithms, indicators, and features. Hence, the addressed research questions of this dissertation are as follows:

R1 – *Which assessment approaches, methods, research challenges, and gaps are present or have been identified in manual veracity assessment of open-source data?* The main idea is to investigate the open issues in the manual veracity assessment process and consider whether they may impact veracity assessment automation efforts. Research challenges, gaps, and automation propositions are identified based on an analysis of manual veracity assessment approaches and methods of open-source data.

R2 – *How can the veracity assessment of data from open sources be automated in part or in whole?* As manual veracity assessment of the large volumes and varieties of open-source data, e.g., various modalities, structured or unstructured, is unfeasible, support in the form of automated approaches, methods, algorithms, and tools is needed.

R3 – *What features and indicators can be extracted, estimated, and used to automate veracity assessment of open-source data?* The thesis statement argues that extracting, estimating, and utilizing relevant features and indicators is the key to automating and quantifying veracity assessment of open-source data. Since veracity can be assessed in various ways depending on the underlying data and the context and purpose of the assessment, the research question aims to identify methods, algorithms, and tools for extracting and estimating features and indicators.

R4 – *To what degree do specific features and indicators work for veracity assessment automation?* The veracity assessment process can be automated in part or as a whole. Automation can involve automating the processing, extraction, estimation, and analysis of individual features and indicators. Veracity assessment of rumors, trends, sentiments, and subjective topics is of interest because there are usually no hard facts available to compare statements to at the time of publication. Therefore, approaches and methods other than fact-checking need to be used to assess veracity.

1.3 Publications

The thesis includes the following peer-reviewed publications; two journal articles, two conference articles, and one workshop article. An extended summary of each publication is presented in Chapter 3. The publications are also included in their entirety in Part II. The author contributions are presented in Section 1.4. Each publication's contribution to the dissertation and research questions is presented in Section 1.5. The publications are presented in logical order rather than chronological.

P1 – M. García Lozano, U. Franke, M. Rosell, and V. Vlassov, "Towards automatic veracity assessment of open source information," in *2015 IEEE International Congress on Big Data*, IEEE, 2015, pp. 199–206. DOI: 10.1109/BigDataCongress.2015.36

P2 – M. García Lozano, J. Brynielsson, U. Franke, M. Rosell, E. Tjörnhammar, S. Varga, and V. Vlassov, "Veracity assessment of online data," *Decision Support Systems*, vol. 129, no. 113132, 2020. DOI: 10.1016/j.dss.2019.113132

P3 – M. García Lozano, J. Schreiber, and J. Brynielsson, "Tracking geographical locations using a geo-aware topic model for analyzing social media data,"

Decision Support Systems, vol. 99, pp. 18–29, 2017. DOI: 10.1016/j.dss.2017.05.006

P4 – M. García Lozano, H. Lilja, E. Tjörnhammar, and M. Karasalo, "Mama Edha at SemEval-2017 Task 8: Stance classification with CNN and rules," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 481–485. DOI: 10.18653/v1/S17-2084

P5 – M. García Lozano and J. Fernquist, "Identifying deceptive reviews: Feature exploration, model transferability and classification attack," in *2019 European Intelligence and Security Informatics Conference (EISIC)*, IEEE, Nov. 2019, pp. 228–228. DOI: 10.1109/EISIC49498.2019.9108852

1.3.1 Additional peer-reviewed publications

In addition to the publications included in the dissertation, the following selection of peer-reviewed publications, presented in reverse chronological order, are not included in the dissertation but have had an impact on it:

P6 – P. Hansen, M. García Lozano, F. Kamrani, and J. Brynielsson, "Real-time estimation of heart rate in situations characterized by dynamic illumination using remote photoplethysmography," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6094–6103. DOI: 10.1109/CVPRW59228.2023.00649

P7 – M. García Lozano, "Trusting open source information," in *2013 European Intelligence and Security Informatics Conference (EISIC)*, IEEE, 2013, pp. 228–228. DOI: 10.1109/EISIC.2013.77

P8 – M. García Lozano, "Semantic based resource identification, storage and discovery in distributed systems," Licentiate thesis, KTH, 2010. [Online]. Available: https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-27340

P9 – M. García Lozano, F. Moradi, E. Ibarzabal, and E. Tjörnhammar, "A semantic approach to simulation component identification and discovery," in *EMSS 2009, 21st European Modeling and Simulation Symposium*, Sep. 2009, pp. 181–186

P10 – M. García Lozano, P. Hörling, F. Moradi, and E. Tjörnhammar, "Supporting C2 with a service oriented framework for opportunistic sensors and sensor networks," in *International Command and Control Research and Technology Symposium 14th ICCRTS:"C2 and Agility"*, 2009

P11 – M. García Lozano, F. Moradi, and R. Ayani, "SDR: A semantic based distributed repository for simulation models and resources," in *First Asia*

International Conference on Modelling Simulation (AMS'07), 2007, pp. 171–176

P12 – M. Chenine, V. Kabilan, and M. García Lozano, "A pattern for designing distributed heterogeneous ontologies for facilitating application interoperability," in *International Conference on Advanced Information Systems Engineering*, CEUR-WS.org, 2006

1.3.2 Technical reports

In addition to the publications included in the dissertation and additional peer-reviewed publications, the following selection of technical reports, presented in reverse chronological order, are not included in the dissertation but have had an impact on it:

P13 – F. Johansson, A. Horndahl, H. Lilja, M. García Lozano, L. Lundmark, and H. Stiff, "Detection of fabricated media," Swedish Defence Research Agency, Stockholm, Sweden, Tech. Rep. FOI-R--5132--SE, Apr. 2021

P14 – F. Johansson, A. Horndahl, H. Stiff, and M. García Lozano, "Data synthesis using generative models," Swedish Defence Research Agency, Stockholm, Sweden, Tech. Rep. FOI-R--5041--SE, Nov. 2020

1.4 Author contributions

The dissertation author is the main author and editor in all of the publications. She also handled most of the communication with the reviewers.

The dissertation author has done most of the work in the first publication (P1 – Towards automatic veracity assessment of open source information), including the literature search, questionnaire, interview design, implementation, analysis, discussion with the reviewers, and writing. The co-authors collaborated in the development of the veracity assessment framework suggested in the paper.

In the second publication (P2 – Veracity assessment of online data), the dissertation author has carried out most of the work, including the design of the search questions, inclusion and exclusion criteria, filtering of database search hits, review protocol, result evaluation, and gap analysis. The co-authors shared the research load in the paper review process, the review protocol results, and the evaluation.

In discussions with others, the dissertation author proposed and refined the main ideas behind publication number three (P3 – Tracking geographical locations using a geo-aware topic model for analyzing social media data). The implementation was done by a master's thesis student (also one of the co-authors), whom the dissertation author also supervised. The experiment design, setup, and analysis were done in collaboration with the master student.

In the fourth publication (P4 – Mama Edha at SemEval-2017 Task 8: Stance classification with CNN and rules), the dissertation author was the main driving

force in the research. She implemented the convolutional neural network (CNN)[7] and the ensemble method, did the experiments, analyzed, and evaluated the results. The co-authors implemented other parts which were part of the ensemble method.

In the last publication included in the dissertation (P5 – Identifying deceptive reviews: Feature exploration, model transferability, and classification attack), the dissertation author conceived the main ideas and research design. The experiment design, setup, and analysis were done in collaboration with the co-author, a master's student supervised by the dissertation author. The co-author executed the implementation and experiment.

1.5 Dissertation contributions

The research presented in this dissertation aims to enable automated veracity assessment of open-source data by exploring approaches, methods, algorithms, and tools. The focus lies in using, extracting, and estimating features and indicators, which is an essential part of automating veracity assessment. The main contributions are as follows:

C1 – *A compilation of research challenges and gaps regarding the current manual approaches and methods for veracity assessment within the military domain.*

In the first publication (P1 – Towards automatic veracity assessment of open source information), a study was conducted that examined which methods analysts in the military domain use and how they employ them to manually assess the veracity of open-source data. Several one-on-one interviews with analysts were conducted, and participants at a military exercise were asked to answer a questionnaire. During the study, various challenges that analysts face, such as contradictory definitions, lack of benchmarks, difficulties with follow-up assessments, and the use of assessment scales that are not adapted to data from open sources, were identified. The study addresses the first research question (R1 – *Which assessment approaches, methods, research challenges, and gaps are present or have been identified in manual veracity assessment of open-source data?*).

C2 – *A roadmap for working with the identified challenges with the manual veracity assessment challenges and gaps.*

In the first publication (P1 – Towards automatic veracity assessment of open source information), a roadmap for manual veracity assessment research questions was presented, resulting from a qualitative study combined with a literature review. This work addresses the first and second research questions (R1 – *Which assessment approaches, methods, research challenges, and gaps*

[7] A Convolutional Neural Network (CNN) is a type of neural network designed to learn spatial features found in images. It is primarily used for image classification and segmentation tasks.

are present or have been identified in manual veracity assessment of open-source data? and R2 – *How can the veracity assessment of data from open sources be automated in part or in whole?*).

C3 – *A literature review identifying the state-of-the-art approaches, methods, algorithms, and tools for automated veracity assessment of online data.*

In the work presented in the second publication (P2 – Veracity assessment of online data), an extensive literature review identifying state-of-the-art veracity assessment of online data was conducted, and the first, second, and third research questions were addressed (R1 – *Which assessment approaches, methods, research challenges, and gaps are present or have been identified in manual veracity assessment of open-source data?* R2 – *How can the veracity assessment of data from open sources be automated in part or in whole?* and R3 – *What features and indicators can be extracted, estimated, and used to automate veracity assessment of open-source data?*).

C4 – *An identification of main veracity assessment research directions and gap analysis for automated veracity assessment of online data.*

In the work described in the second publication (P2 – Veracity assessment of online data), the main veracity assessment research directions for automating the process were identified, and a gap analysis was performed, addressing the first, second, and third research questions (R1 – *Which assessment approaches, methods, research challenges, and gaps are present or have been identified in manual veracity assessment of open-source data?* R2 – *How can the veracity assessment of data from open sources be automated in part or in whole?* and R3 – *What features and indicators can be extracted, estimated, and used to automate veracity assessment of open-source data?*).

C5 – *Implementation and exploration of an algorithm for tracking topic geolocation over time, an indicator useful for automated veracity assessment of streaming micro social media.*

In the work described in the third publication (P3 – Tracking geographical locations using a geo-aware topic model for analyzing social media data), the indicator geolocation was explored with a focus on how it can be tracked over time and related to specific topics. This addressed the second and third research questions (R2 – *How can the veracity assessment of data from open sources be automated in part or in whole?* and R3 – *What features and indicators can be extracted, estimated, and used to automate veracity assessment of open-source data?*).

C6 – *Implementation and study of a stance classification ensemble method, an indicator useful for veracity assessment of rumors.*

In the fourth publication (P4 – Mama Edha at SemEval-2017 Task 8: Stance classification with CNN and rules), another indicator, i.e., stance classifica-

tion, was studied, addressing the second and third research questions (R2 – *How can the veracity assessment of data from open sources be automated in part or in whole?* and R3 – *What features and indicators can be extracted, estimated, and used to automate veracity assessment of open-source data?*).

C7 – *Automatic veracity assessment of a type of deception, i.e., fake reviews, and evaluated the accuracy of different features for automated veracity assessment*

In the fifth publication (P5 – Identifying deceptive reviews: Feature exploration, model transferability, and classification attack), fake reviews, a type of deception, was studied, addressing the fourth research question (R4 – *To what degree do specific features and indicators work for veracity assessment automation?*).

1.6 Research method and approach

A combination of *exploratory* and *empirical* research has been used for this dissertation.

The exploratory method was predominantly used in two different research approaches. Firstly, it was used in the qualitative research approach for the survey presented in the first publication (P1 – Towards automatic veracity assessment of open source information). Furthermore, it has also been used in the secondary research approach for the systematic literature review presented in the second publication (P2 – Veracity assessment of online data).

The *empirical* method, where both *qualitative* and *quantitative* approaches were employed, has primarily been used for the work presented in the third, fourth, and fifth publications (P3 – Tracking geographical locations using a geo-aware topic model for analyzing social media data, P4 – Mama Edha at SemEval-2017 Task 8: Stance classification with CNN and rules, and P5 – Identifying deceptive reviews: Feature exploration, model transferability, and classification attack).

1.7 Outline

The dissertation is divided into two main parts, where Part I consists of a comprehensive summary (*sv. kappa*) and Part II of the full constituent dissertation papers.

The rest of this part, i.e., Part I, is organized as follows: Chapter 2 provides the background and contextualizes the dissertation research. In Chapter 3, a summary of each publication is provided, and their contributions are clarified. Part I comes to a close with Chapter 4, which concludes the work, and addresses legal and ethical considerations, and limitations. The chapter ends with a discussion on current veracity challenges, along with a discussion and proposals for future work.

Part II contains the complete prints of this dissertation's peer-reviewed journal and conference papers.

Chapter 2

Background and related work

> . . . how can you have an opinion if you are not informed? If everybody always lies to you, the consequence is not that you believe the lies, but rather that nobody believes anything any longer.
>
> *Hanna Arendt (1974)*

This chapter explores the background of veracity assessment and discusses key terminology, approaches, and methodological challenges from an open-source data assessment perspective. It also includes related research in the context of the dissertation papers.

2.1 Veracity and related terminology

In 2012, the veracity concept was proposed as the fourth complementary big data "V" [39, 187, 200]. The three original big data Vs are volume, variety, and velocity [119]. Snow, a long-time employee of IBM, argued in talks and blog posts [200] that trusted data needed to be defined separately due to the easy access to large volumes of heterogeneous data. In the same blog post, Snow also argued that the definition of trusted data was context-dependent and built on the way the data was used. Thus, veracity was introduced as a concept that deals with "uncertain or imprecise data."

Even though there is no unified definition of the veracity concept, some researchers have presented descriptions, definitions, and data veracity frameworks. García Lozano *et al.* compiled an overview of dictionary veracity definitions, see Figure 2.1, where aspects of accuracy, credibility, truthfulness, and quality, outline the dimensions of the veracity term [68]. Some researchers' veracity term proposals

align with the dictionary definitions of Figure 2.1, while others, like Schroeck *et al.*, describe veracity as "data uncertainty," arguing that some data is inherently unreliable and uncertain and that this can be managed by fusing less reliable data to create a more accurate data point [187]. In a report from 2012, IBM state that veracity is about managing "data in doubt" and relate it to "uncertainty due to data inconsistency, incompleteness, ambiguities, and deception" [39]. Debattista *et al.* also argue that veracity deals with data uncertainty due to various factors such as inconsistencies, incompleteness, and deliberate deception [45].

Online dictionary definitions of the term veracity

- Cambridge.org https://dictionary.cambridge.org
  the quality of being true, honest, or accurate
- Merriam-webster.com https://www.merriam-webster.com
  i) conformity with truth or fact; accuracy,
  ii) devotion to the truth; truthfulness,
  iii) power of conveying or perceiving truth,
  iv) something true.
- Dictionary.com https://www.dictionary.com
  i) habitual observance of truth in speech or statement; truthfulness,
  ii) conformity to truth or fact; accuracy,
  iii) correctness or accuracy, as of the senses or of a scientific instrument.
- Vocabulary.com https://www.vocabulary.com
  unwillingness to tell lies

Figure 2.1: Based on a figure by García Lozano *et al.* [68], the text in the figure examines the different meanings of the term "veracity" as defined by various English online dictionaries.

Another group of researchers, such as McArdle and Kitchin, define data veracity more in line with the dictionary definitions and with terms of authenticity, precision, and reliability of collected data [137]. Agarwal *et al.* equate veracity with "data correctness" [3]. Ramachandramurthy *et al.* state that veracity "focuses on information quality" and that "uncertainty in the data leads to unreliable information" [177]. Similarly, Pendyala states that veracity "refers to the quality of data" [165]. While Jamil *et al.* define data veracity in terms of "truth and omission" [99].

In the book by Berti-Équille and Borge-Holthoefer, veracity is discussed and analyzed across a wide range of perspectives from a truth discovery perspective to

a misinformation dynamics model perspective [18]. McArdle and Kitchin debate that "it is difficult to express data veracity across an exhaustive range of domains" since "data is often provided without knowledge of all the possible end uses" [137]. From Berti-Équille and Borge-Holthoefer's, McArdle and Kitchin's, and García Lozano *et al.*'s discussion, analysis, and studies, the conclusion in this dissertation is that veracity is context-dependent and the veracity term definition depends on the data and the perspective of the end-user.

In the veracity term discussion, one should not forget that many of the veracity challenges that researchers working within the Big data field brought up in the early 2010s were not new. Dealing with low-quality, inconsistent, incomplete, uncertain, untruthful, or ambiguous data has also been studied in related research settings such as decision support and information systems [2, 64, 134, 159, 230].

As has been discussed in this text, there are several terms related to veracity and data from open sources, such as trust, quality, disinformation, misinformation, fake news, fake reviews, and rumors, which also deserve some examination. In the following subsections, these terms are explored in more detail. To generalize, the first term, i.e., trust, is mainly related to the data sources and users. The following terms, i.e., quality, disinformation, misinformation, fake news, and fake reviews, are primarily related to the data contents. The last term, i.e., rumor, is associated with the data propagation mechanisms.

2.1.1 Trust

In his PhD thesis, Marsh strived to formalize trust as a computational concept that could be included in artificial agent behavior for decision-making and treated trust as an inverse measure of uncertainty in a person or resource [134]. Golbeck and Hendler used a reputation-based trust measure to assess the degree to which a node in a social network can be trusted [79]. The reputation value is based on the average value a node's neighbors assign it. The value can be assigned on a general level or based on a certain topic. Topical trust was also used by, e.g., Knap and Mlỳnková to identify topic networks in a friend of a friend network [113]. McPherson *et al.* noticed that similarity breeds connection, i.e., one tends to trust similar things, and that homophily could be used to assess trust in network analysis [139].

A challenge in assessing the veracity of open-source data is verifying it when there are disagreeing views or contradictory statements. To handle this, some researchers have tried using *trust* as a measure. Galland *et al.* use trust and corroborating sources to handle disagreeing views [63]. They demonstrate that a simple baseline voting algorithm, corresponding to choosing the assessment of the majority about a fact, works relatively well to differentiate between facts. When they add parameters to the trustworthiness of views, with an assessment of the truth of facts and an estimation of the errors, they achieve an improvement in the precision of the results.

Jøsang *et al.* proposed differentiating between different types of trust, e.g., reliability trust, decision trust, and reputation [106]. The difference between the trust

types influences how the information is assessed, i.e., whether the sources and other corroborating or disagreeing entities can be trusted or relied upon.

Another type of trust is the transitive trust, i.e., A trusts B, B trusts C; thus, A can also trust C. Kuter and Golbeck and Avesani *et al.* used this in social networks to develop trust metrics for transitive trust relationships [13, 118]. This kind of trust metric is also related to the centrality trust measure often used in social networks where accounts that are more retweeted, mentioned, or linked to are viewed as more trusted. Weng *et al.* used the centrality measure to develop trust measures to rank influential Twitter accounts on different topics [233]. Ulicny and Kokar converted Twitter streams to RDF[1] graphs and then designed an eigenvector centrality measure called TunkRank and mapped it to the STANAG[2] 2511 reliability metric [216].

2.1.2 Quality

The concepts of data and information quality are also used in the connotation of veracity. The most widely used definition is the one proposed by Juran in 1974, where quality is defined as "fitness for use" [108]. There are two main implications of this definition. The first is that quality depends on the task. Hence, quality is context- and purpose-dependent, and a user may consider information appropriate for one task but not for another. The second is that the quality of the data is subjective, as users may perceive the quality of the same data differently.

To capture the aspects of data quality that are important to data consumers, Wang and Strong conducted a two-stage survey that resulted in a set of dimensions of data quality, i.e., intrinsic data quality, contextual data quality, representational data quality, and accessibility data quality [230]. These dimensions are, in turn, comprised of a number of attributes, e.g., believability, objectivity, relevance, timeliness, interpretability, and accuracy.

Gil and Ratnakar designed a system called TRELLIS where intelligence analysts could get help selecting quality data within a military setting [77]. The analysts made decisions on which sources were trustworthy and which were not, and their decisions were described in the system. This information was then stored in an RDF graph and could be used to derive an assessment of the source based on the annotations of many individuals. The idea behind the system was that a trusted source also would provide qualitative, i.e., credible data. Statements in the system could also have a subjective "likelihood qualification," which denoted an informal indication of the analyst's reaction to a statement, e.g., surprise, dismissal, saliency, and accuracy.

[1] The resource description framework (RDF) is a World wide web consortium (W3C) standard that is used to represent information on the web by modeling relationships between nodes https://www.w3.org/TR/turtle/.

[2] A STANAG is a standardization agreement between NATO members and according to what the agreement is about they are assigned different numbers.

Viewing veracity assessment from the perspective of the recommender, García Lozano *et al.* designed a framework where trust between individuals in a group, such as intelligence analysts, is used to recommend other data along the lines of the idea "analysts who viewed these data also viewed..." [70].

Bizer and Cyganiak created a quality-aware web-based system called WIQA to filter unwanted or low-quality data [24]. The filtering is done by using different policies expressed as graph patterns and filter conditions based on meta-data, e.g., provenance chains[3] and background information about data sources.

2.1.3 Disinformation and misinformation

Access to open-source data has made the internet a significant source of information, and social media is now a primary method of news consumption [49, 91]. However, the escalating pervasiveness of ambiguities, half-truths, and falsehoods is an increasing problem [28]. In today's world, false data can take many different forms. It can be artificially generated text, manipulated images, deepfake videos, or even false personas, accounts, and organizations. This plethora of methods to create and disseminate false data makes it incredibly difficult to distinguish between authentic and trustworthy data and dishonest and false data.

Two terms are commonly used when discussing falsehoods, i.e., *disinformation* and *misinformation*. Even though the terms are sometimes used interchangeably, disinformation is often used to denote data intentionally supplied with the deliberate aim to mislead and deceive [62]. In turn, misinformation denotes inaccurate information resulting from unintentional mistakes [61]. Some organizations like UNESCO[4] also use the term *mal-information* to distinguish information that is true but removed from context, that is used to harm persons, social groups, organizations, or countries [94].

The open-source generative AI models have made it easy to produce data of all modalities. The ever-increasing pervasiveness of generative models such as large language models (LLM),[5] stable diffusion models,[6] and multimodal models (MM),[7]

[3] The provenance of data refers to its source and ownership. The provenance chain chronicles its ownership and modifications.

[4] UNESCO stands for the United Nations Educational, Scientific and Cultural Organization.

[5] A large language model (LLM) is a type of neural network model with billions of parameters that has been trained on massive amounts of text data to learn the statistical patterns and relationships in natural language. These models are typically based on transformers and can generate longer texts in a wide range of topics that closely resemble human writing, e.g., OpenAI's GPT (Generative Pre-trained Transformer) [175], Meta's Llama2 model [211], and Mistral AI's Mixtral [100].

[6] Stable Diffusion is a latent diffusion model, a kind of deep generative neural network, text-to-image model released in 2022 [180]. It is primarily used to generate detailed images conditioned on text descriptions. It is also capable of a number of different image manipulations.

[7] Multimodal models (MM) combine data from multiple modes or sources of data, such as text, images, video, audio, and sensor data, to make more accurate predictions or generate more comprehensive outputs. One example of a multimodal model is a system that analyzes both speech and facial expressions to determine a person's emotional state [201].

makes it extremely easy to generate large amounts of text, images, videos, and sounds like cloned voices. Since many of these models generate new data by predicting the most probable continuation of data, sometimes called confabulating or hallucinating, it is not certain that the new data is true, accurate, or even correct. Nonetheless, the models are extremely good at generating data that read as plausible, and the ease and accessibility of the models have made them very popular. Adding to this there are easily downloadable apps that create images and videos given only a simple photo or a written description, e.g., zao [4], deepheritage [150], resulting in the easy creation of deepfakes used for malicious intents [40] and the spreading of misinformation and disinformation on social media [6]. This evolution has forced many prominent social media companies to shut down accounts spreading disinformation on a daily basis [189].

In the best of cases, spreading false data is unintentional and due to honest mistakes or misunderstandings. In the worst of cases, the purpose, when directed at an individual, can be to confuse or hurt them by, e.g., blackmailing them or ruining their reputation [132, 152]. The purpose can also be to present oneself in a better light. However, when spreading outright lies directed at another state or organization, the goal can, e.g., be to sow discord and uncertainty among the population or allies or destabilize a government. There are numerous accounts of foreign intelligence entities, private citizens, and organizations that use fake profiles and other forms of deception on social media platforms [87, 205]. They target individuals for recruitment and information gathering, undermine public confidence in voting, and intensify sociopolitical divisions in a country [123, 153].

2.1.4 Fake news

The term *Fake news* is a popular term used to denote disinformation or misinformation presented as news [54]. According to Faris *et al.* it is also used as a slur term by partisan readers and politicians for news that one disagrees with [56]. As is common in this domain, there is no single definition of fake news. The term is often connected to other related concepts, e.g., Shu *et al.* uses it to indicate "maliciously false news" [196], Zubiaga *et al.* uses it to refer to "a specific type of disinformation" [245], while Buntain and Golbeck and Liang *et al.* use it for *rumors* [29, 126]. Some researchers have, however, presented definitions, e.g., Meel and Vishwakarma define it as "False information spread under the guise of being authentic news usually spread through news outlets or internet to gain politically or financially, increase readership, biased public opinion" [140], and Zhou and Zafarani gave two definitions of fake news, i.e., the broader definition "Fake news is false news" and the more narrow definition "Fake news is intentionally and verifiably false news published by a news outlet" [244].

Even though the term fake news was popularized during the second half of the last decade, the concept is by no means new [182]. Pennycook and Rand state that earlier, it was primarily related to the journalists' task of verifying information [169].

The popularity trend over time of the search term *fake news* is displayed in Figure 2.2. As can be seen in the image [81], there are a couple of noticeable peaks in the search term's popularity. Moving from a steady and low degree of popularity the first significant peak occurred during Donald Trump's candidacy in the 2016 US presidential elections. The 2018 peaks coincide with Jair Bolsonaro's candidacy for the Brazilian presidential elections, and the major 2020 peak in March coincides with the COVID-19 outbreak.

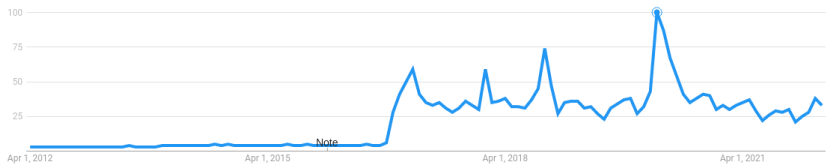

Figure 2.2: Interest worldwide over time for the search term *fake news*. Image from Google Trends [81].

2.1.5 Fake reviews

Another special case of disinformation is the so-called *fake reviews*. According to many, among them Wu *et al.* and He *et al.*, fake reviews are produced to impact consumers' purchase decisions [90, 237]. They also argue that the market for fake reviews is large. This claim is supported by the World Economic Forum that stated in 2021 that "fake online reviews cost $152 billion a year" [133].

The fake reviews can be both positive and negative, i.e., a positive fake review can praise the subject of the review while a negative fake review only discusses how bad it is and can be produced by an adversarial party. In 2012, TripAdvisor was fined for claiming their reviews were *trusted and honest* while the UK Advertising Standards Authority found that approximately 50 million online reviews on their site could not be verified as trusted [213].In 2013, Samsung was discovered to have paid individuals for writing negative online reviews about HTC phones [16]. As a result, Samsung was fined. Additionally, in 2015, Amazon took legal action against more than a thousand individuals for posting fake reviews on their site [65]. In 2021, using official figures and self-reporting by e-commerce sites, e.g., Trip Advisor, Yelp, TrustPilot, and Amazon, about 4% of all online reviews were found to be fake [133]. Luca and Zervas calculated that an extra star in a restaurant's Yelp rating could increase revenue by 5% to 9% [128].

In many of these cases, the producers of fake reviews are people doing it for monetary compensation [52, 65]. However, with the advancement of generative

techniques, it is probable that the need to pay people to write content has been reduced a lot.

2.1.6 Rumor

The term *rumor* often has a negative connotation as it is seen as a tool to spread falsehoods and misinformation [17, 30, 141]. However, the definition of the word *rumor* does not hinge on the veracity of the claims being made. Instead, rumors are understood as "claims whose impact arises from social transmission itself [83], or in the words of Zubiaga *et al.*, a rumor is "an item of circulating information whose veracity status is yet to be verified at the time of posting" [245].

There are several studies that focus on rumor spreading in social media, e.g., Shu *et al.* and Pennycook and Rand, have shown that fake news is more easily spread than true [169, 196]. One study by Vosoughi *et al.* showed that rumors spread up to six times faster and reached further than true news [227]. Other studies have not demonstrated this significant difference [30]. The power of rumors has been especially noticeable in the COVID-19 pandemic fighting, where numerous false rumors regarding prevention and cures have been widely spread [208].

The reasons why false information is spread are not entirely clear. Some studies have seen a disconnect between what people share and what they believe to be true [168]. Early studies by Prasad showed that believability was not a factor that affected rumor spreading, but rather the emotional nature of the situation [173]. A survey of rumors spread on social media by Pennycook *et al.* showed that in addition to accuracy, people shared things according to how interesting, funny, and politically aligned they found the content [168]. Some experiments have found that rumors are distributed more frequently when the level of believability is high [96]. Nevertheless, Shao *et al.*, among others, found that bots often play a disproportionate role in the amplification of content from low-credibility sources in the early stages of article spreading before it goes viral on social networks [192].

2.2 Assessment

In this dissertation, *assessment* refers to the activity of evaluating sources and data with regard to the different aspects of veracity, e.g., accuracy, credibility, truthfulness, and quality. The purpose of an open-source data veracity assessment is often not to differentiate between truth and lies but rather to find a reference point that allows discussion. For example, when assessing (potentially) fabricated data, the purpose might not be to discover any falsehood or truth but rather to *detect* the creation method, propagation rate, or reactions [7, 104, 241]. How the fabricated data is used, and the intent of the producer or source is what, in turn, might constitute a lie, truth, or something else. An example is the use of voice cloning,[8] which is a highly useful technique when making movies to, e.g., allow

[8] Voice cloning is the generation of a synthetic voice designed to mimic a real person's voice.

the actors to speak in foreign languages [114]. However, as it becomes more easily accessible, criminals can also use it for purposes such as fraud, harassment, and blackmailing [202, 235].

It is widely recognized in the literature that manual data assessment is a complex task that involves many factors. Researchers such as Blasch *et al.* and Dragos and Rein point out that defining, implementing, and interpreting assessment scores, as well as effectively using the assessments, pose several challenges [25, 51]. Also, the challenges associated with assessing and incorporating uncertain data are highlighted in the NATO[9] STO[10] technical report [11] on assessment and communication of uncertainty in intelligence. It is discussed that difficulties associated with assessing and integrating uncertain data can lead to inconsistent assessment procedures, and as a result, analysts may tend to ignore assessment procedures or perform them incorrectly.

2.2.1 Manual assessments

In manual data assessment within the military, a scale called the *Admiralty code*, *Admiralty system*, or *NATO system* has been used for a long time to rank items according to a manual veracity assessment [11]. The Admiralty code was created during WWII by the Royal Navy and has undergone little change since then [60, 147]. In NATO, the *Standardization Agreements*, STANAGs, define processes, procedures, terms, and conditions for joint military or technical procedures and equipment, dealing with everything from ammunition standards to communication procedures. Each individual NATO country ratifies the STANAGs and implements them within its armed forces. There are currently over a thousand different STANAGs.

The STANAG 2022 and its updated version no. 2511 outline the admiralty code ranking system for assessing intelligence reports [158]. The two main concepts are the *reliability* of the source and the data's *credibility*. The source reliability rating notation uses an alphabetic coding, A–F, where F is an indication that reliability cannot be judged. Reliability is mainly based on whether the source is previously known and has a proven track record of supplying dependable data, making it trustworthy. A is used to describe a source that is considered completely reliable.

Credibility is ranked using a numeric code ranging from 1 to 6. The assessment is based on a subjective probability assessment of the dependability of the supplied data and the degree of corroboration by other sources. Items that can not be classified are scored with the rating of 6. Reliability and credibility are supposed to be assessed independently, and every combination of A–F and 1–6 is possible.

However, early on, weaknesses of the system were identified [147]. For example, there is no clear methodology on how to apply or interpret the ratings, opening them up to variance in interpretation and the analysts' own biases [185]. On numerous occasions, researchers have presented identified issues with the system [20,

[9]North atlantic treaty organization (NATO) https://www.nato.int/.
[10]NATO's science and technology organization (STO) https://www.sto.nato.int/.

37, 38, 46, 157]. Summarizing, there are several issues with NATO STANAG 2511. The definitions of the core concepts, such as reliability, credibility, and source are ambiguous, missing, and imprecise. Additionally, there are undefined situations, such as dealing with false data created by multiple sources.

To combat some of these issues, some countries have made additions to their assessment frameworks by adding requirements of likelihood and analytical confidence [70]. In the Soviet Union, military information was considered to be trustworthy (*dostovernaya*), probable (*veroyatnaya*), doubtful (*somnitel'naya*), or untrue (*lozhnaya*) [170]. However, these types of nomenclature and additions still present the same type of vulnerabilities as the Admiralty code, and it is worth noting that none of these manual systems are adapted to veracity assessment of data from open sources, e.g., social media or to the automation of assessments.

2.3 Assessment automation

The main question of this dissertation, i.e., how to automate the veracity assessment of data from open sources, is a complex task, and several researchers have tried to suggest approaches.

Pendyala argues that assessing the veracity of data is an NP-hard problem [165], drawing parallels to the misinformation containment problem [210], i.e., how to stop the spreading of misinformation in online social networks by introducing competing information. Lee *et al.* point out that drowning out unwanted information with noise or other information is a well-known redirection technique [121]. The complexity of the veracity assessment task has led to different approaches depending on the context of the data, e.g., rumors, facts, and reviews. There is no one-size-fits-all approach for automating the veracity assessment of all types of open-source data. As previously concluded, veracity is context-dependent, and thus, data and purpose-individualized solutions are needed. Multiple approaches and methods are available, and almost all of them have in common that partial solutions to automatizing are proposed. These partial solutions often aim to assess what, in this dissertation, is called indicators.

In their book, Berti-Équille and Borge-Holthoefer argue that veracity is a quality and truth discovery problem [18], and they introduce a two-dimensional space to classify truth discovery approaches. The approaches are classified according to scope and content type. Along the vertical axis are the three scope types, i.e., content-based, source recommendation-based, and evidence-based. Along the horizontal axis, the content types are divided into structured and semi-structured data, loosely structured data, and finally, textual and multimedia content. Different methods to automate metrics, which also could be called indicators, are surveyed to analyze the data. These methods are categorized as information extraction, truth discovery computation, trust computation, and misinformation dynamics.

Some researchers have proposed frameworks for dealing with veracity, e.g., Lukoianova and Rubin suggested using three main veracity dimensions outlined by

"objectivity, truthfulness, credibility, and their opposites" to assess veracity [130], and García Lozano *et al.* suggested a trust-based framework [70]. García Lozano *et al.*'s framework builds upon a transitive trust idea and combines it with a similarity measure to automate credibility and reliability assessments. Another framework is presented by McArdle and Kitchin, who discuss including crowdsourcing mechanisms to record user observations and fixes in the metadata for improving data quality [137]. However, none of these earlier, i.e., before 2015, veracity framework suggestions were implemented, experimentally tested, or verified.

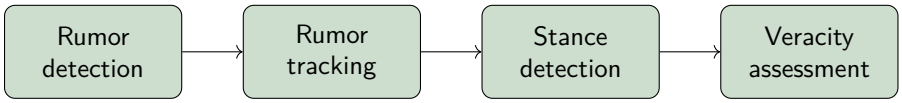

Figure 2.3: Zubiaga *et al.* propose the following rumor veracity assessment architecture modules [245].

Focusing on the detection and veracity assessment of rumors in social media, Zubiaga *et al.* proposed in 2018 a four-step architecture [245]. The steps being: rumor detection, rumor tracking, stance classification, and veracity classification, see Figure 2.3. In the first step, statements need to be analyzed and classified to identify whether they are part of a rumor. The second step requires tracking the spreading of the rumors. In the third step, the reactions regarding support, negation (deny), questioning, or neutrality (comment) (SDQC) to the topics are analyzed. Finally, in the fourth step, following the findings of, e.g., Castillo *et al.* stating that false rumors tend to be questioned more [32], a veracity assessment is completed based on the analysis of the stances combined with other indicators. In another effort to focus on automated fact-checking, the authors view the process as a pipeline where check-worthy claim detection, claim matching, and claim validation are the individual steps [242].

In the systematic literature review paper from 2020, García Lozano *et al.* identified three main approaches to automation of veracity assessment of data from open sources, i.e., utilizing implicit features, employing explicit fact checking, and appealing to an authority [68].

2.3.1 Assessment value

A question one could pose is what value a veracity assessment holds. In an analogy to Shannon's information entropy concept [190], the value of an assessment activity can be compared to the informational value of a communicated message, i.e., it depends on the level to which the message's content is surprising. As a veracity assessment of a subjective statement probably does not provide any new information on, e.g., its truthfulness or quality, an indirect method can thus be more fruitful. Assessing the geographical location of the topic or the stance towards a topic of interest may yield more relevant information [72, 75]. In this dissertation, examining

and assessing this indirect information is called an indicator, and if carefully selected, it can result in a higher informational value. Common to all of the suggested automation approaches is that they rely on a data process that utilizes not only the features of the data in itself or the related metadata but also the surrounding associated indicators.

Two evaluation metrics for data veracity techniques are proposed by Assiri [12]. The first is the *computational* cost, i.e., the amount of time an algorithm or process takes to run. The second is the *communication* cost, i.e., the amount of communication (a function of data volume) that is needed between parties to solve a problem.

Following the logic of focusing the assessment activity on the one expected to give the highest informational value with the lowest computation and communication cost, we focus our attention on indicators.

2.3.2 Indicators

When it is not possible to directly determine the truthfulness of a statement or claim because its veracity status is unknown, such as with many rumors, an indirect approach can be used to obtain a veracity assessment. This indirect approach can be achieved through the use of indicators, which provide complementary data or catch something else that indirectly contributes to the assessment of the statement's veracity.

An example of employing indicators could be in a case where only data from social media is available. Imagine a rumor is spread about an explosion at a chemical plant [33, 206]. As there is little or no verified information available and all discussions about the event come from social media, in the initial phase of the rumor's spreading, it is not feasible to assess the rumor's veracity [246]. One indicator suggesting that the rumor has low veracity could be that many of the accounts that are spreading the rumor exhibit automatized behavior, i.e., are bots [126, 181]. Another indicator could be the geographical location related to the topic or of the sources spreading the rumor [75]. If they are far from the explosion site, those sources cannot be eyewitnesses to the accident. Yet another indicator could be the stance, i.e., people's reaction to what is being said [110]. These indicators imply to some degree that the sources are unreliable, which, in turn, may contribute to the final assessment that the rumor has low veracity.

Single or multiple indicators can be used to enable the veracity assessment process, and these indicators can have a positive or negative effect on the assessment confidence. Depending on the context the indicators can be tailored to analyze and utilize a wide variety of data contents, structure, semantic implications, sentiments, metadata, sources, relations between data, spreading patterns, trends, etc.

2.3.3 Features

Related to indicators is the concept of *features*. In machine learning, features are properties or characteristics of a phenomenon that can be independently measured [57]. Machine learning features can be used to, e.g., identify an indicator. An important task in machine learning is feature engineering, e.g., identification, transformation, and selection [127]. Many methods have been developed to achieve an optimal feature selection. Many have also been explored in relation to veracity, e.g., principal component analysis (PCA),[11] lexical, probabilistic context-free grammars (PCFG),[12] term frequency – inverse document frequency (TF–IDF),[13] and depth of propagation tree [31, 32, 69, 88, 195].

2.4 Stance detection approaches

Stance detection is the automatic classification of the position the producer of a piece of text takes towards a target, typically *favor, against, or neither* [116]. Variations of the classes exist, and sometimes *question, and commenting* are added [246].

Already in 2010, when analyzing tweets related to the Chilean earthquakes, Mendoza *et al.* discovered that false rumors to a higher degree than true tended to be questioned [141]. Along the lines of the original admiralty scale ideas, see Section 2.2.1, Zhang *et al.* concluded that user credibility is an essential factor that impacts information credibility, thus analyzing user credibility helps detect message credibility [243]. The authors also observe regarding a rumor's credibility that a large number of doubtful and inquiring comments often appear in rumor messages, and therefore, opinion analysis or stance analysis on the comments can be used to obtain the message's credibility. Another study concluded that, while the overall tendency for users is to support unverified rumors in the early stages, there is a shift toward supporting true rumors and debunking false rumors as time goes on [246].

In 2017 Derczynski *et al.* proposed a task [47] at the SemEval workshop using the PHEME Twitter dataset [246] to explore stance detection. Several contributions were submitted, and noticeable was that the majority of those that performed well employed some form of neural network. The winning team used a bidirectional long short-term memory (Bi-LSTM)[14] network [115]. García Lozano *et al.* used

[11] Principal component analysis (PCA) is a dimensionality reduction method that is often used in machine learning to reduce a large set of variables into a smaller one that still contains most of the information in the large set.

[12] Probabilistic context-free grammars (PCFG) is used to understand the structure of natural languages.

[13] Term frequency – inverse document frequency (TF–IDF) is a statistical method that measures how important a term is within a document relative to a collection of documents (corpus).

[14] A bidirectional long short-term memory (Bi-LSTM) network is a type of recurrent neural network (RNN) with an additional LSTM layer allowing it to process sequential data both forward and backwards. An LSTM is designed to capture long-term dependencies in sequential data such as language.

a combination of hand-crafted rules and CNNs for stance detection in Twitter rumors [72].

A more fuzzy stance detection approach is used by [110]. Instead of using predefined stance classes, the authors use a similarity measure obtained by using embeddings in the language model BERT[15] [48]. The cosine similarity between the vectors of the news title and body is used as the similarity measure. The higher the cosine similarity, the more credible the articles are considered.

Stance detection has also been used to detect fake news, and the authors, Davoudi *et al.*, simultaneously analyze the propagation tree and stance network news article features to assess the news' veracity [43]. Singla *et al.* divide the stance detection process by focusing on three different sets of attributes [199]. The first set focuses on the data structure and things like the number of retweets, question mark presence, and number. The second set focuses on individual properties like similarities between tweets and distance to the root tweet. The third set focuses on sentiment and emotional response categorized by a number of standards like LIWC[16] [166] and DAL[17] [234].

Sarcasm,[18] similes,[19] and metaphors[20] among other types of rhetorical devices, have been identified as challenging issues that have received less attention in stance detection [199]. Also, many of the stance detection methods, e.g., the third set of attributes in [199], depend on the English language and cannot be easily transferred to other languages or cross-language situations.

2.5 Rumor veracity detection approaches

Detection of rumor veracity can be divided into two main parts; the first, somewhat simplified, is a classification problem of something being a rumor or not, and the second part is the veracity assessment [219, 240, 243]. Focusing on the second part, some researchers simplify the veracity assessment by relying on the *appealing to an authority* [68] method and use features derived from fact-checking sites such as Politifact,[21] factcheck,[22] or state-issued counter-rumors [239]. However, since manual verification steps are needed for debunking rumors, these fact-checking websites are not comprehensive in their topical coverage and can also have long debunking delays. Thus, another method employed by researchers is identifying

[15] BERT is an acronym for "bidirectional encoder representations from transformers" suggested by Devlin *et al.* in 2019 [48]. It is one of the first pre-trained encoder-based transformer models used for natural language understanding (NLU). NLU aims to understand the context and intent, i.e., what is meant, in natural language.

[16] Linguistic inquiry and word count (LIWC) [166].

[17] Dictionary of affect in language (DAL) [234].

[18] A sarcasm is saying one thing but meaning the opposite.

[19] A simile is a figure of speech directly comparing two things.

[20] A metaphor is a figure of speech that refers to one thing by mentioning another.

[21] https://www.politifact.com/.

[22] https://www.factcheck.org/.

features that may indicate low credibility. A system called TweetCred was developed by Gupta *et al.* to evaluate the credibility of tweets in real time [85]. The credibility is assessed using an SVM[23] model, which is semi-supervised and trained with data from six different 2013 crisis events. A set of 45 features is used to calculate the credibility score for each tweet. The feature types and number of features vary with different approaches. A review paper by Varshney and Vishwakarma identifies multiple hand-crafted features for analyzing rumors [220]. The features are divided into 15 categories, i.e., message-based, user-based, topic-based, propagation-based, content-based, network-based, Twitter-based, linguistic-based, temporal, user-behavioral, diffusion-based, structural-based, social features, visual features, and statistical features. Other authors simplify the feature types into fewer and broader types, e.g., content-based features, user-based features, propagation-based features, and other-based features [243]. Another Twitter credibility analysis system was proposed by Alrubaian *et al.* [9]. The system consists of four interconnected components, i.e., a reputation-based component, a credibility-determining classifier engine, a user experience-enhancing component, and a feature-ranking algorithm. Common to most of these systems is that the features are primarily hand-crafted.

Not all scholars are focused on verifying or assessing rumor credibility; some focus on finding rumor origins, rumor-spreading patterns, and other types of features and indicators that can be used for veracity detection, e.g., García Lozano *et al.* study the relation between topics and geographical locations [75]. According to Meel and Vishwakarma, three main models can be used for classifying rumor propagation in social media, i.e., soft computing, epidemiological, and mathematical [140]. Other authors divide rumors according to their temporal characteristics [245]. There are two types of rumors that circulate during breaking news. The first type consists of new rumors that emerge in the early stages, while the second type is made up of long-standing rumors that are discussed for extended periods. The veracity detection methods differ for the two groups. According to a study by Zubiaga *et al.*, users tend to support unverified rumors at the early stages, but over time, they start to support true rumors and debunk false rumors [246]. In another type of approach, Li *et al.*, managed to increase the rumor assessment accuracy by using the user credibility information in combination with other stance detection methods [125]. The authors derive user features, such as whether the account is verified and if the profile includes a location or description.

The complex and time-consuming task of implementing hand-crafted features spurred researchers to look for alternative methods. Following the fantastic results in the early 2010s brought on by the enabling of deep learning, several groups began using neural network variants to assess rumor veracity. In a literature survey, Islam *et al.* divided the used deep learning techniques into three main categories based on

[23] A support vector machine (SVM) is a (usually) supervised learning algorithm (it learns from labeled data) that is used to solve classification, regression, and outlier detection problems by determining boundaries between data points based on predefined classes.

the model, i.e., discriminative models, generative models, and hybrid models [95]. In the pioneering paper from 2016, Ma *et al.* present a method where they use recurrent neural networks (RNN) with gated recurrent units (GRU)[24] to achieve better than state-of-the-art early rumor detection[25] [131]. They also curate a dataset with rumors collected from Twitter and Sina Weibo microblogs. However, the authors do not entirely manage to automatize the whole process, and manual work is required for feature engineering. Employing the dataset curated by Ma *et al.*, Chen *et al.* devise an LSTM and attention model [34]. Guo *et al.* also use LSTM and attention models for early rumor detection [84]. A number of other approaches pop up, e.g., Jin *et al.* use RNNs to assess multimodal rumor data [101], Li *et al.* use GRUs [124], Nguyen *et al.* use CNN and LSTMs for early rumor detection [156], and Alkhodair *et al.* use a combination of word2vec[26] and LSTM-RNN [8].

To conclude, the rumor domain has made significant progress in latter years and there are many good partial solutions to the rumor veracity detection challenge. However, there is still no full solution or system that is able to automatically scan microblogs or other social network forums to assess trending or pervasive rumor veracity without any type of manual interference.

2.6 Fake review detection approaches

Fake reviews are a special case of disinformation. The fake review detection challenge can be simplified as a classification problem where a review is either fake or not [102]. In one of the earliest papers on fake review detection, Jindal and Liu lists three different types of problematic reviews, the first type is called untruthful opinions, i.e., reviews written with the purpose to promote or discredit products unjustly, also known as fake reviews [102]. The second type is reviews that focus more on the brand behind the product than the product itself, and the third type is the non-reviews, e.g., spam-like advertisements, questions, and non-relevant comments. Focusing on the first type of reviews, the authors found by crawling amazon.com a large number of identical or near identical reviews and a group of reviewers that likely had written a large number of these reviews. Using similarity between reviews to assess them as fake is also used by Kauffmann *et al.* in their big data analytics framework [111]. Jindal and Liu also identified three main feature types that can be used to detect a fake review, i.e., the review content, the reviewer, and the features related to the reviewed product [102]. However, the authors emphasized the difficulty of obtaining a relevant dataset. Other researcher, like Ott

[24] Gated recurrent units (GRU) is a variant of RNNs designed to increase the speed performance of LSTM networks with a gating mechanism that allows it to input or forget certain features [36].

[25] The authors define early rumor detection as the detection of topics that have disputed facts.

[26] Word2vec is a group of shallow, two-layer neural network models that are used to produce word embeddings (vector representations) by capturing information about the meaning of a word based on its surrounding words [145].

et al., have thus created pseudo reviews, i.e., reviews produced en masse by people paid to write reviews on services like AMT[27] [162].

One of the challenges in assessing fake reviews is obtaining datasets with genuine fake reviews instead of pseudo reviews. Mukherjee *et al.* showed that fake reviews are harder to spot than pseudo reviews [149]. The authors also showed that models trained using AMT fake reviews had problems detecting genuine fake reviews filtered from Yelp, indicating that pseudo reviews probably do not represent genuine fake reviews.

In a survey from 2018, Patel and Patel concluded that a variety of traditional supervised, semi-supervised, and unsupervised machine learning methods, e.g., Logistic regression, SVM, and K-nearest neighbor, were used to detect fake reviews [164]. For example, García Lozano and Fernquist explored possible classification attacks and used an SVM-based detection method [69].

As deep learning methods began to gain popularity, different types of neural networks started to be used. Fang *et al.* propose what they call a *dynamic knowledge graph-based* method for fake review detection [55]. They create a knowledge graph to tie together four indicators of fake reviews, i.e., reviewer trustworthiness, review honesty, commodity splendid degree,[28] and store reliability. These four indicators combined with a neural network model called a sentence vector/twin-word embedding conditioned bidirectional long short-term memory.

In 2020, Rodrigues *et al.* also surveyed the employed detection approaches and summarized the used techniques as either belonging to the traditional machine learning methods or the newer deep learning techniques [179]. Examples of the employed deep learning approaches are CNN, RNN, LSTM, Bi-LSTM, and GRNN,[29] with various activation functions. In the slightly more extensive survey by Mohawesh *et al.* the authors also summarize some of the public existing datasets that can be used for fake review detection [148]. They classify the datasets according to the construction method, i.e., filtering algorithm, human, Amazon Mechanical Turk, and rule-based. The filtering algorithm method employs the review sites' labeling; however, how the review sites' own filtering algorithms work is unknown. The human method uses a group of people to assess and annotate a set of reviews. Depending on the majority vote, a review is then marked as fake or not. Nevertheless, several studies show that humans are relatively poor at detecting fake reviews [27, 161, 184]. The AMT method relies on crowdsourcing fake reviews by letting people create them on demand. As previously stated, these types of reviews are also called pseudo-reviews and have been shown to have a different distribution

[27] AMT is an acronym for Amazon mechanical turk which is a crowdsourcing marketplace for quick jobs.

[28] The purpose of this measure (commodity splendid degree) is to distinguish between high-quality products and suspicious ones. High-quality products are likely to receive genuine praise from reliable consumers, while suspicious products will attract fake accolades, primarily from spammers.

[29] General regression neural network (GRNN) is a variation of radial basis neural networks, i.e., they use radial basis functions as activation functions in the neurons.

than real fake reviews [149]. In the rule-based method, Mohawesh *et al.* refer to the method employed by Jindal and Liu (described earlier in this section) to create their large Amazon dataset [102]. A criticism of this method is that some reviews might be unjustly categorized as fake due to network failures or user mistakes, creating multiple copies of the same review.

Mohawesh *et al.* also summarize feature types, i.e., meta-data, POS,[30] bag of words (BoW),[31] LIWC, stylometric, semantic, and word embeddings, are used to assess review validity using supervised statistical machine learning models. For unsupervised statistical machine learning model approaches, the features are somewhat different, i.e., LDA,[32] considering duplicates as fakes, content-based features, behavior-based features, and relation-based features. For the semi-supervised statistical machine learning models, a mixture of the two feature types (used in supervised and unsupervised approaches), e.g., LDA, unigram and bigram features, metadata, text content, reviewer features, and reviewer social networks, are used. Looking from the more traditional statistical machine learning methods towards the deep learning approaches, CNNs, RNNs, and LSTMs are the most popular methods used. Examples of features that are used are character levels, pre-trained word2vec, GloVe,[33] and user behavior.

Some trials with GAN models have also been used, but in the trials, the GAN models did not outperform other deep learning methods [5, 207]. The advantages of deep learning methods over traditional machine learning methods are that the need for feature engineering is lower, and the models usually perform better with large datasets.

In 2022 Salminen *et al.* employed large language models to generate fake reviews and show that these models are fairly good at forging reviews, but also that these types of automatically generated reviews are detectable with high accuracy [184]. An advantage of using LLMs to generate fake reviews is that they can generate balanced fake review datasets and training models in languages other than English, the prevalent fake review research language. In 2019, García Lozano and Fernquist experimented with translations of fake reviews back and forth from English to other languages to try and fool the fake review detection accuracy by obfuscating the stylometric clues in the texts [69]. The results show that the accuracy dropped, but not sufficiently to use it as a viable method for fooling automatic fake review detection models.

[30] Part-of-speech (POS) or grammatical tagging is the process of assigning words in a text to their corresponding part of speech, based on both its definition and its context.

[31] A bag of words (BoW) is a representation of a text that is based on an unordered collection of words that captures the number of times each word appears in the text.

[32] LDA is an abbreviation of latent Dirichlet allocation proposed by Blei *et al.* in 2003 [26]. It is a generative probabilistic model that, in a textual context, can be used for latent topic detection in documents of a corpus.

[33] Global vectors for word representation (GloVe) is an algorithm for obtaining vector representations for words [167].

Chapter 3

Summary of publications

> If we knew what it was we were doing, it would not be called research, would it?
>
> *Unknown*

This chapter is comprised of a summary of the publications submitted to this compilation dissertation. The publications are described in a *logical* rather than chronological order. Each publication's relation to the dissertation contributions, presented in Section 1.5, is described as well.

3.1 P1 – Towards automatic veracity assessment of open source information

Publication P1 – "Towards automatic veracity assessment of open source information" [70] has two main parts consisting of i) a literature review regarding the use and the discovered challenges with assessment scales, and ii) interviews and an online survey with military personnel regarding their use and experience with open-source data assessment and assessment scales. The purpose is to find which challenges need addressing and approaches that may be taken to assess open-source data's veracity automatically. It serves as a complement to other literature [19, 50, 216–218] on the topic of issues related to state-of-the-art in veracity automation within the military domain.

The NATO STANAG 2511 (an update of STANAG 2022 sometimes also called the admiralty scale) assessment scale is used to assess source (reliability) and data (credibility) according to a six graded scale [158]. The admiralty scale is a scale that is highly popular among many military organizations around the globe. Sources are ranked from A to F based on their assessed reliability. Sources are ranked from A to F based on their reliability. A is used for a source that is regarded as completely

reliable, while E is used for unreliable sources. F is used for sources that have not been used before and whose reliability cannot be determined.

The level of credibility is judged from 1 to 6, where one is used for data that has been confirmed, and five is used for data deemed improbable, e.g., data contradicted by other data coming from reliable sources. Level 6 is used for data that cannot be judged. Some of the variations and additions that have been made to the scale are also studied and discussed.

The paper contains a literature review that is focused on other researchers' experiments and discovered issues using the NATO STANAG 2511 [20, 37, 38, 46, 157]. In summary, the primary identified challenges are of two types, the first being that the core concepts have missing, ambiguous, or imprecise definitions. The second is that there are many undefined situations and little guidance on how the recommendation should be applied. This leads to variation in the interpretation of the scale's use and application method and susceptibility to the analysts' bias and level of competence. Some countries and service branches have added to the NATO 2511 assessment scale with, e.g., a likelihood and judgment confidence scale, to ameliorate the found deficiencies.

Due to the challenges and assessment scale variations found, the work was also directed toward investigating how veracity assessment is implemented in practice, focusing on open-source data. The task was handled through a three-pronged approach: i) a questionnaire, ii) interviews at the 2013 Combined Joint Staff Exercise in Sweden, and iii) meetings with Swedish intelligence analysts. The questionnaire had four main parts directed to the respondents: i) experience, ii) opinions on definitions and attributes, iii) questions regarding source judgments, and iv) thoughts on how a source and data is regarded and judged. The questionnaire, interviews, and meetings resulted in many insights into how open-source data veracity assessment is implemented. One of those insights is that open-source data is seldom rated higher than C3. There exists a discrepancy in how a source is viewed. Some saw, e.g., the web as a single source, while others saw websites as different sources. Also, how the respondents would evaluate and give feedback on an assessment highlighted the need for traceability and the use of the so-called *gut feeling*. The constant lack of time was also a significant problem. It was perceived as complex and time-consuming to assess sources and data continuously; thus, follow-ups were seldom done. Finally, a significant challenge is that the assessments are subjective, making the analysts vulnerable to their individual biases and mental models.

For an automation approach to be accepted by an end-user and seen as qualitative and trustworthy, the questionnaire and interview results indicated that an assessment needs, among other things, to be traceable regarding sources, data, and methods. Also, the assessment scale needs to be well-defined and unambiguous. Further on, an assessment should preferably have a confidence value and possibly a timestamp, i.e., a best-before date. In the paper, homophily, i.e., similarities, was argued to be a good starting point for assessing source credibility and data reliability. Three main types of metrics were identified in the paper that can be used to obtain indicators, i.e., content, meta-data, and rating-based metrics.

This publication's main contributions in relation to the dissertation are the results from the survey, interviews, and compilation of challenges for open-source data assessment using NATO STANAG 2511, see contributions C1– *Compilation of challenges* and C2 – *Roadmap* in Section 1.5. Also, the research results emphasize the need for systematization and automation of the veracity assessment, leading to predictable and consistent behavior in the assessment process.

3.2 P2 – Veracity assessment of online data

Publication P2 – "Veracity assessment of online data" [68] is a structured literature review of papers related to veracity assessment of online data, targeted towards social media and open-source data [68]. The aim is to understand research trends and determine future research needs. The paper serves as a complement to other related secondary research approaches [41, 45, 78, 99, 178, 221, 225, 232].

The research question studied in the paper is: "Which approaches, methods, algorithms, and tools are used or proposed for automatic veracity assessment of open-source data?" A structured literature review method was used to answer the research question employing the guidelines proposed by Kitchenham and Charters [112].

In the study, five research databases were queried with the same 17 search questions, generating a total of 5047 hits. After a rigorous trimming process using inclusion and exclusion criteria, 107 publications remained. The inclusion and exclusion criteria resulted in filtering the database hits based on topic, language, and publication year. After the filtering, an arduous process of reviewing and assessing each publication based on a previously agreed-upon review protocol took place. The review protocol consists of three parts. The first is an administrative part with questions regarding, e.g., authors and their affiliations. The second part is the research question, which is further divided into six groups, i.e., approaches, methods, algorithms, tools, data, and miscellaneous questions about issues indirectly related to the main research question. The third and final part of the review protocol is the qualitative assessment part. The analyst had a chance to state their impression of the reviewed publication, contributions, and quality. The review protocol results were then analyzed and summarized with a gap analysis as the final touch.

Three main research direction approaches for veracity assessment were identified. They were called i) utilizing implicit features, ii) employing explicit fact-checking, and iii) appeal to authority approach. Although these approaches are frequently utilized together, the implicit features approach is the one most widely used. The underlying idea is that claims that are not completely true in some way possess distinct characteristics that set them apart from truthful claims. For instance, Igawa *et al.*, Kumar *et al.*, and Popat *et al.* examined indications provided by stylometric text features [93, 117, 172]. In their research, Shao *et al.*, Yan *et al.*, and Vosoughi *et al.* studied temporal distributions [191, 226, 238]. The distribution patterns within social networks were analyzed by Abbasi and Liu, Shao *et al.*, and

Yan *et al.*, as stated in their research [1, 191, 238]. Others, like Saez-Trumper and Tološi *et al.* employed user account features [183, 209].

The approach of explicit fact-checking, which involves comparing a claim to an existing and confirmed body of knowledge, is rare, but there are examples such as those by Shiralkar *et al.* and Levchuk and Blasch [122, 194]. Another approach, the appeal to authority approach, is even rarer. This approach involves considering a claim as true if it is made by an authoritative source, as suggested by Jain *et al.* [97], or if it is part of a general consensus, i.e., if a majority agrees with it, as explained by Namihira *et al.* [151]. However, there is a challenge in determining whether a source is authoritative enough or whether a vocal opinion can be regarded as an authority.

The most widely used method for verifying the truthfulness of a given text is to apply supervised learning techniques to analyze it. As a result, many research papers tend to have a limited focus, as they concentrate on addressing a specific issue by examining a single type of data from a single source. During the last decade, significant developments have taken place in the field of machine learning, particularly in deep learning techniques. However, most papers have focused on supervised methods, with few utilizing these advancements.

Several research gaps were identified, including i) the general lack of research employing multiple sources and data types, ii) the lack of standard definitions of core terminology, iii) the prevailing scarcity of implementation, data gathering, and sharing details, leading to low reproducibility, iv) general poverty in datasets suitable for benchmarking, v) the low degree of use of modern ML techniques such as deep learning and transfer learning, and vi) the general shortcoming in used methods and data, not adapted to scalable situations or handling of streamed data.

A significant conclusion is that the overall veracity assessment problem is complex. It requires a combination of data sources, data types, indicators, and methods. Only a very few of the studied papers take on such a broad scope, thus demonstrating the veracity assessment domain's relative immaturity.

The main contributions of this publication in relation to the dissertation are the analyzed research results from the publication reviews and conclusions from those results, i.e., contributions C3 – *Literature review identifying the state-of-the-art*, and C4 – *Gap enalysis* defined in Section 1.5. Also, the research results support the central conclusion of the publication, i.e., automation of open-source data veracity assessment is a complex and challenging task. In addition, the research results showcase the popularity of utilizing implicit features for veracity assessment and highlight the viability of using features and indicators.

3.3 P3 – Tracking geographical locations using a geo-aware topic model for analyzing social media data

Publication P3 – "Tracking geographical locations using a geo-aware topic model for analyzing social media data" [75] describes the design and implementation of a

geographic location-aware topic model, which is used to analyze streaming, social media data such as tweets. The innovative geo-location approach presented in the paper focuses on identifying and tracking a topic's geo-location, e.g., a music tour is related to a temporally changing geographical location depending on where the latest concert is supposed to take place. The paper can be viewed as a complement to other geo-location efforts, which mainly focus on discovering either the account holder's or tweet's geo-location, i.e., where the tweet was sent from [215, 228, 229].

Lau *et al.* [120] built on a batch document processing idea previously presented by AlSumait *et al.* [10]. However, using a dynamic vocabulary and updating the priors α and β of new time slices using θ and φ from the previous time slice Lau *et al.* were able to adapt the model to streaming data [120]. This model, which we denote *streaming* LDA (SLDA), has several advantages over regular LDA. For example, its time complexity is independent of the number of time slices, and it is sensitive to emerging trends since it has a dynamic vocabulary, allowing the detection of trendy words. SLDA is the model we use as a basis for our geo-location model.

The method used consists of four main steps in a data workflow process. The first step is data collection, followed by data pre-processing. The third and fourth steps are data analysis and result evaluation. The method is evaluated using the American presidential primary elections of 2016 as a study case and then comparing the results with a keyword-based approach. The data collection step was done with Apache Flume. It can stream large amounts of tweets and, for experimental purposes, store them in an Apache Hadoop-compliant distributed resilient storage. The data pre-processing consists of several steps, e.g., retweet handling, named entity recognition, tokenization, sentence splitting, and lemmatization. The SLDA model was implemented in Java utilizing Apache Spark. The pre-processed tweets are then further processed in a streaming LDA window. The result is then used to discover topics, trends, and their geographical correlation.

This paper's contributions to the dissertation are the research results providing examples of automatized indicators, i.e., topic and geographical location; see contribution C5 – *Indicator exploration, topic geolocation*in Section 1.5. The indicators are not only automatized but also combined to achieve added value. Furthermore, the research results demonstrate that online discussions can be tracked geographically over time.

3.4 P4 – Mama Edha at SemEval-2017 Task 8: Stance classification with CNN and rules

Publication P4 – "Mama Edha at SemEval-2017 Task 8: Stance classification with CNN and rules" [72] is about exploring methods for stance classification (support, deny, query, or comment) for messages in Twitter conversation threads related to rumors. The publication was submitted to a workshop competition task for rumor veracity detection (closed-world assumption), i.e., Task 8 RumourEval: Determining rumor veracity and support for rumors, Subtask A (SDQC) [47]. A dataset

(gathered in the EU-project PHEME) with eight rumors disseminated through Twitter was used for training and testing. The dataset was manually annotated according to a scheme involving assessments by social media researchers, journalists, and through crowdsourcing [246]. The stance classification task is of interest since it may provide a basis, as an indicator, for rumor veracity assessment. This publication serves as a complement to other stance classification efforts [15, 59, 89].

A data workflow with multiple parallel data pipelines was employed to solve the task of stance classification. The three main parts of the data pipeline are: i) data processing, ii) feature engineering, and iii) stance classification.

The first step of data processing was to extract chosen parts of the raw data into new subsets of attributes tailored for each classifier type. Among the extracted data attributes were the tweet's text content and metadata related to the tweets and the users, e.g., time of posting and number of followers. Regarding the division of training, development, and test data, the SemEval Task organizers' data split was used.

In feature engineering, most of the raw metadata attributes could be used as features for classification without further processing. However, the tweet texts required some work due to their noisy nature with abbreviations, contractions, links, and other symbols. Some pre-processing steps were, e.g., splitting contractions, removing and replacing mentions and links with other suitable and generic symbols, and stemming.

As part of the research, it was of interest to investigate whether the grammatical structure could be helpful as a replacement for the tweets themselves. An English pre-trained model for the initial tweet to part of speech (POS) tagging was used to do this. Also, since it was shown to perform better on the shorter Tripadvisor comments, the same rule encoding as Feng *et al.* used [58] was used, and the one-level neighborhood semantic rules, i.e., the $\hat{r}*$ notational case was chosen. Thus, it was hypothesized that Twitter texts would share shorter grammatical structures than more extended ones.

In the final step, i.e., the stance classification, the approach was to utilize a conglomerative approach with three main classifiers and use their individual strengths to rank the predictions in the final classification verdict. The idea is that different types of classifiers may have a predisposition to different concepts and may, therefore, be able to complement each other, resulting in a better prediction capability for the joint classifier. Furthermore, the accuracy of applying the ensemble approach to pure messages using relabeled data, grammatical structure, and combinations was evaluated. Combining convolutional neural networks with automatic rule mining and manually written rules, the ensemble classification approach achieved a final accuracy of 74.9% on the competition's test dataset.

This publication's relation to the dissertation is the exploration of suitable indicators to predict the veracity of an unverified rumor; see contribution C6 – *Indicator exploration, stance classification* in Section 1.5. The research is an example of an approach towards automated veracity assessment employing indicators and data

features. The research results lead to the conclusion that combining different types of classifiers results in better stance prediction capability.

3.5 P5 – Identifying deceptive reviews: Feature exploration, model transferability, and classification attack

The main focus of publication P5 – "Identifying deceptive reviews: feature exploration, model transferability and classification attack" [69], is on automatic identification of deceptive, both positively and negatively biased, reviews. Hence, a deceptive review SVM-based classification model is built. The performance impact of using different feature types (TF–IDF, word2vec, PCFG) is also explored, and the transferability of trained classification models applied to review datasets of other types of products is studied. Furthermore, the classifier robustness, i.e., the accuracy impact against attacks by stylometry obfuscation through machine translation, is also studied. This publication complements other automatic review classification efforts on the same datasets [58, 161, 162].

Two different datasets, called AMT and CSI, were used to train the classification models. The AMT dataset is the same that was used by Ott *et al.* [161]. It consists of 1600 truthful and deceptive hotel reviews written in English. While the deceptive reviews were created by paying people to write them using Amazon's Mechanical Turk service, the truthful reviews were mined from hotel review sites such as TripAdvisor, Expedia, and Yelp. All of the reviews are either of a positive or a negative nature. The second dataset is the CLiPS stylometry investigation corpus (CSI) [223]. It is a Dutch corpus containing both essays and reviews. The review segment of the dataset includes 1,298 reviews, comprising both truthful and deceptive reviews. The reviews, which can be positive or negative, are written by students and cover various topics such as musicians, food chains, books, smartphones, and movies.

Since a linear kernel SVM is a good algorithm for text classification [105, 214], it was chosen as the classification model to use as a basis for the experiments. The workflow pipeline consists of three main parts, i.e., i) data pre-processing, ii) feature selection, and iii) classification. The data pre-processing primarily consists of creating variants of the texts by employing stemming and stop-words in different combinations. In the second workflow pipeline step, three main feature types were explored. The first feature type is the term frequency-inverse document frequency [176]. The second feature type is the output vectors of the word embedding tool word2vec [146]. The third feature type used is the probabilistic context-free grammar (PCFG) method [107]. PCFG focuses on the sentences' structural buildup, while the other feature types focus on the words themselves and their context. Three types of experiments were conducted for the classification: i) feature exploration, ii) classification model transferability, and iii) translation impact. The purpose was to address the following three research questions:

- How do different features affect the performance of deceptive review classification using a linear support vector machine (SVM) based classification model?[1]
- How well do trained models transfer in terms of classification accuracy between different product review datasets?
- How does machine translation influence the stylometric, i.e., linguistic patterns of deception and truthfulness, and what consequences does that have on the classification accuracy?

In conclusion, i) a veracity assessment accuracy of over 91% (AMT), 84% (CSI), and 85% (combined) is achieved using a combination of the TF–IDF and PCFG features, ii) perhaps, less surprising, the trained classification models do not perform well when applied to datasets containing reviews of different products than what the model was trained on, leading to an accuracy drop of around 53%, and iii) on a happier note, the veracity accuracy results are only slightly impacted by using machine translation and can thus not be used as a viable veracity classification attack method.

This publication's contribution to the dissertation is the provision of an example of a fully automatic veracity assessment process. The research targets the classification of positive and negative truthful and deceptive reviews; see contribution C7 – *Automatic veracity assessment, i.e., the study of a type of deception* in Section 1.5. The research results lead to the conclusion that classification model transferability is low with an SVM and that there is no general learning of deceptive review detection regarding the feature types; the conclusion is that the used words are more significant for detecting deception than the constructed grammar rules. Furthermore, on a more positive note, the research results also lead to the conclusion that obfuscation of the texts through the use of translation between languages to confuse the classifier has a low impact.

[1] For better clarity, the question has been slightly reformulated compared to the original phrasing.

Chapter 4

Concluding remarks

> New beginnings are often
> disguised as painful endings.
>
> *Unknown*

This chapter comprises a summary and conclusion of the research work, including the thesis statement, research questions, contributions, and legal and ethical considerations that were taken into account during the research. Additionally, sections analyzing the dissertation scope and the work's validity and reliability are included. The chapter concludes with a discussion of potential future research directions.

4.1 Conclusions

When dealing with data-intensive operations, the quality of the used data is crucial. High-quality and relevant data is often costly to obtain in terms of time and effort. Although a vast amount of open-source data is available, it is mostly unverified and often includes noise such as ambiguities, untruths, biases, and contradictions. Additionally, it is common to encounter fake news, influence operations, and disinformation. Therefore, assessing the veracity of open-source data before it's use is essential. Even if the veracity level is low, the assessment results can provide valuable insights into how the data can be viewed. However, the sheer amount and heterogeneity of open-source data makes manual assessment difficult. Thus, there is a need for automated support in the form of approaches, methods, algorithms, and tools to verify the veracity of data from open sources.

The work presented in this dissertation aims to investigate the **thesis statement**, first stated in Section 1.1, i.e.,

> *The careful extraction, estimation, and use of relevant features and indicators are key to automating and quantifying veracity assessment of data from open sources.*

Four detailed research questions are identified and used to address the thesis statement hypothesis. These questions enable a comprehensive and thorough analysis, exploration of the topic from different angles, and in-depth understanding of the subject matter.

The first research question: R1 – *Which assessment approaches, methods, research challenges, and gaps are present or have been identified in manual veracity assessment of open-source data?* is mainly investigated in the work described in the first publication [70]. Based on an empirical study of veracity assessment within the military domain, the study concludes that there are several challenges in the manual veracity assessment process. These challenges include issues resulting from the analyst's perceived lack of time to do veracity assessment, the subjective nature of the assessments, and the ambiguity and fuzziness of the assessment scale and terminology. The study also identifies gaps in the process, such as the absence of traceability in assessments, which means that there is often no record of the sources, data, and methods used for the assessment. To ensure that the ranking scale used in assessments is well-defined and unambiguous, it is important that what has been assessed and how the assessment should be interpreted is transparent to the end user. Furthermore, the reliability of sources is a significant concern. For a source to be considered reliable, it needs to be perceived as objective and should have produced similar data over a long period. Access to data and motivation are also essential factors to consider when assessing a source's capacity to provide reliable data.

The research described in the second publication [68] explored the current state-of-the-art in automated veracity assessment of data from open sources, addressing **the second research question**: R2 – *How can the veracity assessment of data from open sources be automated in part or in whole?* The main conclusions from that study indicate the presence of three main directions for veracity assessment automation approaches, i.e., utilizing implicit features, employing explicit fact-checking, and the appeal to authority approach. The implicit features approach is based on the idea that it is possible to identify some differences between veridical and non-veridical claims. The difference may be found not only by analyzing the claim itself but also by looking beyond it to other aspects that may indicate the level of veracity. The second approach uses external data to support or refute a claim through comparison with existing knowledge. The third approach, i.e., appeal to authority, is based on the idea that a claim can be considered trustworthy if an authoritative source can independently verify it.

In addition, the gap analysis in the second publication [68] reveales a general need for more approaches and methods that can be adapted to multiple sources and data types, reproducibility challenges, and under-utilization of recent advancements in machine learning. The need for approaches adapted to multiple sources and data types is especially relevant, particularly with the emergence of multimodal systems. Furthermore, a need for more efforts to target online data streams is revealed.

The third research question: R3 – *What features and indicators can be extracted, estimated, and used to automate veracity assessment of open-source data?*,

is addressed using several approaches, resulting in multiple publications. In the first publication [70], three main metrics that can be used for veracity assessment are identified, i.e., content, metadata, and rating-based data.

A conclusion from the second publication [68] is that the research converges towards three main veracity assessment approaches, where the one most commonly used focuses on utilizing implicit features, meaning both features and indicators. In the first publication, following the lines of appeal to authority and analogical thinking, a similarity-based framework is proposed to automate veracity assessment. A suggestion of using a combination of topic detection and trust methods is given, and in the research presented in the third publication, topic detection as an indicator is explored in more detail.

The conclusions from the work presented in the third publication [75] demonstrate that latent Dirichlet allocation is an effective model for tracking geographical trends and topical locations in streaming data (the work focuses on assessing streaming social media). Also, by evaluating their textual context, conditional random fields[1] are found to partially solve the problem of geographical disambiguation, i.e., predicting if Paris refers to a place or a person. However, the disambiguation of different places with the same name remains a topic for future work.

In the research presented in the fourth publication [72], another type of indicator, i.e., stance detection, is explored. The conclusions from that work emphasize the importance of feature engineering and that neural network-based methods are viable methods to process large amounts of data. The work employs convolutional neural networks, but other researchers have obtained better results using long short-term memory networks. As to the research question of what features and indicators can be used as input for automated veracity assessment, the work described in this dissertation has successfully explored some features and indicators that show promise and have been suggested by other researchers in the field. The question, however, remains open-ended.

The fourth research question: R4 – *To what degree do specific features and indicators work for veracity assessment automation?*, is mainly explored in the work presented in the final paper [69]. The explored domain is short reviews like those found in online forums about products, movies, and food chains. Several feature types and preprocessing methods are explored, e.g., TF–IDF, word2vec, PCFG, and even translation. The study results show that the best feature types are TF–IDF and PCFG, suggesting that feature types like TF–IDF and PCFG use the word context to a greater extent than other feature types and that the context is important to detect fake reviews. The experiments also show that it is possible to distinguish between true and false reviews, regardless of whether they are positive or negative. The results are very promising for models trained in one domain, but perhaps not so surprising, testing the models with reviews from another domain

[1] Conditional random fields (CRF) are used for predicting labels while considering the context. For example, in natural language processing (NLP), it is of interest to tag words with their grammatical structure or recognize named entities.

does not yield uplifting results. This leaves the issue of assessing the veracity of open-source data when no previous or similar data is available for training a model. Other types of indicators must then be used to assess the veracity.

The main contributions of the work presented in this dissertation can be summarized as follows: i) a compilation of challenges with manual methods of veracity assessment, ii) a road map for addressing the identified challenges, iii) identification of the state-of-the-art and gap analysis of veracity assessment of open-source data, iv) exploration of indicators such as topic geo-location tracking over time and stance classification, and v) evaluation of various feature types, model transferability, and style obfuscation attacks and the impact on accuracy for automated veracity assessment of a type of deception: fake reviews. The research results on approaches, methods, algorithms, features, indicators, and open-source data show that automated veracity assessment is viable. Indicators may be extracted and calculated directly from the data, meta-data, or other related data. The research results also suggest that automated methods can handle and assess the veracity of large volumes of open-source data. The research presented in this dissertation uses natural language processing, machine learning, deep learning, opinion mining, and topic detection methods for automated veracity assessment of open-source data.

4.2 Legal and ethical considerations

The European union general data protection regulation (GDPR)[2] governs how the personal data of individuals in the EU may be processed and transferred. It is a regulation that the author of the dissertation believes impacts the veracity domain the most. The regulation has several essential parts, such as consent, personal data, privacy impact assessment, and the right to be informed. Naturally, countries and organizations such as the European Union may also have other relevant regulations that the dissertation author is unaware of.

According to GDPR, processing personal data is generally prohibited unless allowed by law or the data subject has consented to the processing. In addition to consent, five other bases are mentioned in Article 6(1) of the GDPR. Of these five; *vital interests of the data subject, public interest, and legitimate interest* are deemed the most relevant to the dissertation topic. However, the legal assessment of their applicability to the automated veracity assessment task is something we have yet to study and is thus left as an open task.

As per the 2024 EU AI Act,[3] which, among other things, discusses transparency obligations for providers and users of certain AI systems and generative AI models, users must be transparent in their creation and dissemination of deepfakes. This means that anyone who creates or shares a deepfake must disclose its source and provide details about the techniques used. The aim of this requirement is to provide consumers with information about the content they are viewing and to reduce the

[2] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.
[3] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

risk of manipulation. However, individuals with malicious intent are unlikely to reveal the use of generative models for deepfakes or generated texts, and therefore, detection approaches and methods are still required.

Due to the inherent purpose of obtaining automatic data assessments and, by extension, their sources, ethical considerations must also be considered when developing such systems. Depending on which normative ethical theory one follows, e.g., *deontology* vs. *consequentialism*, one needs to consider whether the morality of our actions lies in the actions themselves or in the consequences of such actions. For example, is the action of assessing a source justified if the main purpose is assessing the veracity of a rumor? It is important to consider that the result of such an assessment can be uncertain and have low confidence. Which might lead to the source being unfairly deemed untrustworthy and potentially shut down. On the other hand, sources that consistently spread misinformation and disinformation can pose a significant threat [98, 174]. Studies also show that people are much more likely to spread hateful messages, misinformation, and disinformation than true stories [14, 227, 231].

With the advancements in artificial intelligence, it has become easier to generate data of all modalities. For instance, there are software programs that can generate text or clone voices, suites designed to manage multiple social media accounts, and downloadable apps that can generate large amounts of texts [175, 211], create deepfake images and videos from a simple photo or written description [4, 40, 150]. In the best-case scenario, spreading false data is unintentional and may result from honest mistakes or misunderstandings. In the worst-case scenario, it may be directed at an individual to hurt them, such as through blackmailing them or ruining their reputation [132, 152]. As a result, researchers have started working on deepfake detection methods to combat this issue [109, 136, 188]. A risk is that nobody believes anything or anyone any longer. Alternatively, the lie is believed, and the recipients do not accept the proof that it is a lie or false information.

Lately, there has been much criticism regarding the big social media platforms' handling of polarizing and illegal content and their management of accounts spreading misinformation and hateful messages [174, 186]. Another problem is the bias of efforts put into combating disinformation. For example, according to the Facebook whistle-blower Frances Haugen, only nine percent of Facebook's users live in English-speaking countries, but 87% of the company's budget for combating disinformation is placed on English-speaking areas [76]. More efforts must also be made to combat misinformation in other languages.

AI systems are susceptible to discrimination based on gender or race, which poses an ethical challenge. This is because systems are designed to classify content and users with sensitive labels without their knowledge. For instance, job applicants and health care recipients have been affected by such failings [42, 160, 198]. Due to these issues, various ethical guidelines have been compiled for AI systems. For example, the European Commission released 2018 their ethical guidelines [53] for trustworthy AI systems. The guidelines have seven key requirements for the development, deployment, and use of AI systems, i.e., i) human agency and oversight, ii)

technical robustness and safety, iii) privacy and data governance, iv) transparency, v) diversity, non-discrimination, and fairness, vi) environmental and societal well-being and vii) accountability. Other actors, such as large international companies, universities, and researchers, have also had similar initiatives and published their ethics guidelines for AI and social media-related work [171, 212, 236]. Some reoccurring topics are, e.g., *consent, privacy, avoiding bias, and looking to the societal well-being.*

In the research presented in this dissertation, the legal and ethical concerns for the veracity assessment domain of other researchers are studied in publication P2 – Veracity assessment of online data [68]. However, this topic had minimal discussion and is marked as an understudied area.

4.3 Validity and reliability

Open-source data comes in many shapes and forms. The work presented in this dissertation has mainly employed data from micro social media services or open benchmark datasets with unstructured textual data. The veracity assessment research focuses on situations where there are no known facts against which to compare results, and the topics are inclined toward subjective opinions. To ensure the ecological validity of the research, domain experts have been consulted, including the interviews and questionnaires used in the work of the first publication and the related research study presented in the second publication.

This compilation thesis is comprised of five peer-reviewed publications. The research explores different aspects of automated veracity assessment and is presented in the publications of the following types: (two) journals, (two) conferences, and (one) workshop contributions. The peer review of the included publications ensures the validity of this dissertation. Also, the research is based on well-known and established methods, i.e., a combination of exploratory and empirical research, see Section 1.6, which have been thoroughly described in the papers. Experiments are conducted and evaluated rigorously, and when necessary, experiment results are evaluated against benchmarks. Therefore, conclusions follow from the analyzed research results. In the structured literature survey, the method employed and the results are transparently described to ensure reliability.

4.4 Dissertation scope

Placing this dissertation's work into context allows us to find the scope and point out areas for future work. García Lozano *et al.* identified three main types of research approaches [68], i.e., utilizing *implicit features*, employing *explicit fact checking*, and the *appeal to authority*, see Section 3.2. Different methods for knowledge elicitation from open-source intelligence are discussed by Pastor-Galindo *et al.* [163]. The methods are *correlation, classification, outlier detection, clustering, regression, and tracking patterns.* Correlation methods are used to detect relationships between

pieces of data. In classification methods, the data is categorized into predefined classes. Outlier detection is used to find anomalies in the data. Clustering is used to divide the data into related groups or categories. Regression is used to predict values or outcomes from the data. Tracking patterns are used to detect patterns and regularities in the data. Combining the two to form a simple matrix of the research domain and placing the dissertation publications into that context is one way to do a scope analysis of the dissertation's contributions and obtain the areas suitable for future work, see Table 4.1.

Table 4.1: A scope analysis by placing the dissertation publications into the context of open-source data knowledge elicitation methods vs. research approaches.

Research approaches / Knowledge elicitation	Implicit features	Explicit fact checking	Appeal to authority
Correlation	P5		
Classification	P4, P5		
Outlier detection			
Clustering	P4		
Regression			
Tracking patterns	P3	P3	P1, P2

What can be observed in Table 4.1 is that the dissertation contributions primarily lie in the *implicit features* (three publications) and *appeal to authority* (two publications) research directions. Also, the publications are mostly distributed along the *tracking patterns* (three publications) and *classification* (two publications) knowledge elicitation tiers. Little research has been done within this dissertation's scope in areas such as the *outlier detection* and *regression* tiers, and they remain challenges for future work.

4.5 Discussion and future work

This dissertation has delved into the topic of automating the veracity assessment of open-source data and explores viable approaches to achieve this. The research shows that using features and indicators is of utmost importance in automating veracity assessment. Detecting and assessing problematic content, rumors, and news early when working with open-source data is critical to ensure reliable data. Analysis of state-of-the-art and conducted research demonstrates that pattern matching and similarity are recurring themes that play a crucial role in approaches and methods for automating the veracity assessment of open-source data. The research conducted with feature engineering extensively utilizes these themes to create reliable and accurate automated methods. It is necessary to note that the studied indicators serve a different purpose than the features. These indicators can be used as a

basis for logical reasoning in determining the veracity of the data, which is crucial for achieving high confidence in the assessments.

The emergence of generative artificial intelligence models has provided a natural venue for future work. One identified challenge is that these generative techniques can be used to self-feed, i.e., one model creates text that is then used as input to train another model, a second version of the first model, and so on. This can produce self-oscillation situations resulting in problems such as false rumors [224]. Luccioni *et al.* found that bias and stereotypes are amplified in diffusion models [129]. Approaches and methods to combat this are needed. Another major challenge with these models is that researchers have managed to locate and edit factual associations in LLMs, and even after modifications, the models can still pass tests [44, 142, 143], which raises concerns about their reliability. While this capability has the potential for good in correcting false associations, it also has the potential for abuse, such as adding disinformation, bias, or other adversarial data to the model.

A 2024 Stanford report [135] discusses that there are no standardized evaluations for LLMs, which raises questions about their reliability in critical settings. However, it's worth noting that even encyclopedias (printed and online) contain errors [92, 203]. As the popularity of these models increases and more data is encoded and generated, better approaches and methods for automatic veracity assessment, e.g., explicit fact-checking, need to be further researched.

The challenges for explicit fact-checking are multiple, from automatic data encoding into knowledge graphs and the fusion of multimodal data to detecting bias, contradiction, and disinformation. Also, automated methods for the detection of fact-checking worthy claims are needed.

Other, more qualitative areas related to veracity assessment can also be considered venues for future work. These areas involve investigating the various aspects of how veracity assessments can be expressed, communicated, or visualized, the psychological impact a veracity assessment value might have on the receiver, and determining the most effective ways to communicate these values in an easily understood and accepted manner. For example, different visualization techniques can be utilized to highlight dubious text passages and communicate veracity assessments more effectively.

Another area for future work is determining how and when to combine veracity assessments. This process involves merging multiple veracity assessments from different sources or using various methods to arrive at a more accurate overall assessment. The challenge lies in finding the most effective ways to combine veracity assessments (veracity assessment fusion), as well as determining how to assign weights to the assessments from different methods in order to achieve a more precise result or maximize the confidence value.

Another potential area of future research is the *time* aspects of veracity assessments. This includes the duration for which a veracity assessment is valid and under what conditions it remains accurate. Furthermore, it includes the impact

that time has on veracity assessments, such as how the accuracy of an assessment changes over time and how to account for this in the assessment process.

To conclude, based on the scope analysis, the obtained results of the feature and indicator experiments, and the recent advances made with deep learning, several potential directions for future work exist, and automated veracity assessment will remain a highly relevant domain to continue exploring.

Bibliography

[1] M.-A. Abbasi and H. Liu, "Measuring user credibility in social media," in *Social Computing, Behavioral-Cultural Modeling and Prediction*, A. M. Greenberg, W. G. Kennedy, and N. D. Bos, Eds., Berlin, Heidelberg: Springer, 2013, pp. 441–448. DOI: 10.1007/978-3-642-37210-0_48.

[2] A. Abdul-Rahman and S. Hailes, "Supporting trust in virtual communities," in *Proceedings of the 33rd annual Hawaii international conference on system sciences*, IEEE, 2000, 9–pp. DOI: 10.1109/HICSS.2000.926814.

[3] B. Agarwal, A. Ravikumar, and S. Saha, "A novel approach to big data veracity using crowdsourcing techniques and bayesian predictors," in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2016, pp. 1020–1023. DOI: 10.1109/ICMLA.2016.0184.

[4] Agence France-Presse in Shanghai, "Chinese deepfake app Zao sparks privacy row after going viral," *The Guardian*, no. 02, Sep. 2019. [Online]. Available: https://www.theguardian.com/technology/2019/sep/02/chinese-face-swap-app-zao-triggers-privacy-fears-viral (visited on 04/24/2024).

[5] H. Aghakhani, A. Machiry, S. Nilizadeh, C. Kruegel, and G. Vigna, "Detecting deceptive reviews using generative adversarial networks," in *2018 IEEE Security and Privacy Workshops (SPW)*, IEEE, 2018, pp. 89–95. DOI: 10.1109/SPW.2018.00022.

[6] E. Aïmeur, S. Amri, and G. Brassard, "Fake news, disinformation and misinformation in social media: A review," *Social Network Analysis and Mining*, vol. 13, no. 1, p. 30, 2023. DOI: 10.1007/s13278-023-01028-5.

[7] R. Alkhalifa and A. Zubiaga, "Capturing stance dynamics in social media: Open challenges and research directions," 2021. arXiv: 2109.00475.

[8] S. A. Alkhodair, S. H. Ding, B. C. Fung, and J. Liu, "Detecting breaking news rumors of emerging topics in social media," *Information Processing & Management*, vol. 57, no. 2, p. 102 018, 2020. DOI: 10.1016/j.ipm.2019.02.016.

[9] M. Alrubaian, M. Al-Qurishi, M. M. Hassan, and A. Alamri, "A credibility analysis system for assessing information on Twitter," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 661–674, 2016. DOI: 10.1109/TDSC.2016.2602338.

[10] L. AlSumait, D. Barbará, and C. Domeniconi, "On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking," in *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, IEEE, 2008, pp. 3–12. DOI: 10.1109/ICDM.2008.140.

[11] "Assessment and Communication of Uncertainty in Intelligence to Support Decision-Making," STO Technical Report TR-SAS-114, Jun. 2020, p. 384.

[12] F. Assiri, "Methods for Assessing, Predicting, and Improving Data Veracity: A survey," *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, vol. 9, no. 4, pp. 5–30, 2020. DOI: 10.14201/ADCAIJ202094530.

[13] P. Avesani, P. Massa, and R. Tiella, "A trust-enhanced recommender system application: Moleskiing," in *Proceedings of the 2005 ACM symposium on Applied computing*, ACM, 2005, pp. 1589–1593.

[14] M. Avram, N. Micallef, S. Patil, and F. Menczer, "Exposure to social engagement metrics increases vulnerability to misinformation," 2020. arXiv: 2005.04682.

[15] R. Bar-Haim, I. Bhattacharya, F. Dinuzzo, A. Saha, and N. 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*, 2017, pp. 251–261.

[16] D. Bates. "Samsung must pay $340,000 after paying people to write bad reviews," Mail Online. (2013), [Online]. Available: www.dailymail.co.uk/sciencetech/article-2476630/Samsung-ordered-pay-340-000-paid-people-write-negative-online-reviews-HTC-phones.html (visited on 04/15/2022).

[17] A. J. Berinsky, "Rumors and Health Care Reform: Experiments in Political Misinformation," *British Journal of Political Science*, vol. 47, no. 2, pp. 241–262, 2017. DOI: 10.1017/S0007123415000186.

[18] L. Berti-Équille and J. Borge-Holthoefer, "Veracity of Data: From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics," *Synthesis Lectures on Data Management*, vol. 7, no. 3, pp. 1–155, 2015. DOI: 10.2200/S00676ED1V01Y201509DTM042.

[19] J. Besombes, L. Cholvy, and V. Dragos, "A semantic-based model to assess information for intelligence," *AerospaceLab*, 2012.

[20] J. Besombes and A. d'Allonnes, "An extension of STANAG 2022 for information scoring," in *Information Fusion, 2008 11th International Conference on*, IEEE, 2008, pp. 1–7.

[21] C. Best, "Challenges in open source intelligence," in *2011 European Intelligence and Security Informatics Conference (EISIC)*, IEEE, Sep. 2011, pp. 58–62. DOI: 10.1109/EISIC.2011.41.

[22] R. A. Best Jr and A. Cumming, "Open source intelligence (OSINT): Issues for congress," Congressional Research Service, Library of Congress, Washington, DC, USA, Rep. ADA488690, Dec. 2007.

[23] C. M. Bishop, Ed., *Pattern recognition and machine learning*. New York, USA: Springer-Verlag, 2006. DOI: 10.1007/978-0-387-45528-0.

[24] C. Bizer and R. Cyganiak, "Quality-driven information filtering using the WIQA policy framework," *Journal of Web Semantics*, vol. 7, no. 1, pp. 1–10, 2009. DOI: 10.1016/j.websem.2008.02.005.

[25] E. Blasch, K. B. Laskey, A.-L. Jousselme, V. Dragos, P. C. Costa, and J. Dezert, "URREF reliability versus credibility in information fusion (STANAG 2511)," in *Proceedings of the 16th International Conference on Information Fusion*, IEEE, 2013, pp. 1600–1607.

[26] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[27] C. F. Bond Jr and B. M. DePaulo, "Accuracy of deception judgments," *Personality and social psychology Review*, vol. 10, no. 3, pp. 214–234, 2006. DOI: 10.1207/s15327957pspr1003_2.

[28] S. Bradshaw, H. Bailey, and P. N. Howard, "Industrialized disinformation: 2020 global inventory of organized social media manipulation," Oxford Internet Institute, Tech. Rep., 2021.

[29] C. Buntain and J. Golbeck, "Automatically Identifying Fake News in Popular Twitter Threads," in *2017 IEEE International Conference on Smart Cloud (SmartCloud)*, New York, NY: IEEE, 2017, pp. 208–215. DOI: 10.1109/SmartCloud.2017.40.

[30] G. Cai, H. Wu, and R. Lv, "Rumors detection in Chinese via crowd responses," in *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, China: IEEE, 2014, pp. 912–917. DOI: 10.1109/ASONAM.2014.6921694.

[31] J. Camacho, G. Macia-Fernandez, J. Diaz-Verdejo, and P. Garcia-Teodoro, "Tackling the Big Data 4 vs for anomaly detection," in *2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Toronto, ON, Canada: IEEE, 2014, pp. 500–505. DOI: 10.1109/INFCOMW.2014.6849282.

[32] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th International Conference on World Wide Web - WWW '11*, Hyderabad, India: ACM Press, 2011, p. 675. DOI: 10.1145/1963405.1963500.

[33] A. Chen, "The Agency," *The New York TimesMagazine*, 2015. [Online]. Available: https://www.nytimes.com/2015/06/07/magazine/the-agency.html (visited on 04/17/2022).

[34] T. Chen, X. Li, H. Yin, and J. Zhang, "Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection," in *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2018 Workshops, BDASC, BDM, ML4Cyber, PAISI, DaMEMO, Melbourne, VIC, Australia, June 3, 2018, Revised Selected Papers 22*, Springer, 2018, pp. 40–52.

[35] M. Chenine, V. Kabilan, and M. García Lozano, "A pattern for designing distributed heterogeneous ontologies for facilitating application interoperability," in *International Conference on Advanced Information Systems Engineering*, CEUR-WS.org, 2006.

[36] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014. arXiv: 1406.1078.

[37] L. Cholvy, "Information evaluation in fusion: A case study," in *Proceedings of the conference IPMU 2004*, Citeseer, 2004.

[38] L. Cholvy and V. Nimier, "Information evaluation: Discussion about STANAG 2022 recommendations," DTIC Document, Tech. Rep., 2004.

[39] I. Claverie-Berge, *Solutions big data IBM*, presentation slides, Mar. 2012.

[40] S. Cole, "This horrifying app undresses a photo of any woman with a single click," *Vice*, no. 27, p. 06, 2019. [Online]. Available: https://www.vice.com/en/article/kzm59x/deepnude-app-creates-fake-nudes-of-any-woman (visited on 04/24/2024).

[41] N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," in *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, ser. ASIST '15, St. Louis, Missouri, USA: American Society for Information Science, 2015, 82:1–82:4. DOI: 10.1002/pra2.2015.145052010082.

[42] J. Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," *Reuters*, 2018. [Online]. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[43] M. Davoudi, M. R. Moosavi, and M. H. Sadreddini, "DSS: A hybrid deep model for fake news detection using propagation tree and stance network," *Expert Systems with Applications*, vol. 198, p. 116 635, 2022.

[44] N. De Cao, W. Aziz, and I. Titov, "Editing factual knowledge in language models," 2021. arXiv: 2104.08164.

[45] J. Debattista, C. Lange, S. Scerri, and S. Auer, "Linked 'big' data: Towards a manifold increase in big data value and veracity," *2015 IEEE/ACM 2nd International Symposium on Big Data Computing (BDC)*, pp. 92–98, 2015. DOI: 10.1109/BDC.2015.34.

[46] T. Delavallade and P. Capet, "Information evaluation as a decision support for counter-terrorism," in *NATO symposium on C3I in Crisis, Emergency and Consequence Management, IST*, vol. 86, 2009.

[47] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. Wong Sak Hoi, and A. Zubiaga, "SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours," in *Proceedings of SemEval*, Association for Computational Linguistics, 2017. DOI: 10.18653/v1/S17-2006.

[48] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2019. arXiv: 1810.04805.

[49] "Digital 2022: Global Overview Report." in collab. with S. Kemp, DataReportal. (Jan. 26, 2022), [Online]. Available: https://datareportal.com/reports/digital-2022-global-overview-report (visited on 04/09/2022).

[50] V. Dragos, "Shallow semantic analysis to estimate HUMINT correlation," in *Information Fusion (FUSION), 2012 15th International Conference on*, IEEE, 2012, pp. 2293–2300.

[51] V. Dragos and K. Rein, "Integration of soft data for information fusion: Pitfalls, challenges and trends," in *17th International Conference on Information Fusion (FUSION)*, IEEE, 2014, pp. 1–8.

[52] J. Enoch. "Can you trust online reviews? Here's how to find the fakes," NBC News. (Feb. 27, 2019), [Online]. Available: https://www.nbcnews.com/business/consumer/can-you-trust-online-reviews-here-s-how-find-fakes-n976756 (visited on 04/16/2022).

[53] "Ethics guidelines for trustworthy AI," 2018. [Online]. Available: https://ec.europa.eu/futurium/en/ai-alliance-consultation.

[54] *Fake news*, in *Wikipedia*, Apr. 7, 2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Fake_news&oldid=1081403379 (visited on 04/09/2022).

[55] Y. Fang, H. Wang, L. Zhao, F. Yu, and C. Wang, "Dynamic knowledge graph based fake-review detection," *Applied Intelligence*, vol. 50, pp. 4281–4295, 2020.

[56] R. M. Faris, R. Hal, E. Bruce, B. Nikki, Z. Ethan, and B. Yochai, "Partisanship, propaganda, and disinformation: Online media and the 2016 U.S. presidential election," Berkman Klein Center for Internet & Society Research Paper, 2017.

[57] *Feature (machine learning)*, in *Wikipedia*, Nov. 25, 2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Feature_(machine_learning)&oldid=1057151262 (visited on 04/17/2022).

[58] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, Association for Computational Linguistics, 2012, pp. 171–175.

[59] W. Ferreira and A. 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*, Association for Computational Linguistics, 2016.

[60] "Field Manual FM 30-5, Combat Intelligence," U.S. Department of the Army, Washington D.C., USA, 1963.

[61] L. Floridi and P. Illari, Eds., *The Philosophy of Information Quality* (Synthese Library). Cham: Springer, 2014, vol. 358. DOI: 10.1007/978-3-319-07121-3.

[62] T. J. Froehlich, "A not-so-brief account of current information ethics: The ethics of ignorance, missing information, misinformation, disinformation and other forms of deception or incompetence," *BiD*, vol. 39, p. 14, 2017.

[63] A. Galland, S. Abiteboul, A. Marian, and P. Senellart, "Corroborating information from disagreeing views," in *Proceedings of the third ACM international conference on Web search and data mining*, 2010, pp. 131–140.

[64] D. Gambetta *et al.*, "Can we trust trust," *Trust: Making and breaking cooperative relations*, vol. 13, pp. 213–237, 2000.

[65] A. Gani, "Amazon sues 1,000 'fake reviewers'," *The GuardianTechnology*, Oct. 18, 2015. [Online]. Available: https://www.theguardian.com/technology/2015/oct/18/amazon-sues-1000-fake-reviewers (visited on 04/15/2022).

[66] M. García Lozano, "Semantic based resource identification, storage and discovery in distributed systems," Licentiate thesis, KTH, 2010. [Online]. Available: https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-27340.

[67] M. García Lozano, "Trusting open source information," in *2013 European Intelligence and Security Informatics Conference (EISIC)*, IEEE, 2013, pp. 228–228. DOI: 10.1109/EISIC.2013.77.

[68] M. García Lozano, J. Brynielsson, U. Franke, M. Rosell, E. Tjörnhammar, S. Varga, and V. Vlassov, "Veracity assessment of online data," *Decision Support Systems*, vol. 129, no. 113132, 2020. DOI: 10.1016/j.dss.2019.113132.

[69] M. García Lozano and J. Fernquist, "Identifying deceptive reviews: Feature exploration, model transferability and classification attack," in *2019 European Intelligence and Security Informatics Conference (EISIC)*, IEEE, Nov. 2019, pp. 228–228. DOI: 10.1109/EISIC49498.2019.9108852.

[70] M. García Lozano, U. Franke, M. Rosell, and V. Vlassov, "Towards automatic veracity assessment of open source information," in *2015 IEEE International Congress on Big Data*, IEEE, 2015, pp. 199–206. DOI: 10.1109/BigDataCongress.2015.36.

[71] M. García Lozano, P. Hörling, F. Moradi, and E. Tjörnhammar, "Supporting C2 with a service oriented framework for opportunistic sensors and sensor networks," in *International Command and Control Research and Technology Symposium 14th ICCRTS:"C2 and Agility"*, 2009.

[72] M. García Lozano, H. Lilja, E. Tjörnhammar, and M. Karasalo, "Mama Edha at SemEval-2017 Task 8: Stance classification with CNN and rules," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 481–485. DOI: 10.18653/v1/S17-2084.

[73] M. García Lozano, F. Moradi, and R. Ayani, "SDR: A semantic based distributed repository for simulation models and resources," in *First Asia International Conference on Modelling Simulation (AMS'07)*, 2007, pp. 171–176.

[74] M. García Lozano, F. Moradi, E. Ibarzabal, and E. Tjörnhammar, "A semantic approach to simulation component identification and discovery," in *EMSS 2009, 21st European Modeling and Simulation Symposium*, Sep. 2009, pp. 181–186.

[75] M. García Lozano, J. Schreiber, and J. Brynielsson, "Tracking geographical locations using a geo-aware topic model for analyzing social media data," *Decision Support Systems*, vol. 99, pp. 18–29, 2017. DOI: 10.1016/j.dss.2017.05.006.

[76] M. Gelin, "Maria Ressa: Facebook är som gödsel för demokratins kollaps," *Dagens Nyheter*, 2021.

[77] Y. Gil and V. Ratnakar, "Trusting information sources one citizen at a time," in *International Semantic Web Conference*, Springer, 2002, pp. 162–176.

[78] A. L. Ginsca, A. Popescu, M. Lupu, *et al.*, "Credibility in information retrieval," *Foundations and Trends® in Information Retrieval*, vol. 9, no. 5, pp. 355–475, 2015. DOI: 10.1561/1500000046.

[79] J. Golbeck and J. Hendler, "Accuracy of metrics for inferring trust and reputation in semantic web-based social networks," in *Engineering knowledge in the age of the semantic web*, Springer, 2004, pp. 116–131.

[80] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[81] Google, *Google trends fakenews interest image*. [Online]. Available: https://trends.google.com (visited on 04/09/2022).

[82] J. Gottfried and E. Shearer, "News use across social media platforms 2016," Pew Research Center, Washington, D.C., USA, Rep. May 2016.

[83] A. M. Guess and B. A. Lyons, "Misinformation, disinformation, and online propaganda," *Social media and democracy: The state of the field, prospects for reform*, pp. 10–33, 2020.

[84] H. Guo, J. Cao, Y. Zhang, J. Guo, and J. Li, "Rumor detection with hierarchical social attention network," in *Proceedings of the 27th ACM international conference on information and knowledge management*, 2018, pp. 943–951.

[85] A. Gupta, P. Kumaraguru, C. Castillo, P. Meier, *et al.*, "TweetCred: A real-time web-based system for assessing credibility of content on twitter," 2014. arXiv: 1405.5490.

[86] P. Hansen, M. García Lozano, F. Kamrani, and J. Brynielsson, "Real-time estimation of heart rate in situations characterized by dynamic illumination using remote photoplethysmography," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6094–6103. DOI: 10.1109/CVPRW59228.2023.00649.

[87] K. Hao, *Deepfake "Amazon workers" are sowing confusion on Twitter. that's not the problem*, 2021. [Online]. Available: https://www.technologyreview.com/2021/03/31/1021487/deepfake-amazon-workers-are-sowing-confusion-on-twitter-thats-not-the-problem/ (visited on 04/24/2024).

[88] M. Harpalani, M. Hart, S. Signh, R. Johnson, and Y. Choi, "Language of vandalism: Improving Wikipedia vandalism detection via stylometric analysis," p. 6, Jun. 19–24, 2011.

[89] K. S. Hasan and V. Ng, "Stance classification of ideological debates: Data, models, features, and constraints," in *Proceedings of the sixth international joint conference on natural language processing*, 2013, pp. 1348–1356.

[90] S. He, B. Hollenbeck, and D. Proserpio, "The market for fake reviews," *Marketing Science*, vol. 41, no. 5, pp. 896–921, 2022. DOI: 10.2139/ssrn.3664992.

[91] A. Hermida, "Twittering the news: The emergence of ambient journalism," *Journalism Practice*, vol. 4, no. 3, pp. 297–308, Aug. 2010. DOI: 10.1080/17512781003640703.

[92] L. Holman Rector, "Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles," *Reference services review*, vol. 36, no. 1, pp. 7–22, 2008.

[93] R. A. Igawa, S. B. Jr, K. C. S. Paulo, G. S. Kido, R. C. Guido, M. L. P. Júnior, and I. N. da Silva, "Account classification in online social networks with LBCA and wavelets," *Information Sciences*, vol. 332, pp. 72–83, 2016, ISSN: 0020-0255. DOI: 10.1016/j.ins.2015.10.039.

[94] C. Ireton and J. Posetti, Eds., *Journalism, Fake News & Disinformation: Handbook for Journalism Education and Training*. Paris, France: United Nations Educational, Scientific and Cultural Organization, 2018.

[95] M. R. Islam, S. Liu, X. Wang, and G. Xu, "Deep learning for misinformation detection on online social networks: A survey and new perspectives," *Social Network Analysis and Mining*, vol. 10, no. 1, pp. 1–20, 2020.

[96] M. E. Jaeger, S. Anthony, and R. L. Rosnow, "Who hears what from whom and with what effect: A study of rumor," *Personality and Social Psychology Bulletin*, vol. 6, no. 3, pp. 473–478, 1980. DOI: 10.1177/014616728063024.

[97] S. Jain, V. Sharma, and R. Kaushal, "Towards automated real-time detection of misinformation on Twitter," in *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Sep. 2016, pp. 2015–2020. DOI: 10.1109/ICACCI.2016.7732347.

[98] J. Jaiswal, C. LoSchiavo, and D. Perlman, "Disinformation, misinformation and inequality-driven mistrust in the time of COVID-19: Lessons unlearned from AIDS denialism," *AIDS and Behavior*, vol. 24, pp. 2776–2780, 2020.

[99] N. B. C. E. Jamil, I. B. Ishak, F. Sidi, L. S. Affendey, and A. Mamat, "A systematic review on the profiling of digital news portal for big data veracity," *Procedia Computer Science*, vol. 72, pp. 390–397, 2015, The Third Information Systems International Conference 2015. DOI: 10.1016/j.procs.2015.12.154.

[100] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand, *et al.*, "Mixtral of experts," 2024. arXiv: 2401.04088.

[101] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 795–816.

[102] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proceedings of the 2008 international conference on web search and data mining*, 2008, pp. 219–230.

[103] F. Johansson, A. Horndahl, H. Lilja, M. García Lozano, L. Lundmark, and H. Stiff, "Detection of fabricated media," Swedish Defence Research Agency, Stockholm, Sweden, Tech. Rep. FOI-R--5132--SE, Apr. 2021.

[104] F. Johansson, A. Horndahl, H. Stiff, and M. García Lozano, "Data synthesis using generative models," Swedish Defence Research Agency, Stockholm, Sweden, Tech. Rep. FOI-R--5041--SE, Nov. 2020.

[105] F. Johansson, L. Kaati, and A. Shrestha, "Timeprints for identifying social media users with multiple aliases," *Security Informatics*, vol. 4, no. 1, pp. 1–11, 2015. DOI: 10.1186/s13388-015-0022-z.

[106] A. Jøsang, R. Ismail, and C. Boyd, "A survey of trust and reputation systems for online service provision," *Decision support systems*, vol. 43, no. 2, pp. 618–644, 2007.

[107] D. Jurafsky and J. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 1st. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2000.

[108] J. Juran, *Quality control handbook*, 3rd ed. New York, U.S.A.: McGraw-Hill, 1974.

[109] S. Kalf, *What does a feminist approach to deepfake pornography look like*, 2019. [Online]. Available: http://mastersofmedia.hum.uva.nl/blog/2019/10/24/what-does-a-feminist-approach-to-deepfake-pornography-look-like/ (visited on 04/24/2024).

[110] H. Karande, R. Walambe, V. Benjamin, K. Kotecha, and T. Raghu, "Stance detection with BERT embeddings for credibility analysis of information on social media," *PeerJ Computer Science*, vol. 7, e467, Apr. 14, 2021. DOI: 10.7717/peerj-cs.467.

[111] E. Kauffmann, J. Peral, D. Gil, A. Ferrández, R. Sellers, and H. Mora, "A framework for big data analytics in commercial social networks: A case study on sentiment analysis and fake review detection for marketing decision-making," *Industrial Marketing Management*, vol. 90, pp. 523–537, 2020. DOI: 10.1016/j.indmarman.2019.08.003.

[112] B. A. Kitchenham and S. M. Charters, "Guidelines for performing systematic literature reviews in software engineering, version 2.3," Keele University and Durham University, United Kingdom, Technical Report EBSE-2007-01, Jul. 2007.

[113] T. Knap and I. Mlỳnková, "Towards topic-based trust in social networks," in *Ubiquitous Intelligence and Computing*, Springer, 2010, pp. 635–649.

[114] W. Knight, "This AI makes Robert De Niro perform lines in flawless German," *Wired*, vol. 19, no. 05, 2021.

[115] E. Kochkina, M. Liakata, and I. Augenstein, "Turing at SemEval-2017 Task 8: Sequential Approach to Rumour Stance Classification with Branch-LSTM," Apr. 24, 2017. arXiv: 1704.07221.

[116] D. Küçük and F. Can, "Stance Detection: A Survey," *ACM Computing Surveys*, vol. 53, no. 1, pp. 1–37, Jan. 31, 2021. DOI: 10.1145/3369026.

[117] S. Kumar, R. West, and J. Leskovec, "Disinformation on the web: Impact, characteristics, and detection of Wikipedia hoaxes," in *Proceedings of the 25th International Conference on World Wide Web*, Montreal, Quebec, Canada: ACM, 2016, pp. 591–602. DOI: 10.1145/2872427.2883085.

[118] U. Kuter and J. Golbeck, "Sunny: A new algorithm for trust inference in social networks using probabilistic confidence models," in *Proceedings of the national conference on artificial intelligence*, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, vol. 22, 2007, p. 1377.

[119] D. Laney, "3D data management: Controlling data volume, velocity and variety," *Application delivery strategies, File*, vol. 949, 2001.

[120] J. H. Lau, N. Collier, and T. Baldwin, "On-line trend analysis with topic models: #twitter trends detection topic model online," in *COLING*, 2012, pp. 1519–1534.

[121] J. H. Lee, N. Santero, A. Bhattacharya, E. May, and E. S. Spiro, "Community-based strategies for combating misinformation: Learning from a popular culture fandom," *Harvard Kennedy School Misinformation Review*, 2022.

[122] G. Levchuk and E. Blasch, "Probabilistic graphical models for multi-source fusion from text sources," in *2015 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, May 2015, pp. 1–10. DOI: 10.1109/CISDA.2015.7208640.

[123] D. H. Levin, "When the great power gets a vote: The effects of great power electoral interventions on election results," *International Studies Quarterly*, vol. 60, no. 2, pp. 189–202, Feb. 2016. DOI: 10.1093/isq/sqv016.

[124] L. Li, G. Cai, and N. Chen, "A rumor events detection method based on deep bidirectional GRU neural network," in *2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC)*, IEEE, 2018, pp. 755–759. DOI: 10.1109/ICIVC.2018.8492819.

[125] Q. Li, Q. Zhang, and L. Si, "Rumor detection by exploiting user credibility information, attention and multi-task learning," in *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019, pp. 1173–1179. DOI: 10.18653/v1/P19-1113.

[126] G. Liang, W. He, C. Xu, L. Chen, and J. Zeng, "Rumor Identification in Microblogging Systems Based on Users' Behavior," *IEEE Transactions on Computational Social Systems*, vol. 2, no. 3, pp. 99–108, Sep. 2015. DOI: 10.1109/TCSS.2016.2517458.

[127] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. New York, NY: Springer Science & Business Media, Dec. 6, 2012. DOI: 10.1007/978-1-4615-5689-3.

[128] M. Luca and G. Zervas, "Fake it till you make it: Reputation, competition, and Yelp review fraud," *Management science*, vol. 62, no. 12, pp. 3412–3427, 2016. DOI: 10.1287/mnsc.2015.2304.

[129] S. Luccioni, C. Akiki, M. Mitchell, and Y. Jernite, "Stable bias: Evaluating societal representations in diffusion models," in *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. [Online]. Available: https://openreview.net/forum?id=qVXYU3F017.

[130] T. Lukoianova and V. L. Rubin, "Veracity roadmap: Is big data objective, truthful and credible?" *Advances In Classification Research Online*, vol. 24, p. 1, 2014. DOI: 10.7152/acro.v24i1.14671.

[131] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, AAAI Press, 2016.

[132] "Malicious actors almost certainly will leverage synthetic content for cyber and foreign influence operations," Federal Bureau of Investigation, Washington D.C., USA, White paper 210310-001, Mar. 10, 2021. [Online]. Available: https://www.ic3.gov/Media/News/2021/210310-2.pdf (visited on 04/09/2022).

[133] J. Marciano. "Fake online reviews cost $152 billion a year. Here's how e-commerce sites can stop them," World Economic Forum. (Aug. 10, 2021), [Online]. Available: https://www.weforum.org/agenda/2021/08/fake-online-reviews-are-a-152-billion-problem-heres-how-to-silence-them/ (visited on 04/15/2022).

[134] S. P. Marsh, "Formalising trust as a computational concept," Ph.D. dissertation, University of Stirling, Stirling, United Kingdom, Apr. 1994.

[135] N. Maslej, L. Fattorini, R. Perrault, V. Parli, A. Reuel, E. Brynjolfsson, J. Etchemendy, K. Ligett, T. Lyons, J. Manyika, J. C. Niebles, Y. Shoham, R. Wald, and J. Clark, "The AI index 2024 annual report," AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, USA, Tech. Rep., Apr. 2024.

[136] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward," *Applied intelligence*, vol. 53, no. 4, pp. 3974–4026, 2023.

[137] G. McArdle and R. Kitchin, "Improving the veracity of open and real-time urban data," *Built Environment*, vol. 42, no. 3, pp. 457–473, 2016. DOI: 10.2139/ssrn.2643430.

[138] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, pp. 115–133, 1943.

[139] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, pp. 415–444, 2001.

[140] P. Meel and D. K. Vishwakarma, "Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities," *Expert Systems with Applications*, vol. 153, p. 112 986, Sep. 2020. DOI: 10.1016/j.eswa.2019.112986.

[141] M. Mendoza, B. Poblete, and C. Castillo, "Twitter under crisis: Can we trust what we RT?" In *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*, Washington D.C., District of Columbia: ACM Press, 2010, pp. 71–79. DOI: 10.1145/1964858.1964869.

[142] K. Meng, D. Bau, A. Andonian, and Y. Belinkov, "Locating and editing factual associations in GPT," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 359–17 372, 2022.

[143] K. Meng, A. S. Sharma, A. J. Andonian, Y. Belinkov, and D. Bau, "Mass-editing memory in a transformer," in *The Eleventh International Conference on Learning Representations*, 2023.

[144] T. Mikolov, "Statistical language models based on neural networks," Ph.D. dissertation, Brno University of Technology, 2012.

[145] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013. arXiv: 1301.3781.

[146] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[147] M. S. Miron, S. M. Patten, and S. M. Halpin, "The structure of combat intelligence ratings," US Army Research Institute for the Behavioral and Social Sciences, Arlington, Technical paper 286, Sep. 1978, p. 97.

[148] R. Mohawesh, S. Xu, S. N. Tran, R. Ollington, M. Springer, Y. Jararweh, and S. Maqsood, "Fake reviews detection: A survey," *IEEE Access*, vol. 9, pp. 65 771–65 802, 2021. DOI: 10.1109/ACCESS.2021.3075573.

[149] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "Fake review detection: Classification and analysis of real and pseudo reviews," UIC-CS-03-2013. Technical Report, Tech. Rep., 2013.

[150] My heritage, *Animate your family photos*, 2021. [Online]. Available: https://www.myheritage.se/deep-nostalgia (visited on 04/24/2024).

[151] Y. Namihira, N. Segawa, Y. Ikegami, K. Kawai, T. Kawabe, and S. Tsuruta, "High precision credibility analysis of information on twitter," in *2013 International Conference on Signal-Image Technology Internet-Based Systems*, Dec. 2013, pp. 909–915. DOI: 10.1109/SITIS.2013.148.

[152] F. Navarro. "Deepfake videos being used to blackmail people," Komando.com. (Jan. 1, 2019), [Online]. Available: https://www.komando.com/security-privacy/deepfake-porn-videos-are-now-being-used-to-publicly-harass-ordinary-people/526877/ (visited on 04/09/2022).

[153] NCSC Newsroom, *FBI and NCSC release new movie to increase awareness of foreign intelligence threats on professional networking sites and other social media platforms*, Sep. 2020. [Online]. Available: https://www.dni.gov/index.php/ncsc-newsroom/3479-nevernight-press-release (visited on 04/24/2024).

[154] N. Newman, R. Fletcher, K. Eddy, C. T. Robertson, and R. K. Nielsen, "Reuters institute digital news report 2023," Reuters Institute for the Study of Journalism, University of Oxford, Oxford, United Kingdom, Tech. Rep., 2023. DOI: 10.60625/risj-p6es-hb13.

[155] N. Newman, R. Fletcher, A. Schulz, S. Andı, and R. K. Nielsen, "Reuters institute digital news report 2020," Reuters Institute for the Study of Journalism, University of Oxford, Oxford, United Kingdom, Tech. Rep., 2020. DOI: 10.60625/risj-048n-ap07.

[156] T. N. Nguyen, C. Li, and C. Niederée, "On early-stage debunking rumors on Twitter: Leveraging the wisdom of weak learners," in *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part II 9*, Springer, 2017, pp. 141–158.

[157] V. Nimier, "Information evaluation: A formalisation of operational recommendations," in *Fusion 2004: Seventh International Conference on Information Fusion*, 2004.

[158] North Atlantic Treaty Organization, *Standard: NATO - STANAG 2511, intelligence reports - ed 1*, Jan. 2003.

[159] D. E. O'Leary, "Blog mining-review and extensions: From each according to his opinion," *Decision Support Systems*, vol. 51, no. 4, pp. 821–830, 2011. DOI: 10.1016/j.dss.2011.01.016.

[160] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019. DOI: 10.1126/science.aax2342.

[161] M. Ott, C. Cardie, and J. 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 (NAACL2013)*, Association for Computational Linguistics, 2013, pp. 497–501.

[162] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," 2011. arXiv: 1107.4557.

[163] J. Pastor-Galindo, P. Nespoli, F. G. Mármol, and G. M. Pérez, "The not yet exploited goldmine of OSINT: Opportunities, open challenges and future trends," *IEEE Access*, vol. 8, pp. 10 282–10 304, 2020. DOI: 10.1109/ACCESS.2020.2965257.

[164] N. A. Patel and R. Patel, "A survey on fake review detection using machine learning techniques," in *2018 4th international Conference on computing Communication and automation (ICCCA)*, IEEE, 2018, pp. 1–6. DOI: 10.1109/CCAA.2018.8777594.

[165] V. Pendyala, *Veracity of Big Data*. Berkeley, CA: Apress, 2018. DOI: 10.1007/978-1-4842-3633-8.

[166] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: LIWC 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[167] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[168] G. Pennycook, Z. Epstein, M. Mosleh, A. A. Arechar, D. Eckles, and D. G. Rand, "Shifting attention to accuracy can reduce misinformation online," *Nature*, vol. 592, no. 7855, pp. 590–595, Apr. 22, 2021. DOI: 10.1038/s41586-021-03344-2.

[169] G. Pennycook and D. G. Rand, "The psychology of fake news," *Trends in Cognitive Sciences*, vol. 25, no. 5, pp. 388–402, May 2021. DOI: 10.1016/j.tics.2021.02.007.

[170] M. R. E. Peterson, "The Soviet combat intelligence process: An integrative interpretation," U.S. Army Russian Institute APO, NY, USA, Student research Report ADA098541, Jun. 1980, p. 49. [Online]. Available: https://apps.dtic.mil/sti/pdfs/ADA098541.pdf (visited on 04/16/2022).

[171] S. Pichai, "AI at Google: Our principles," 2018. [Online]. Available: https://www.blog.google/technology/ai/ai-principles/.

[172] K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum, "Credibility assessment of textual claims on the web," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, ser. CIKM '16, Indianapolis, Indiana: ACM, 2016, pp. 2173–2178. DOI: 10.1145/2983323.2983661.

[173] J. Prasad, "The psychology of rumour: A study relating to the great Indian earthquake of 1934," *British Journal of Psychology. General Section*, vol. 26, no. 1, pp. 1–15, Jul. 1935. DOI: 10.1111/j.2044-8295.1935.tb00770.x.

[174] N. Purnell and J. Horwitz, *The facebook files: Facebook services are used to spread religious hatred in India, internal documents show*, 2021.

[175] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," Jun. 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (visited on 04/24/2024).

[176] A. Rajaraman and J. Ullman, *Mining of massive datasets*. Cambridge University Press Cambridge, 2012, vol. 1.

[177] S. Ramachandramurthy, S. Subramaniam, and C. Ramasamy, "Distilling big data: Refining quality information in the era of yottabytes," *The Scientific World Journal*, vol. 2015, 2015.

[178] C. Reuter, M.-A. Kaufhold, and R. Steinfort, "Rumors, fake news and social bots in conflicts and emergencies: Towards a model for believability in social media," in *Proceedings of the 14th ISCRAM Conference – Albi, France*, 2017.

[179] J. C. Rodrigues, J. T. Rodrigues, V. L. K. Gonsalves, A. U. Naik, P. Shetgaonkar, and S. Aswale, "Machine & deep learning techniques for detection of fake reviews: A survey," in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, IEEE, 2020, pp. 1–8. DOI: 10.1109/ic-ETITE47903.2020.063.

[180] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695. DOI: 10.1109/CVPR52688.2022.01042.

[181] N. Rosenfeld, A. Szanto, and D. C. Parkes, "A kernel of truth: Determining rumor veracity on Twitter by diffusion pattern alone," *Proceedings of The Web Conference 2020*, pp. 1018–1028, Apr. 20, 2020. DOI: 10.1145/3366423.3380180. arXiv: 2002.00850.

[182] V. L. Rubin, Y. Chen, and N. K. Conroy, "Deception detection for news: Three types of fakes," *Proceedings of the Association for Information Science and Technology*, vol. 52, no. 1, pp. 1–4, Jan. 2015. DOI: 10.1002/pra2.2015.145052010083.

[183] D. Saez-Trumper, "Fake tweet buster: A webtool to identify users promoting fake news on Twitter," in *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, ser. HT '14, Santiago, Chile: ACM, 2014, pp. 316–317, ISBN: 978-1-4503-2954-5. DOI: 10.1145/2631775.2631786.

[184] J. Salminen, C. Kandpal, A. M. Kamel, S.-g. Jung, and B. J. Jansen, "Creating and detecting fake reviews of online products," *Journal of Retailing and Consumer Services*, vol. 64, p. 102 771, 2022. DOI: 10.1016/j.jretconser.2021.102771.

[185] M. G. Samet, "Subjective interpretation of reliability and accuracy scales for evaluating military intelligence," US Army Research Institute for the Behavioral and Social Sciences, AD/A-003 260, Jan. 1975, p. 35.

[186] J. Scheck, N. Purnell, and J. Horwitz, *The Facebook files: Facebook employees flag drug cartels and human traffickers. the company's response is weak, documents show.* 2021.

[187] M. Schroeck, R. Shockley, J. Smart, D. Romero-Morales, and P. Tufano, "Analytics: The real-world use of big data," *IBM Global Business Services*, Oct. 2012. [Online]. Available: https://www.bdvc.nl/images/Rapporten/GBE03519USEN.PDF (visited on 04/24/2024).

[188] M. Schroepfer, "Creating a data set and a challenge for deepfakes," *Facebook artificial intelligence*, 2019.

[189] S. Shane and M. Isaac, "Facebook says it's policing fake accounts. but they're still easy to spot," *New York Times*, vol. 3, 2017. [Online]. Available: https://www.nytimes.com/2017/11/03/technology/facebook-fake-accounts.html (visited on 04/24/2024).

[190] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[191] C. Shao, G. L. Ciampaglia, A. Flammini, and F. Menczer, "Hoaxy: A platform for tracking online misinformation," in *Proceedings of the 25th International Conference Companion on World Wide Web*, ser. WWW '16 Companion, Montreal, Quebec, Canada: International World Wide Web Conferences Steering Committee, 2016, pp. 745–750. DOI: 10.1145/2872518.2890098.

[192] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, "The spread of low-credibility content by social bots," *Nature Communications*, vol. 9, no. 1, Dec. 2018. DOI: 10.1038/s41467-018-06930-7.

[193] E. Shearer and E. Grieco, "Americans are wary of the role social media sites play in delivering the news," Pew Research Center, Washington, D.C., USA, Rep., Oct. 2019.

[194] P. Shiralkar, A. Flammini, F. Menczer, and G. L. Ciampaglia, "Finding streams in knowledge graphs to support fact checking," in *2017 IEEE International Conference on Data Mining (ICDM)*, Nov. 2017, pp. 859–864. DOI: 10.1109/ICDM.2017.105.

[195] S. Shojaee, M. A. A. Murad, A. B. Azman, N. M. Sharef, and S. Nadali, "Detecting deceptive reviews using lexical and syntactic features," in *2013 13th International Conference on Intellient Systems Design and Applications*, Salangor, Malaysia: IEEE, Dec. 2013, pp. 53–58. DOI: 10.1109/ISDA.2013.6920707.

[196] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," Sep. 2, 2017. arXiv: 1708.01967.

[197] C. Silverman, "This analysis shows how viral fake election news stories outperformed real news on Facebook," *BuzzFeed news*, vol. 16, Nov. 2016. [Online]. Available: https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook (visited on 04/24/2024).

[198] T. Simonite, "A health care algorithm offered less care to black patients," *WIRED*, [Online]. Available: https://www.wired.com/story/how-algorithm-favored-whites-over-blacks-health-care/ (visited on 04/24/2024).

[199] A. Singla, A. Jaiswal, S. Aggarwal, P. Sohni, P. Arora, and N. Sachdeva, "Rumour stance classification on textual social media content using machine learning," in *Advancements in Interdisciplinary Research: First International Conference, AIR 2022, Prayagraj, India, May 6–7, 2022, Revised Selected Papers*, Springer, 2023, pp. 313–322.

[200] D. Snow, *Adding a 4th V to BIG data - veracity*, 2012. [Online]. Available: http://dsnowondb2.blogspot.se/2012/07/adding-4th-v-to-big-data-veracity.html (visited on 04/24/2024).

[201] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image and Vision Computing*, vol. 65, pp. 3–14, 2017.

[202] C. Stupp, "Fraudsters used AI to mimic CEO's voice in unusual cybercrime case," *The Wall Street Journal*, vol. 30, no. 08, 2019. [Online]. Available: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402 (visited on 04/24/2024).

[203] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser, "Information quality work organization in Wikipedia," *Journal of the American society for information science and technology*, vol. 59, no. 6, pp. 983–1001, 2008. DOI: 10.1002/asi.20813.

[204] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[205] SVT, *Vårdinnovation hade en påhittad kommunikationschef*, 2021. [Online]. Available: https://www.svt.se/nyheter/lokalt/skane/vardinnovation-anvande-sig-av-en-pahittad-kommunikationschef (visited on 04/24/2024).

[206] A. Szal. "Report: Russian 'internet trolls' behind Louisiana chemical explosion hoax," Manufacturing.net. (Jun. 3, 2015), [Online]. Available: www.manufacturing.net/operations/news/13099148/report-russian-internet-trolls-behind-louisiana-chemical-explosion-hoax (visited on 04/17/2022).

[207] X. Tang, T. Qian, and Z. You, "Generating behavior features for cold-start spam review detection with adversarial learning," *Information Sciences*, vol. 526, pp. 274–288, 2020. DOI: 10.1016/j.ins.2020.03.063.

[208] S. Tasnim, M. M. Hossain, and H. Mazumder, "Impact of rumors and misinformation on COVID-19 in social media," *Journal of Preventive Medicine and Public Health*, vol. 53, no. 3, pp. 171–174, May 31, 2020. DOI: 10.3961/jpmph.20.094.

[209] L. Tološi, A. Tagarev, and G. Georgiev, "An analysis of event-agnostic features for rumour classification in Twitter," in *Proceedings of the Tenth International AAAI Conference on Web and Social Media*, 2016, pp. 151–158. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13197/12859.

[210] A. Tong, D.-Z. Du, and W. Wu, "On misinformation containment in online social networks," in *32nd Conference on Neural Information Processing Systems*, Montreal, Canada, 2018, p. 11.

[211] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, *Llama 2: Open foundation and fine-tuned chat models*, 2023. arXiv: 2307.09288.

[212] L. Townsend and C. Wallace, "Social media research: A guide to ethics," *University of Aberdeen*, pp. 1–16, 2016. [Online]. Available: https://www.gla.ac.uk/media/Media_487729_smxx.pdf.

[213] Travelmail, "TripAdvisor told to stop claiming reviews are 'trusted and honest'," *Mail OnlineTravel*, Feb. 1, 2012. [Online]. Available: https://www.dailymail.co.uk/travel/article-2094766/TripAdvisor-told-stop-claiming-reviews-trusted-honest-Advertising-Standards-Authority.html (visited on 04/15/2022).

[214] M. Tsikerdekis and S. Zeadally, "Multiple account identity deception detection in social media using nonverbal behavior," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 8, pp. 1311–1321, Aug. 2014, ISSN: 1556-6013. DOI: 10.1109/TIFS.2014.2332820.

[215] M.-H. Tsou, J.-A. Yang, D. Lusher, S. Han, B. Spitzberg, J. M. Gawron, D. Gupta, and L. An, "Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing): A case study in 2012 US presidential election," *Cartography and Geographic Information Science*, vol. 40, no. 4, pp. 337–348, 2013.

[216] B. Ulicny and M. Kokar, "Toward formal reasoning with epistemic policies about information quality in the twittersphere," in *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, IEEE, 2011, pp. 1–8.

[217] B. Ulicny, C. Matheus, and M. Kokar, "A semantic wiki alerting environment incorporating credibility and reliability evaluation," in *Proceedings of the 5th International Conference on Semantic Technologies for Intelligence, Defense, and Security (STIDS 2010), Fairfax, VA*, 2010.

[218] B. Ulicny, G. Powell, C. Matheus, M. Coombs, and M. Kokar, "Priority intelligence requirement answering and commercial question-answering: Identifying the gaps," in *Proceedings of the 15th International Command and Control Research and Technology Symposium (ICCRTS '10)*, 2010.

[219] D. Varshney and D. K. Vishwakarma, "A unified approach for detection of Clickbait videos on YouTube using cognitive evidences," *Applied Intelligence*, vol. 51, no. 7, pp. 4214–4235, Jul. 2021. DOI: 10.1007/s10489-020-02057-9.

[220] D. Varshney and D. K. Vishwakarma, "A review on rumour prediction and veracity assessment in online social network," *Expert Systems with Applications*, vol. 168, p. 114 208, 2021. DOI: 10.1016/j.eswa.2020.114208.

[221] A. Vartapetiance and L. Gillam, "Deception detection: Dependable or defective?" *Social Network Analysis and Mining*, vol. 4, no. 1, p. 166, Mar. 2014. DOI: 10.1007/s13278-014-0166-8.

[222] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[223] B. Verhoeven and W. Daelemans, "CLiPS stylometry investigation (CSI) corpus: A Dutch corpus for the detection of age, gender, personality, sentiment and deception in text.," in *LREC*, 2014, pp. 3081–3085.

[224] J. Vincent, *Google and Microsoft's chatbots are already citing one another in a misinformation shitshow*, Mar. 2023. [Online]. Available: https://www.theverge.com/2023/3/22/23651564/google-microsoft-bard-bing-chatbots-misinformation (visited on 04/24/2024).

[225] M. Viviani and G. Pasi, "Credibility in social media: Opinions, news, and health information – a survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 7, no. 5, 2017. DOI: 10.1002/widm.1209.

[226] S. Vosoughi, M. 'N. Mohsenvand, and D. Roy, "Rumor Gauge: Predicting the veracity of rumors on Twitter," *ACM Trans. Knowl. Discov. Data*, vol. 11, no. 4, 50:1–50:36, Jul. 2017. DOI: 10.1145/3070644.

[227] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, Mar. 9, 2018. DOI: 10.1126/science.aap9559.

[228] C. Wang, J. Wang, X. Xie, and W.-Y. Ma, "Mining geographic knowledge using location aware topic model," in *Proceedings of the 4th ACM workshop on Geographical information retrieval*, ACM, 2007, pp. 65–70. DOI: 10.1145/1316948.1316967.

[229] F. Wang, H. Wang, K. Xu, R. Raymond, J. Chon, S. Fuller, and A. Debruyn, "Regional level influenza study with geo-tagged Twitter data," *Journal of Medical Systems*, vol. 40, no. 8, pp. 1–8, 2016. DOI: 10.1007/s10916-016-0545-y.

[230] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *Journal of management information systems*, vol. 12, no. 4, pp. 5–33, 1996.

[231] Y. Wang, M. McKee, A. Torbica, and D. Stuckler, "Systematic literature review on the spread of health-related misinformation on social media," *Social science & medicine*, vol. 240, p. 112 552, 2019. DOI: 10.1016/j.socscimed.2019.112552.

[232] H. Webb, P. Burnap, R. Procter, O. Rana, B. C. Stahl, M. Williams, W. Housley, A. Edwards, and M. Jirotka, "Digital wildfires: Propagation, verification, regulation, and responsible innovation," *ACM Trans. Inf. Syst.*, vol. 34, no. 3, 15:1–15:23, Apr. 2016. DOI: 10.1145/2893478.

[233] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "TwitterRank: Finding topic-sensitive influential twitterers," in *Proceedings of the third ACM international conference on Web search and data mining*, ACM, 2010, pp. 261–270. DOI: 10.1145/1718487.1718520.

[234] C. Whissell, "Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language," *Psychological reports*, vol. 105, no. 2, pp. 509–521, 2009.

[235] K. Wiggers, *Voice cloning experts cover crime, positive use cases, and safeguards*, 2020. [Online]. Available: https://venturebeat.com/2020/01/29/ftc-voice-cloning-seminar-crime-use-cases-safeguards-ai-machine-learning/ (visited on 04/24/2024).

[236] M. L. Williams, P. Burnap, and L. Sloan, "Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation," *Sociology*, vol. 51, no. 6, pp. 1149–1168, 2017. DOI: 10.1177/0038038517708140.

[237] Y. Wu, E. W. Ngai, P. Wu, and C. Wu, "Fake online reviews: Literature review, synthesis, and directions for future research," *Decision Support Systems*, vol. 132, p. 113 280, May 2020. DOI: 10.1016/j.dss.2020.113280.

[238] S.-R. Yan, X.-L. Zheng, Y. Wang, W. W. Song, and W.-Y. Zhang, "A graph-based comprehensive reputation model: Exploiting the social context of opinions to enhance trust in social commerce," *Information Sciences*, vol. 318, pp. 51–72, 2015, ISSN: 0020-0255. DOI: 10.1016/j.ins.2014.09.036.

[239] F. Yang, Y. Liu, X. Yu, and M. Yang, "Automatic detection of rumor on Sina Weibo," in *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, ACM, 2012, p. 13.

[240] S. Yu, M. Li, F. Liu, and M. Scarpiniti, "Rumor identification with maximum entropy in micronet," *Complexity*, vol. 2017, Jan. 2017. DOI: 10.1155/2017/1703870.

[241] D. H. Zanette, "Dynamics of rumor propagation on small-world networks," *Physical review E*, vol. 65, no. 4, p. 041 908, 2002.

[242] X. Zeng, A. S. Abumansour, and A. Zubiaga, "Automated fact-checking: A survey," *Language and Linguistics Compass*, vol. 15, no. 10, 2021. DOI: 10.1111/lnc3.12438.

[243] Q. Zhang, S. Zhang, J. Dong, J. Xiong, and X. Cheng, "Automatic detection of rumor on social network," in *Natural Language Processing and Chinese Computing*, Springer, 2015, pp. 113–122. DOI: 10.1007/978-3-319-25207-0_10.

[244] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Computing Surveys*, vol. 53, no. 5, pp. 1–40, 2020. DOI: 10.1145/3395046.

[245] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and resolution of rumours in social media: A survey," *ACM Computing Surveys*, vol. 51, no. 2, pp. 1–36, 2018. DOI: 10.1145/3161603.

[246] A. Zubiaga, M. Liakata, R. Procter, G. Wong Sak Hoi, and P. Tolmie, "Analysing how people orient to and spread rumours in social media by looking at conversational threads," *PLOS ONE*, vol. 11, no. 3, N. Masuda, Ed., e0150989, 2016. DOI: 10.1371/journal.pone.0150989.

Part II

Publications