Computational Modeling of Science Communication with Natural Language Processing

by

Jiaxin Pei

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information)
in the University of Michigan
2024

Doctoral Committee:

Associate Professor David Jurgens, Chair
Professor Sol Hart
Associate Professor Ágnes Horvát
Assistant Professor Misha Teplitskiy

Jiaxin Pei
pedropei@umich.edu
ORCID iD: 0000-0002-1849-7962

© Jiaxin Pei 2024

TABLE OF CONTENTS

LIST OF FIGURES iv

LIST OF TABLES vi

ABSTRACT vii

CHAPTER

1 Introduction **1**

1.1 An Overview of Motivations 2
1.1.1 Why Study Science Communication? 2
1.1.2 Why Do We Need New Computational Methods for Science Communication 3
1.1.3 Proposed Approach 4
1.2 Contributions of Studies 5
1.2.1 Certainty and Uncertainty in Science Communication 5
1.2.2 Modeling Nuanced Information Change 5
1.2.3 Modeling Public Perception of Scientific Information in Media 6

2 Related work **7**

2.1 Models of Science Communication 7
2.1.1 The Deficit Model 7
2.1.2 The Engagement Model 8
2.2 The Structure of Science Communication 9
2.3 Presentations of Scientific Information 10
2.4 NLP for Science Communication 12

3 Modeling Certainty and Uncertainty in Science Communication **14**

3.1 Introduction 14
3.2 Modeling Certainty 16
3.3 Data 18
3.4 Annotating Certainty 19
3.5 Predicting Certainty 22
3.6 Certainty in Science Communications 23
3.7 Discussion 28
3.8 Conclusion 28

4 Modeling Information Change in Science Communication . . . 29
4.1 Introduction . . . 29
4.2 Related Work . . . 31
4.3 SPICED . . . 32
4.3.1 Data Collection . . . 32
4.3.2 Data Sampling . . . 32
4.3.3 Finding Matching Annotation . . . 34
4.4 Scientific Information Change Models . . . 36
4.4.1 Experimental setup . . . 37
4.4.2 Results . . . 38
4.5 Application: Modeling Information Change in Science Communication . . . 39
4.5.1 RQ1: Comparing Media Outlets . . . 39
4.5.2 RQ2: Comparing Social Media Accounts . . . 40
4.5.3 RQ3: What Information Changes . . . 41
4.6 Conclusion . . . 41
5 Modeling Public Perception of Science in News Media . . . 44
5.1 Introduction . . . 44
5.2 Framework for Modeling the Public Perception of Scientific Information in Media 46
5.3 Building a New Dataset to Model Public Perception of Scientific Information . . . 49
5.3.1 Raw Dataset . . . 49
5.3.2 Crowdsourcing with Representative Samples . . . 51
5.3.3 Quality Check and Post Processing . . . 52
5.4 Computational Modeling of Public Perceptions . . . 54
5.5 Applications . . . 56
5.5.1 RQ1: Public Perceptions of Scientific Information Across and Within Fields . . . 57
5.5.2 RQ2: Background Factors Affecting Perceptions of Scientific Information 57
5.5.3 RQ3: Estimated Perceptions and Public Engagement with Science . . . 60
5.6 Discussions . . . 63
5.7 Ethics . . . 64
5.8 Conclusions . . . 64
6 Conclusion . . . 66
6.1 Summary of Findings . . . 66
6.2 Contributions . . . 67
6.3 Limitations . . . 68
6.4 Future Work . . . 69
BIBLIOGRAPHY . . . 71

LIST OF FIGURES

FIGURE

3.1 Certainty is a multi-dimensional construct. The certainty of a scientific finding can be perceived holistically at the sentence level from its description. However, scientific findings may involve multiple aspects that may each be described as certain (black↑) or uncertain (black↓) (aspect-level certainty). 15

3.2 (Top) The number of hedges shows a moderate correlation with sentence-level certainty, while (Bottom) hedges vary in frequency across aspects. 20

3.3 Relative sentence-level certainty when each aspect is certain/uncertain. The overall certainty of scientific findings is majorly affected by PROBABILITY and SUGGESTION, while are less affected by other aspects like NUMBER and EXTENT 21

3.4 Binary-F1 for predicting aspect-level certainty. SciBERT outperforms all the other baselines. 23

3.5 Controlling for multiple factors in RQ2 (e.g., topic, news outlet), the marginal effects show the relative probability of finding each aspect described in the abstract (left) versus news (right), revealing that some aspects like numeric certainty are much more likely to be described in one source. 25

3.6 Averaged marginal effects of aspect-level certainty in abstract findings on the level of certainty for matched news findings. 26

3.7 Averaged marginal effects of journal impact factors on level of certainty. Findings appearing on journals with lower impact factors are associated with higher levels of certainty. 27

3.8 Averaged marginal effects of number of authors on level of certainty. Findings written by more authors are associated with a higher level of certainty, while such a pattern does not persist in science news. 28

4.1 We are interested in measuring the information similarity of statements about scientific findings between different sources, including scientific papers, news, and tweets, shown here with real examples. The finding in this figure comes from Fang et al. (2016) and the news quote is from this Reuters story. 30

4.2 (a) Mean Squared Error (MSE, ↓ better) and (b) Pearson correlation (r, ↑ better) on the test set of SPICED. Grey = zero-shot transfer models, red = MLM models fine-tuned on SPICED,blue = SBERT models fine-tuned on SPICED. Results are averaged across 5 random seeds. Best results are given in bold. 36

4.3 Scientific findings covered by Press Release and SciTech generally have less informational changes compared with findings presented in General Outlets 40

4.4 Organizational Twitter accounts keep more original information from the paper finding while verified users and those with more followers change more information when tweeting about a scientific finding. 40
4.5 Journalists tend to downplay the certainty and strength of findings in abstracts, but overstate findings discussed in limitations sections. 42

5.1 Public perception of scientific information varies within and across scientific fields. . . 58
5.2 Numbers in each cell show the regression coefficient. Experience with science and individuals' backgrounds could both affect people's perception of scientific information. People's science news consumption frequency and trust in science have the largest effect sizes. 59
5.3 Posts with higher estimated perception values receive higher scores on Reddit. 62
5.4 Posts with higher estimated perception values receive more comments on Reddit. . . . 62
5.5 When sharing the same science news story, posts with higher estimated importance, surprisingness and fun receive higher scores on Reddit. Posts that require more specialized knowledge to understand receive lower scores. 63
5.6 When sharing the same science news story, posts with higher estimated importance, surprisingness, fun, and controversy receive more comments on Reddit. Posts that require more specialized knowledge to understand receive less comments. 63

LIST OF TABLES

TABLE

3.1 Sentence-level certainty performance 22

4.1 Annotated information matching score (IMS) and the similarity score estimated by SBERT (Reimers and Gurevych, 2019) for selected finding pairs from SPICED. These examples demonstrate that simple similarity scores may not reflect whether the two sentences are covering the same scientific finding. 33

4.2 The average normalized edit distance between matching pairs for various datasets shows that SPICED includes more pairs that are lexically dissimilar. For SPICED and STSB, pairs are considered matching if their similarity score is greater than 3. For SNLI, pairs are considered matching if the label is "entailment". 35

5.1 The twelve-dimension framework to measure public perceptions of scientific information in media. 50

5.2 Krippendorff's α and Cronbach's α for rating the public perceptions of science news stories. 52

5.3 Model performances as measured by Pearson's r over different test sets. The model generally performs well on News and Reddit posts but performs poorly on tweets. "Paper" refers to the Arxiv abstract newsworthiness dataset created by Nishal and Diakopoulos (2022). We calculated the correlation between our model's predictions and crowd ratings for the corresponding category. Overall our model also shows good performance on the Arxiv abstract newsworthiness dataset. * means that an absolute correlation score is used because the two measured dimensions are conceptually opposite. 55

5.4 Krippendorff's α and Cronbach's α for rating the public perceptions of Reddit and Twitter posts. We observe similar or higher levels of reliability and internal consistency compared with news annotations. 56

ABSTRACT

From the evolution of human beings to state-of-the-art artificial intelligence, knowledge created by scientists has been actively communicated to the public and has further driven the development of human society. Science communication is the process of conveying scientific information produced by the science community to the general public. Science communication not only aims to promote public understanding of scientific facts but also is tasked with encouraging public engagement with scientific information. For science communicators, effectively communicating science is challenging as it involves many key steps, including modeling the public's interests, identifying newsworthy information, and translating technical jargon into plain language that people can easily understand and engage with. My dissertation focuses on building natural language processing models and conducting large-scale analyses to understand how scientific information is communicated to the public. More specifically, in the three proposed studies, I developed NLP models that can automatically analyze the textual certainty in scientific claims, the holistic information change in science communications, and the public perception of science news stories. Further large-scale analysis demonstrates that the proposed methods could be used to analyze science communication at scale and provide valuable insights about effective science communication.

CHAPTER 1

Introduction

> Nothing in science has any value to society if it is not communicated.
>
> —— *Anne Roe*

The ability to create and communicate knowledge is one of the key factors that distinguish human beings from other species. From learning how to light a fire in a cold valley to mastering the art of launching starships, human beings create, accumulate, and leverage knowledge to improve our world for everyone's benefit. While knowledge can arise from the daily experiences of laypeople, in modern society, knowledge production increasingly relies on professional scientists dedicated to pushing the boundaries of human understanding (Wolpert, 1994). Whether it's understanding how to generate electricity or maintaining emotional well-being, scientific discoveries drive changes in human society, and individuals rely on these findings to guide their daily decisions.

Scientific findings are the result of scientific research and represent knowledge generated through rigorous methodologies. Unlike personal life experiences, which are often influenced by individual biases, scientific findings are produced by trained and dedicated professional scientists. Therefore, while science itself may contain uncertainties (Brillouin, 2014), scientific knowledge in the public domain is generally considered trustworthy and serves as a foundation for guiding policy design and daily decisions across various domains (Castell et al., 2014; Besley, 2013). Furthermore, the public exhibits significant interest in scientific topics, sometimes even surpassing that in sports news and politics (Durant et al., 1989).

Despite the widespread public interest in learning about science, laypeople, in most cases, do not directly engage with scientific information produced by scientists. Instead, "science communication", the process of communicating scientific information to the general public is usually needed to fulfill the public interest in science (Burns et al., 2003). However, science communication faces numerous challenges due to the natural diversity among the lay public and the complexity of scientific information (Weigold, 2001; Scheufele, 2013). Firstly, any communication process involves information change, including science communication. In practice, this entails translating scientific jargon into accessible language that anyone can understand (Franklin, 2008). Yet, striking a

balance between clarity and accuracy is challenging even for professional journalists, leading to occasional miscommunication (Bell, 1994; Shuchman and Wilkes, 1997; Boiarsky, 2016). Moreover, given the vast volume of scientific information communicated daily, manual fact-checking and examination become nearly impossible at scale. Thus, monitoring information change in science communication becomes a critical issue.

Secondly, the general public encompasses individuals with diverse identities, educational backgrounds, and levels of experience with science. Therefore, effective science communication requires understanding the diverse needs of the lay public (Falk et al., 2007). However, public interest and reception of science could be influenced by various factors, making it challenging for even professional journalists to accurately gauge the public's interests (Atkin et al., 1983).

In my dissertation, I aim to address the challenges in science communication by developing computational methods that enable large-scale analysis of information change and public perception of scientific information in the media. Specifically, I have developed natural language processing models to (1) analyze the certainty of scientific findings (Chapter 3), (2) quantify the degree of information change for scientific findings across research papers, news stories, and social media posts (Chapter 4), and (3) estimate the public perception of scientific information in media (Chapter 5). These methods are further applied to large-scale datasets to answer key research questions in science communication.

1.1 An Overview of Motivations

Understanding and promoting effective science communication is a crucial endeavor for our society, yet it encounters numerous challenges given the intricate nature of scientific information and the stakeholders involved (Fischhoff, 2012a). Despite existing efforts in modeling scientific information, additional research is required to address several fundamental challenges in science communication. In this section, I will discuss these core challenges in both research and practice within science communication and introduce the proposed approach of this dissertation.

1.1.1 Why Study Science Communication?

This dissertation focuses on developing computational methods to study science communication. To begin, it is crucial to discuss why we should study science communication. In this section, I discuss the significance of science communication and outline the current challenges it faces.

Science communication is the process of communicating scientific findings and knowledge to the public (Trench, 2008). The term "public" encompasses individuals across society beyond domain experts, including parents, government officials, and students (Burns et al., 2003). At the

individual level, people's daily decision-making relies on various forms of scientific knowledge. Misunderstanding these scientific facts could lead to undesirable outcomes (Dewey, 1929). For example, the misunderstanding of the COVID-19 vaccine may lead to vaccine hesitation and further increase an individual's health risk (Garett and Young, 2021). At the societal level, a collective understanding of science, also known as science literacy, has long been considered a critical foundation for modern society (Laugksch, 2000; National Academies of Sciences et al., 2016). Therefore, engaging the public with accurate and important scientific findings is vital for both individuals and society as a whole.

However, science communication is not a trivial task and effectively communicating science to the public encounters many challenges (Bruine de Bruin and Bostrom, 2013). The first challenge lies in understanding the public's interest in science. Communication is contextual and audience-specific (Bauer, 1958). However, the audience for scientific information spans across the entire society, making it difficult to fully grasp their interests and needs (Falk et al., 2007). The second challenge involves balancing faithfulness and understandability of the scientific information. Science communicators usually need to translate complex jargon into accessible language for laypeople. However, simplifying information risks losing nuance, and presenting information in a way that is engaging and accurate is challenging (Gyasi, 2013). Thirdly, handling scientific uncertainties represents a significant challenge. Science naturally involves uncertainty because it tackles with problems that were usually unknown. In practices, scientists actively communicate the uncertainties in research articles (Fischhoff, 2012a; Brillouin, 2014). However, people tend to avoid uncertain information when making decisions (Ellsberg, 1961; Fischhoff et al., 2011), making it crucial for communicators to effectively convey uncertainties about science.

1.1.2 Why Do We Need New Computational Methods for Science Communication

Effectively communicating science to the public and comprehensively understanding current science communication practices face numerous challenges. Despite extensive research exploring science communication from various angles, several important limitations persist. Here, I discuss the major research gaps in the existing literature.

One of the major issues in the current science communication research is that the prior literature may not be able to generate actionable suggestions for the communication of science. Such issues can be majorly attributed to two reasons: (1) research topic: existing studies usually focus on specific scientific topics like climate change or COVID-19, therefore the produced findings may not generalize to other topics. As pointed out by National Academies of Sciences, Engineering, and Medicine (2017) "The studies that make up the literature in this field are fragmented, issue-specific,

and anchored in different disciplines, and often address the specific topic of science communication only obliquely". (2) research method: existing research usually relies on humans analyzing a small sample of science news articles. Different research may generate conflicting findings on the same topic (Murray et al., 2019). For example, the question of whether journalists overstate the certainty of scientific findings has long been debated. Some studies indicate that news reports often remove qualifiers and present uncertain findings as more definitive (Weiss and Singer, 1988; Fahnestock, 1986). Conversely, another study suggests that journalists can also portray scientific claims as more uncertain than they are (Friedman et al., 1999). This debate highlights the need for further research to develop knowledge and tools that can transcend specific fields and be applied more broadly.

Computational methods offer a potential solution by enabling the examination of science communication at scale. These methods make it feasible to analyze large-scale science communication datasets across various topics and fields. However, despite recent advances in Natural Language Processing (NLP), existing computational methods are limited in their applicability to science communication. For instance, while there are studies on modeling linguistic hedges (Hyland, 2005), hedges may not fully explain the variations of uncertainties in science communication. Additionally, although general-purpose language models exhibit promising performance across many NLP tasks, domain-specific fine-tuned models are often more accurate and efficient. Therefore, there is a continued need for new computational methods specifically designed to capture the nuances of science communication at scale.

1.1.3 Proposed Approach

To facilitate better understanding and practices for science communication, in this dissertation, I propose to computationally model science communication using natural language processing approaches. I develop new natural language models capable of analyzing large-scale datasets and addressing research questions pertinent to science communication. The first study focuses on modeling uncertainty in scientific claims. The second study focuses on a more fundamental problem in quantifying the holistic information change of scientific information. The third study models the public perceptions of scientific information presented in the media.

Science communication is a multifaceted process involving various stakeholders and procedures. These three studies form a cohesive series aimed at enhancing our understanding of science communication. The first two studies concentrate on "what is being communicated," transitioning from a specific aspect of information (certainty) to a holistic measure of information change. The third study extends beyond communication content to explore the audience's perceptions of scientific information in the media. Collectively, these studies contribute to a deeper understanding of

both content and the audience of science communication.

1.2 Contributions of Studies

In this dissertation, I propose three studies aiming at promoting the understanding and effectiveness of public science communication.

1.2.1 Certainty and Uncertainty in Science Communication

Certainty and uncertainty are essential components of science communication. However, how to model (un)certainty has long been a challenging question in both linguistics and science communication research. This study aims to answer the following questions: (1) Are hedges a good proxy to measure certainty in scientific texts? (2) How to model certainty in science communications? (3) Does the certainty of scientific findings change in science communication? (4) What factors affect the certainty of scientific findings in news and abstracts? In this study, I create (i) a new dataset and method for measuring certainty in scientific findings and (ii) an NLP model for automatic textual certainty prediction. I apply this model over 431K scientific findings to study a series of research questions in science communication. Our analysis shows that 1) hedges are not able to fully capture both sentence-level and aspect-level certainty in scientific findings 2) findings from news hold lower sentence-level certainty, contradicting existing studies that journalists tend to make science sounds more certain 3) findings from paper abstracts varies with journal impact and team size: low-impact journals and large teams often present scientific findings with higher sentence-level certainty. Overall, our study suggests that NLP methods could be used to effectively model certainty in science communications.

1.2.2 Modeling Nuanced Information Change

Science communication is a complex process and many nuances can be changed in the communication process. Can we quantify the holistic change of information in the science communication process? In the second study, I build a computational framework to analyze the degrees of information change for pairs of scientific findings. More specifically, in the second study, I present the Scientific Paraphrase and Information Change Dataset (SPICED), the first paraphrase dataset of scientific findings annotated for the degree of information change. SPICED contains 6,000 scientific finding pairs extracted from news stories, social media discussions, and full texts of original papers. Furthermore, I show that models trained on SPICED can reveal large-scale trends in the degrees to which people and organizations communicate new scientific findings in media.

1.2.3 Modeling Public Perception of Scientific Information in Media

The ultimate goal of science communication is to promote public engagement with scientific information in a competitive and complex media environment (Scheufele, 2013; Stilgoe et al., 2014). One key step is to understand how the public perceives the scientific information presented in the media. In the third study, I propose to model the public perception of science news stories through crowdsourcing and NLP. I propose a new categorization framework encompassing 12 nuanced dimensions to measure the public perception of scientific information in the media. Furthermore, I built a large-scale dataset covering 1,506 news articles annotated by 2,101 participants from the United States and the United Kingdom. Extensive analysis of this dataset shows that: (1) the public perception of science news stories is field-dependent but also varies at the story level. (2) experience with science news and trust in science are the two strongest predictors for an individual's ratings of science news stories. Additionally, to automatically estimate the perceived science news values, I fine-tuned a RoBERTa-Large model to automatically rate the news values of each science news story. Leveraging this model, I analyzed a large dataset of science news stories shared on Reddit. I found that the estimated public perceptions are positively associated with public engagement with scientific information on social media. Reddit posts estimated with higher perceived values are more likely to receive upvotes and comments. This result suggests that the proposed model could help science communicators better predict potential audience perceptions and promote public engagement with scientific information.

The next chapters are ordered in the following order. Chapter 2 introduces prior literature on science communication. I will discuss (1) theories around science communication (2) existing computational approaches in modeling science communication and (3) the motivation of my dissertation and the contributed studies. Chapter 3 covers the first study on modeling certainty and uncertainty of scientific findings using NLP. Chapter 4 discusses the second study on modeling information changes in the communication of science across news and social media platforms. Chapter 5 discusses the third study on modeling public perceptions of science news stories.

CHAPTER 2

Related work

Effective communication of science to the public is crucial for the advancement of human society. As a result, science communication has garnered wide interest from diverse fields such as communication (Weingart and Guenther, 2016), sociology (D'andrea and Declich, 2005), psychology (Garvey and Griffith, 1971), and the broader scientific community (Bruine de Bruin and Bostrom, 2013). In this section, I provide a literature review of existing work on science communication. Although the term "science communication" may appear intuitive, its underlying implications and scope can vary largely depending on the specific context (Weigold, 2001). I begin by discussing two classic models for science communication, as they provide basic theoretical frameworks for understanding science communication practices and research. Based on the two classific models, I then introduce definitions and the basic structure of science communication. Lastly, I discuss the current challenges in science communication and then summarize the proposed approach and contributed studies.

2.1 Models of Science Communication

"Science communication" broadly refers to the practices of communicating scientific information to the general public. While the term "science communication" appears straightforward, in real practice, science communications may be operated with diverse underlying assumptions and goals (Weigold, 2001). Existing literature has majorly discussed two models of science communication: the deficit model and the engagement model (Bubela et al., 2009).

2.1.1 The Deficit Model

The deficit model is one of the most classic models of science communication, and despite all its critics, it has been widely accepted by many scientists, journalists, and policymakers (Besley, 2013; Suldovsky, 2016). The deficit model assumes that the public is ignorant about science and,

because of this lack of scientific knowledge, they may have negative attitudes towards science and its applications (Kearnes et al., 2006). Therefore, the goal of science communication is to educate the general public about scientific facts and improve science literacy (Jones, 2008). As Dickson (2005) suggests, this model assumes that "by providing sufficient information about modern science and technology to overcome this lack of knowledge—or 'knowledge deficit'—the public will change its mind and decide that both science and the technology that emerges from it are 'good things." In this sense, science communication is "directed at filling in the deficit in knowledge, with the hope that if members of the public only understood the scientific facts, they would be more likely to see the issues as experts do (Bubela et al., 2009)."

The deficit model has been criticized for years because it is not contextualized with social-cultural factors as well as how science is presented to the public (Sturgis and Allum, 2004; Bubela et al., 2009; Evans and Durant, 1995). More specifically, critics of the deficit model majorly focus on the following points: (1) It oversimplifies the connection between prior scientific knowledge and the public's attitude towards science, neglecting other important contextual factors like identity and trust (Siipi and Ahteensuu, 2011). (2) it majorly considers science communication as an educational process to promote science literacy through science media outlets and therefore overlooks the public reception of scientific information in a competing media environment (Bubela et al., 2009; Nisbet and Goidel, 2007).

2.1.2 The Engagement Model

Given all the critics of the deficit model, the public engagement model, as an alternative, gradually attracts more attention from researchers (Stilgoe et al., 2014; Leshner, 2003). Unlike the deficit model, the engagement model acknowledges the contextualized nature of science communication and does not assume public illiteracy as the sole explanation for negative attitudes towards science (Scheufele, 2013; Scheufele and Krause, 2019). Moreover, the engagement model views science communication as a dialogue involving a diverse range of participants (Nisbet and Scheufele, 2009). Accordingly, the goal of science communication is not to educate the public but to facilitate public engagement with scientific information. Moreover, public opinions can play a crucial role in shaping public policy and research priorities (Bubela et al., 2009). Existing research indicates that promoting public engagement with science can empower citizens (Powell and Lee Kleinman, 2008) and enhance trust in technology and scientists (Besley et al., 2008).

2.2 The Structure of Science Communication

Science communication is a complex process involving many stakeholders and procedures. Broadly speaking, the ultimate goal of science communication is to promote public understanding of and engagement with science (Stilgoe et al., 2014). It is important to thoroughly discuss "What is the public," "What is being communicated," and "Who are the communicators." In this section, drawing from existing literature, I discuss the basic components and structures of science communication.

The Audience of Science Communication Broadly speaking, science communication refers to any practices that are trying to communicate scientific information to the public, which could include government officials, relevant industries, the general public, or even other scientists (Weigold, 2001). Therefore, from a broad sense, the audience of science communication could be "every person in society" (Burns et al., 2003). However, in existing literature, the scope of science communication might be narrowed to the communication to the general public, or the lay public (Weingart and Guenther, 2016), which means a group of people who are "non-expert in a particular field" (Burns et al., 2003). Therefore, one important underlying assumption of science communication is that the audiences have varied levels of prior knowledge about science, and translation is needed in this process. It is important to note that such an assumption is different from the basic assumptions of the deficit model, which considers public illiteracy as the major cause of negative attitudes towards science. The deficit model simply assumes that the public is illiterate about science. However, in reality, it is very likely that science communication may actually be only able to reach a small subgroup of people who are already aware of and interested in certain topics (Goidel and Nisbet, 2006). Therefore, understanding the variations of public understanding of science represents a key step to contextualize science communication and further promote public engagement with science (Fischhoff, 2013; Nisbet and Scheufele, 2009). A broader definition of science communication may also encompass scholarly communication, which involves scientists communicating scientific information to their peers (Sompel et al., 2004; Rowlands et al., 2004). However, existing studies on science communication generally focus on the narrower definition which excludes the communication within the science community.

What Is Being Communicated The information being communicated in science communication is usually considered "scientific information." Scientific information is largely produced by professional scientists and first appears in scholarly publications (Hyland and Salager-Meyer, 2008). It is possible that scientists might communicate new scientific findings without scholarly publications. For example, research findings could be presented in seminars or discussed with

other scientists, these findings haven't gone through the peer-review process and are rarely communicated to the general public. Therefore, the original scientific information is largely coming from scholarly publications including conference proceedings, journals, or workshop publications (Evans, 2008). One grey area that receives an increasing amount of attention and discussion is pre-prints (Kaiser, 2017; Berg et al., 2016). While many pre-prints have not officially gone through peer review, the media sometimes picked the pre-print papers to cover due to their relevance to society (Fleerackers et al., 2024). This is especially common for timely topics like COVID-19 and AI.

Who Are the Communitators Science communication is a complex process of conveying scientific information to the general public and involves many challenges for communicators (Valdeón, 2018; Fischhoff, 2013). Traditionally, journalists, especially those covering science news, play a key role in translating scientific information for the general public (Weigold, 2001), and science journalism has been the primary focus of existing science communication research (Angler, 2017). Typical science journalism practices include identifying the topic or specific research to cover, searching for relevant literature, interviewing authors or relevant experts in the field, and finally, drafting a news story for audiences, usually the general public (Dunwoody, 2021). In addition to journalists, press release officers working for journals or universities also play a key role in publicizing new scientific findings (Woloshin and Schwartz, 2002). Press release offices usually have in-house journalists who work closely with scientists to create a story about their research work. Press releases serve as an important source of translated information in the science news ecosystem, and many public journalists may rely on press release services like EurekAlert[1] to identify new topics or findings to cover (De Semir et al., 1998). Science news plays a crucial role in engaging the public with scientific information (Wade and Schramm, 1969; Funk et al., 2017). In recent years, with the rise of social media platforms like Twitter and TikTok, social media has also become an important platform for science communication. Social media platforms not only provide a medium for scientists to directly communicate their research to the public (Collins et al., 2016) but also create a unique environment where the public can easily communicate and engage with science (López-Goñi and Sánchez-Angulo, 2018).

2.3 Presentations of Scientific Information

The ultimate goal of science communication is to promote the public understanding of and engagement with scientific information (Fischhoff, 2013). To effectively communicate scientific information, communicators usually need to consider how to present the information to the audience. Here,

[1] https://www.eurekalert.org

we discussed several key dimensions of presentations that existing research has majorly explored.

Faithfulness and Accuracy Faithfulness is one of the primary dimensions that science communicators need to consider when communicating science to the public (Tichenor et al., 1970). However, how to faithfully communicate science to the public is challenging because of both the communicators and the audiences. On the one hand, while science journalists are trained professionals for news reporting, they may not have enough background knowledge to understand all the details of science (Krieghbaum, 1940). On the other hand, lay people may also face difficulties in understanding all the original scientific information (Peters et al., 2014b) and science communicators usually need to simplify complex science in a way that lay people can easily understand and engage with (Scheufele, 2013). Therefore, how to present science in a readable and faithful way becomes a big challenge (National Academies of Sciences, Engineering, and Medicine, 2017). Furthermore, science communicators including news outlets and social media users usually face a competitive media environment and are motivated to compete for the attention of readers (Davenport and Beck, 2001). Existing research suggests that exaggerated content is more likely to receive attention and the hype about science may originate from media coverage, social media mentions, and press releases (Caulfield and Condit, 2012; Intemann, 2022). Therefore, the hype of science also becomes another big issue against faithful science communication.

Engagement with Diverse Audiences Engaging the public with scientific information is an important goal of science communication (Stilgoe et al., 2014; Armeni, 2004). However, despite the public interest in learning about science, scientific information itself may not be easily understandable and engaging to lay people (Peters et al., 2014b; Board, 2012). Therefore, science communicators need to present science in a way that lay people can easily engage with. However, in reality, the diversity of the public can lead to diverse perceptions and attitudes toward science (Bubela et al., 2009; Nisbet and Goidel, 2007). Therefore, a fundamental issue in science communication lies in how to effectively communicate scientific knowledge to engage with diverse audience groups. Despite all the efforts from news outlets and other organizations, the coverage of science may still not fulfill the needs and interests of the general public (Durant et al., 1989). For example, the media may focus excessively on covering specific fields such as biomedical science (Badenschier and Wormer, 2011) and a survey by the Pew Research Center found that 20 percent of US adults consider "covering too many research findings that are not really important" to be a significant issue in media coverage of science (Funk et al., 2017). Therefore, how to better understand the public need for science and engaging with different audience groups is one of the biggest challenges for science communication (National Academies of Sciences, Engineering, and Medicine, 2017).

Limitations and Uncertainties One important goal of science is to provide factual knowledge to help people make daily life decisions (Scheufele, 2013). While people generally expect science to provide factual knowledge, in real practice, science often involves various types of uncertainties (Fischhoff, 2012a), disagreements (Murray et al., 2019; Lamers et al., 2021), and limitations (Sullivan, 2018; Rosen, 1996). Therefore, the challenge lies in how to communicate science to the public, considering all the limitations and uncertainties inherent to scientific research (Fischhoff, 2012a). Existing research suggests that the current science journalism practices may fail to communicate the limitations of scientific findings (Fischhoff, 2012b). This may be because science news practices sometimes lead journalists to avoid communicating uncertainties to the general public, fearing that it might lead to distrust in science. However, according to another research, communicating uncertainties may actually not affect people's trust in science (Van Der Bles et al., 2020). Despite this, given the audiences' ability to understand scientific uncertainty, how to contextualize science and faithfully communicate scientific uncertainties is still a challenging and unanswered question.

2.4 NLP for Science Communication

Recent advances in NLP offer huge opportunities to scale up science communication. In this section, I discuss the relevant literature on using NLP methods to model science communication in various settings.

Scientific Text Understanding Understanding scientific information is a fundamental task for the computational modeling of science communication. On the other hand, due to the complexities of scientific information, understanding scientific text has also been widely considered as one important setting to evaluate NLP models' language understanding capabilities. As a foundational approach for scientific text understanding, Beltagy et al. (2019a) developed a pre-trained language model for scientific text and attained promising performances on many NLP tasks. Other task-specific models have also been proposed to understand nuanced information in scientific texts, including tasks such as named-entity recognition (Li et al., 2016; Dogan et al., 2014), relation extraction (Kringelum et al., 2016), abstract sentence classification (Gonçalves et al., 2018), semantic text similarity (Magerman et al., 2010), framing detection (Jurgens et al., 2018; Prabhakaran et al., 2012) and citation worthiness (Wright and Augenstein, 2021a). Research on modeling scientific text provides important foundations for analyzing science communication at scale.

Misinformation Detection and Fact Checking Scientific findings provide important factual knowledge for individual and collective decision-making. On the contrary, scientific misinfor-

mation may mislead the public and lead to enormous societal costs. For example, during COVID-19, misinformation regarding vaccines creates vaccine hesitance (Garett and Young, 2021), which may further lead to increased mortality rates (Hoxha et al., 2022). Therefore, combating scientific misinformation has become an important issue, and automatic scientific fact-checking or rumor detection has attracted extensive attention from the NLP research community (Lin et al., 2021). A typical automatic scientific claim fact-checking pipeline usually includes models on detecting claims (Wuhrl and Klinger, 2021; Levy et al., 2014), retrieving relevant evidence (Rinott et al., 2015), and final fact verification (Soleimani et al., 2019). Recent advances in large language models also open new opportunities to leverage LLMs for zero-shot scientific fact-checking (Wright et al., 2022a; Tan et al., 2023), but also comes with the risk of creating new misinformation (Augenstein et al., 2023).

Analyzing Science News Stories Existing studies have also developed computational models to directly model science communication. August et al. (2020b) build a dataset and model to analyze writing strategies in press releases. Wright and Augenstein (2021b) proposed semi-supervised methods to detect exaggerations in health-related news stories. Despite the importance of automatic modeling of science news articles, NLP models developed for analyzing science news stories remain rare.

CHAPTER 3

Modeling Certainty and Uncertainty in Science Communication

3.1 Introduction

Expressing certainty about what is known is a necessary characteristic of scientific work as science involves producing knowledge about what was previously unknown (Friedman et al., 1999; Smithson, 2012). Given the natural aversion to uncertainty, existing studies have found that presenting uncertainty in science communications influences people's perception of scientific findings and trust in science (Gustafson and Rice, 2019; Fischhoff, 2012b; Van Der Bles et al., 2020). Therefore, understanding how journalists and scientists communicate certainty and uncertainty is critical for understanding the current practices of science journalism and further provides better guidance for uncertainty communication (National Academies of Sciences, Engineering, and Medicine, 2017).

Multiple studies in Linguistics, NLP, and the Science of Science literature have examined how certainty is expressed. These studies have modeled certainty in multiple ways, including epistemic modality (Vold, 2006), semantic uncertainty (Szarvas et al., 2012), verbal uncertainty (Hart and Childers, 2004), factuality (Saurí and Pustejovsky, 2009), and hedging (Hyland, 1996). In practice, most uses of uncertainty rely on hedging as a coarse characterization of the overall uncertainty (Farkas et al., 2010). However, as suggested by Rubin et al. (2006), certainty itself is a complex construct and has to be modeled from multiple dimensions. The complex and subjective nature of certainty makes annotation challenging, often resulting in moderate-to-low annotator agreement (Henriksson and Velupillai, 2010; Rubin, 2007), and motivating a better and more practical way to model and annotate certainty in text.

We propose to study certainty from the commonly-used sentence-level and, in parallel, introduce a new dimension of *aspect-level certainty*, providing a fine-grained description of how certainty is communicated in text. This approach is analogous to work in sentiment analysis that models both holistic (Meena and Prabhakar, 2007), and aspect-level valence (Schouten and Frasincar, 2015)—and their interactions. Based on existing categorizations of certainty, we compile

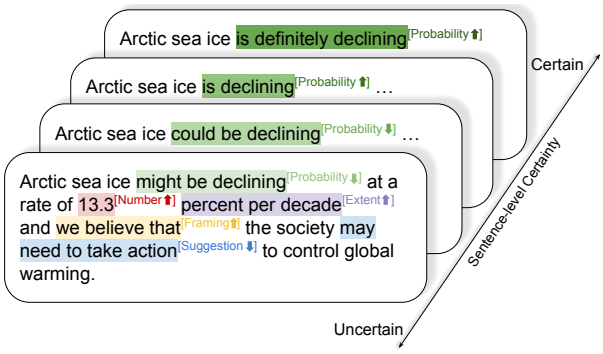


Figure 3.1: Certainty is a multi-dimensional construct. The certainty of a scientific finding can be perceived holistically at the sentence level from its description. However, scientific findings may involve multiple aspects that may each be described as certain (black↑) or uncertain (black↓) (aspect-level certainty).

six aspects of scientific findings including: NUMBER, EXTENT, PROBABILITY, FRAMING, CONDITION, and SUGGESTION. Following carefully designed annotation guidelines and after extensive annotator training, we introduce a new annotated corpus of 2200 scientific findings equally, sampled from news and scientific abstracts for both sentence-level and aspect-level certainty, and attained reliable inter-annotator agreements. Analysis with this dataset suggests that the number of hedges can only partially explain the variance of the overall sentence-level of certainty (Pearson's r=0.55). Therefore, to better model certainty in scientific findings, we fine-tuned a SciBERT (Beltagy et al., 2019a) for the two tasks and it achieves 0.63 Pearson's r for sentence-level certainty and an average of 0.66 binary-F1 for aspect-level certainty.

Our paper offers the following three contributions. First, we provide the first dataset of scientific findings annotated with both sentence-level and aspect-level certainty and fine-tune neural language models to predict certainty in scientific findings. Second, using our best-performing model, we infer the sentence-level and aspect-level certainty for 431K scientific findings in news and abstracts and show that the sentence-level certainty of findings in abstracts is associated with journal impact factor and team size. Regression analysis reveals that low-impact journals and large teams often present scientific findings with higher sentence-level certainty. Third, using 6586 findings from abstracts paired with their description in news, we find that news reports a finding with lower certainty than its corresponding description in the abstract. Fine-grained regression analysis over aspect-level certainty further reveals that journalists describe some key aspects with less certainty. Despite some studies suggest that news reports tend to describe uncertain findings as more certain (e.g., Weiss and Singer, 1988; Fahnestock, 2009), our study indicates that journalists may actually play down the certainty when reporting scientific findings.

3.2 Modeling Certainty

Communicating certainty involves qualifying parts of a statement to indicate what is known and what the speaker's beliefs are. Multiple approaches have proposed models for this property of language. The majority of these studies have focused on the degree or level of certainty (Rubin et al., 2006). Within this approach, hedges have been considered effective proxies for uncertainty (Farkas et al., 2010) across multiple fields including linguistics (Recasens et al., 2013), economics (Ahir et al., 2018), and psychology (Tausczik and Pennebaker, 2010), where the frequent use of hedges indicates more uncertainty and absence indicates certainty. Other work has proposed examining which aspects of language contribute to the perception of uncertainty. For example, Rubin et al. (2006) proposed a four-dimension model of certainty, including perspective, focus, timeline, and level, and each dimension contains several sub-categories. Following, we synthesize multiple approaches for measuring sentence-level and aspect-level certainty and propose a representative categorization.

Sentence-level Certainty Prior research has assumed that the level of certainty for a finding is presented, perceived, and further analyzed within one or several sentences (Holmes, 1982; Henriksson and Velupillai, 2010; Rubin, 2007). This aggregate perception represents a unified perception of various information expressed in the given piece of text and is the primary judgment of certainty along a continuum from uncertain to certain (Rubin et al., 2006). This perception of this overall level of certainty is known to influence people's following actions in many contexts (Corley and Wedeking, 2014; Wood and Eagly, 2009). Therefore, in modeling certainty, we include a sentence-level estimate of the certainty for a scientific finding's description.

Aspect-level Certainty In describing a finding, authors may contextualize which aspects are more or less certain. Many categorizations have been proposed to model the complex nature of these descriptions of certainty and uncertainty including types of uncertainty (Szarvas et al., 2012; Thunnissen, 2003), sources of uncertainty (Politi et al., 2007), issues of uncertainty (Han et al., 2011), forms of uncertainty (French, 1995), focus and perspective of certainty (Rubin, 2007). While prior work focuses on various aspects of certainty, their categorizations are often domain-specific (Politi et al., 2007; Han et al., 2011) or overly limited in scope to cover common aspects appearing in scientific findings (Szarvas et al., 2012; Rubin, 2007). Here, we synthesize prior approaches and propose six representative aspects of scientific findings which could involve certainty or uncertainty, with a goal of creating a comprehensive scheme that captures most of what is seen in knowledge-intensive corpora.

NUMBER refers to certainty towards specific quantities. For example, "approximately 250 indi-

viduals participated in this study" is uncertain towards NUMBER. Numerical information is vitally important in science communication as it is found to be the best way to promote scientific understanding in situations like climate change (Budescu et al., 2009) and health (Peters et al., 2014a). Accordingly, the imprecision of numbers or the inaccuracy of calculations are usually considered as a form of uncertainty (French, 1995). How to effectively communicate numerical information in scientific findings has been identified as one of the major challenges of science communication (Peters et al., 2014b). Identifying certainty regarding NUMBER in scientific findings could help to understand how journalists and scientists communicate certainty about numbers and inspire better ways to communicate this information.

EXTENT refers to certainty about the proportion/ratio of properties that make up an object/event or the extent of a change. For example, "This bridge is mainly composed of agate" and "We observe a moderate increase of suicide in Winter" involves uncertainty towards EXTENT. EXTENT can be described with numbers in certain situations. For example, "The average sea level across the world increased by approximately 30%" expresses extent via a number. However, unlike numbers that focus on specific quantities, EXTENT focuses on the components of an object, substance, or the extent of a change/effect. Previously, EXTENT was not explicitly proposed as a source of uncertainty, although some studies have brought up similar ideas. For example, french1995uncertainty considers "Uncertainty about how much [the] impacts matter" as a form of uncertainty and phillips2009disclose propose uncertainty about the strength or validity of evidence about risks, which may not be described with specific quantities. Existing studies suggest that journalists may misreport the extent of scientific findings to which it is supported by evidence (Dixon and Clarke, 2013), motivating its inclusion here.

PROBABILITY refers to certainty about the probability that something will occur, has occurred, or is associated with another factor. For example, "This medicine could possibly cure cancer" and "A is possibly associated with B" involves uncertainty about PROBABILITY. PROBABILITY has been widely recognized as one major source of uncertainty (Howard, 1988; Mosleh and Bier, 1996; Politi et al., 2007) and how to communicate probabilities effectively has long been an important question in science communication (Budescu et al., 2012; Sinayev et al., 2015).

CONDITION refers to the situation where something depends on a specific condition, and the condition involves certainty or uncertainty (Szarvas et al., 2012). Scientific findings are often qualified by specific conditions under which the result is valid, which may themselves be certain or uncertain (Friedman et al., 1999). For example, "Cancer could be cured if the medicine can be made shelf-stable" is uncertain regarding CONDITION.

FRAMING refers to the certainty about how scientists or journalists themselves frame or interpret the scientific finding. For example, "We suspect A has effects on B" involves the uncertainty from the authors while "We conclude that A has effects on B" frames the finding with conviction.

This aspect is related to expressions about epistemic uncertainty(the speaker having or lacking knowledge Szarvas et al., 2012; Fox and Ülkümen, 2011) and psychological uncertainty in the Psychology literature (Windschitl and Wells, 1996). In the news, journalists actively add their interpretations about the original information (Lin et al., 2006) and different framing may further affect people's perception of the overall certainty of the presented information (Soni et al., 2014). Therefore, identifying the certainty about FRAMING could help us to better understand how journalist's framing affects people's perceptions of scientific findings.

SUGGESTION refers to certainty or uncertainty about the implications or future actions for the public or science community. Scientific findings do not only describe facts, but can also communicate practical implications for people's daily life (Batteux et al., 2021). For example, "Patients probably need more medicine to cure this disease" involves uncertainty regarding future actions. Uncertainty about SUGGESTION was previously identified as dynamic uncertainty in Szarvas et al. (2012).

A single scientific finding may include multiple aspects with their own certainties. For example, "The vaccine is effective for 76% of chances" is uncertain regarding PROBABILITY while is certain about the specific NUMBER. Similarly, "The scientists need to do more research to understand the effect of A on B" indicates the uncertainty about PROBABILITY but is certain about SUGGESTION.

3.3 Data

To study certainty, we construct a dataset of scientific findings reported in news and research articles. News data comes from Altmetrics, which tracks mentions of scientific articles in news outlets. We restrict our analysis to U.S.-based outlets where we could retrieve the full text of the article and where the DOI for the scientific article was recorded in the Microsoft Academic Graph (MAG), which provides metadata on the article (e.g., authors, abstract, and publication venue). A total of 128,942 news/article pairs were collected, spanning 273 different news outlets and 57,807 different scientific articles.

For scientific articles, we extract the findings from the abstract reported in the MAG using the abstract parser developed by Prabhakaran et al. (2016), which labels sentences as background, method, introduction, result, and conclusion. We use sentences labeled as result or conclusion in our analysis. For news, we adopt a heuristic approach and identify all sentences containing a discovery-related keyword (e.g., find, conclude). We retain the subjective clause after the verb as the finding. This process finally leads to 608,694 unique scientific findings from abstracts and 106,612 unique scientific findings from news reports. Among the 128,942 news-paper pairs, 52,406 have both identified findings from news and paper abstracts.

3.4 Annotating Certainty

We annotate scientific findings for both their sentence-level certainty and the presence and certainty of each of the six aspects. Given subjectivity in perception, the same expression of certainty may evoke different perceptions (Druzdzel, 1989). Prior work in annotating uncertainty has generally reported low to moderate inter-annotator agreement (Henriksson and Velupillai, 2010; Rubin, 2007); for example, Rubin (2007) reports Cohen's κ=0.41 when annotating news certainty with a five-level Likert scale. To mitigate these challenges, we carefully designed the annotation procedures which are described in this section.

Annotation Setup Annotators were recruited from a US university and received initial one-hour training. All annotators are fluent in English and have extensive experience in reading scientific news and research articles. Annotators who attained high IAA with our gold standard were retained and then went through four additional rounds of pilot annotation and discussions (2 rounds for sentence-level and 2 for aspect-level) to build consensus. All annotators were paid $15/hr for training and annotation.

Annotation was performed in three phases. In the **first phase**, the initial data was sampled in a way to be more balanced across levels of certainty. Markers of certainty are not equally distributed throughout scientific communication (Rubin et al., 2006); for example, 87% of the data labeled in Henriksson and Velupillai (2010) were found to be very certain. Therefore, in the initial data, we sample 1000 findings equally from news and paper abstracts where 50% scientific findings containing no hedges, 35% with one hedge, and 15% with 2+ hedges. The hedge words are collected from Hyland (2005).

The annotators were first asked to rate how certain they perceived the finding on a six-point Likert scale as the sentence-level certainty. Aspect-based ratings were performed in a separate round so as not to potentially bias annotators towards focusing their sentence-level judgments on the basis of aspects. For each aspect, annotators were asked to assess whether that aspect was present and, if so, whether the language for the aspect was certain or uncertain. For instances that are clearly not scientific findings, the annotators were instructed to label it as BAD-TEXT. Each finding was rated by at least two annotators for sentence-level and, due to increased variance observed during training, three annotators for aspect-level. In this phase, 2349 sentence-level and 3209 aspect-level annotations were collected for 1000 findings. Annotators had a high agreement in both sentence-level and aspect-level tasks. For sentence-level, annotators attained a Krippendorff's α=0.67, which is substantially higher than IAA for the closest comparable task (Rubin, 2007, Cohen's κ=0.41).[1] The final sentence-level rating is computed as the average across all an-

[1]While Cohen's κ and Krippendorff's α are not directly comparable, both are chance-corrected scores in [-1,1]

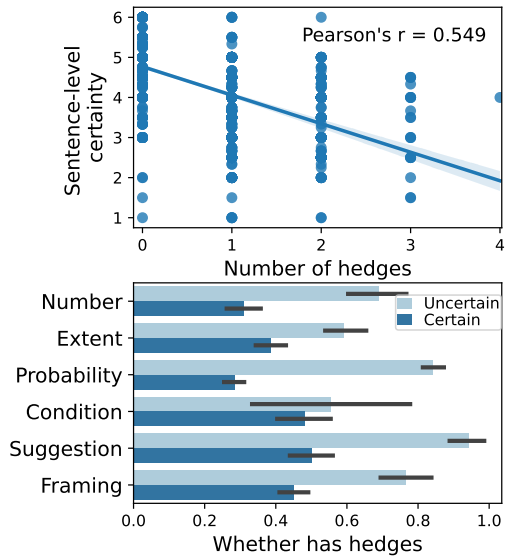


Figure 3.2: (Top) The number of hedges shows a moderate correlation with sentence-level certainty, while (Bottom) hedges vary in frequency across aspects.

notators' scores. For aspect-level certainty, the average Krippendorff's α is 0.57 for the six aspects, indicating moderate to high agreement. We take the majority label as the final label for aspect-level certainty.

In the **second phase**, given the label imbalance in both sentence-level and aspect-level certainty, we sample additional data for annotation based on model predictions using different sampling strategies for sentence-level and aspect-level certainties. For sentence-level certainty, we fine-tuned a roberta-base classification model and predicted all the extracted findings with sentence-level certainty. We further sampled 400 findings with low model confidence equally from news and abstracts for the second-phase annotation. For aspect-level certainty, we fine-tuned a SciBERT classification model and up-sampled 600 findings for less frequent aspects, including SUGGESTION, EXTENT, CONDITION, and FRAMING.

Given that the first and second phase data do not reflect the distribution of certainty in a natural sample, we further randomly sampled 200 findings from all the extracted findings as to the **third phase**. Given that the annotators were capable of doing reliable annotations after the first phase, annotators independently annotated 1200 findings for the second and third phases.

Results The final annotated dataset contains 6958 labels for 2200 findings. After removing findings that are labeled as BAD-TEXT, we obtained 1551 findings labeled with sentence-level certainty and 1760 findings labeled with aspect-level certainty, among which 1144 findings are labeled with

and we still view rough comparisons of magnitude useful to provide insight into annotators' relative agreement.

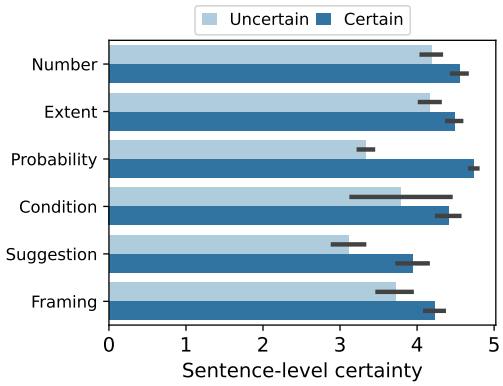


Figure 3.3: Relative sentence-level certainty when each aspect is certain/uncertain. The overall certainty of scientific findings is majorly affected by PROBABILITY and SUGGESTION, while are less affected by other aspects like NUMBER and EXTENT

both sentence-level and aspect-level certainty. Previously, Szarvas et al. (2012) considers CONDITION as one type of uncertainty, however, in the annotated data, less than 10% of CONDITION is labeled as uncertain. This difference indicates that previously proposed types of uncertainty may not be considered as uncertain in knowledge-intensive corpus like scientific findings and demonstrates the value of aspect-level certainty.

To what degree do hedges capture certainty? Comparing the sentence-level certainty with the number of hedges (fig:hedge-analysis, top) shows only a *moderate* correlation between hedging and certainty, r=0.55, despite their widespread use as a proxy. For example, "Further research is necessary to understand whether this is a causal relationship" contains zero hedges but explicitly expresses strong uncertainty towards the causal relationship, suggesting that many descriptions of certainty are not well capture by simple hedge-based lexicons. Further, authors vary in how frequently they employ hedges when describing the different aspects of certainty (fig:hedge-analysis, bottom). This variance in their distribution suggests that hedges are less effective as proxies for capturing uncertainty for all aspects.

What is the association between aspect-level and sentence-level certainty? Figure 3.3 shows the relative sentence-level certainty for each aspect. Uncertainties about PROBABILITY and SUGGESTION are associated with a sharp decrease of sentence-level certainty. However, the uncertainties about NUMBER and EXTENT are only associated with a small decrease of sentence-level certainty. This result indicates that the descriptions of aspects vary in how they contribute to the perception of the overall certainty of scientific findings.

Model	r on full test set	r on random set
LR + Hedges	−0.02	−0.02
LR + BoW	0.55	0.44
RoBERTa-base	0.62 ±0.051	0.55 ±0.067
SciBERT	**0.63 ±0.061**	**0.57 ±0.065**

Table 3.1: Sentence-level certainty performance

3.5 Predicting Certainty

In this section, we build models to predict sentence- and aspect-level certainty in scientific findings to support downstream analyses of certainty. We test two linear baseline models and two deep-learning models based on neural language models. As linear baselines, we include a model using bag-of-words (BoW) features and another based on the frequency of each hedging word. For neural models, we use SciBERT (Beltagy et al., 2019a), and RoBERTa model (Liu et al., 2019a) as the base models and fine-tune them over our annotated dataset. For both sentence-level and aspect-level certainty, the data labeled in phase 1 and phase 2 are split 8:1:1 into training, validation, and test. To better reflect the expected performance generalization, the test is made from the random set annotated in Phase 3 and the 10% test partition of Phases 1 and 2. For all the models, we also report their performance on the random test set to demonstrate their performance over the natural samples.

Sentence-level Certainty We formulate sentence-level certainty prediction as a regression task for all the models and Table 3.1 shows the model performance. We find that a linear weighting of the hedges is unable to predict the overall sentence-level certainty when tested on the random sample, largely due to the relatively low ratio of findings containing hedges. In comparison, linear regression with bag-of-words features is better able to capture overall certainty with Pearson's r=0.55, suggesting that other cues in addition to hedges also affect the overall certainty in the natural sample. Compared with the two baselines, the two neural models based on pre-trained language models achieve better performance. Both neural models are run five times with different random seeds, showing that the performance improvements over the baselines are statistically significant ($p<0.05$ paired t-test). The SciBERT model performs slightly better than the RoBERTa-base model, indicating domain-specific pre-training is helpful though not to the point of significance. We use the best-performing SciBERT model (r=0.70) as the regressor for sentence-level certainty in the following analyses.

Aspect-level Certainty For each aspect-level certainty, we predict whether a scientific finding is *NotPresent*, *Certain* or *Uncertain*. For the two neural models, we use shared pre-trained language

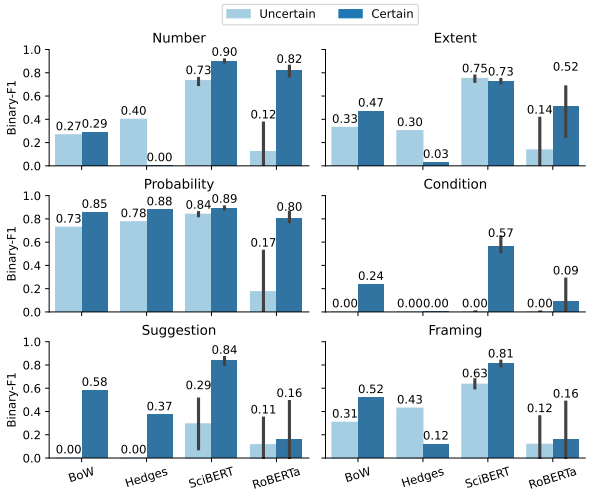


Figure 3.4: Binary-F1 for predicting aspect-level certainty. SciBERT outperforms all the other baselines.

models but independent classification heads for each aspect. Figure 3.4 shows the binary-F1 scores for predicting aspect-level certainty. The SciBERT model consistently outperforms other baselines over the six aspects, indicating that aspect-level certainty prediction requires more domain-specific and context information. However, given that uncertainties about CONDITION and SUGGESTION are relatively rare in the annotated dataset, the SciBERT model does not capture well the uncertainty about CONDITION and shows high variance when predicting uncertainty about SUGGESTION. In the following analysis, we use the best-performing SciBERT model (mean F1=0.71) as the classifier for aspect-level certainty.

3.6 Certainty in Science Communications

Certainty is a core aspect of science communication (Friedman et al., 1999) and presenting certainty in different forms (i.e., aspects) may further affect people's perception and future action about a series of issues including climate change (Fortner et al., 2000) and the Covid-19 vaccine (Batteux et al., 2021). Our models and dataset enable us to study how journalists and scientists present certainty in science communications. Here, we focus on the following five research questions. **RQ1:** Are findings in science news more certain than those in paper abstracts? **RQ2:** Do journalists and scientists differ in their use of aspect-level certainty? **RQ3**: Does aspect-level certainty in abstract findings affect the sentence-level certainty in news findings? **RQ4**: Does journal impact factor affect the certainty of scientific findings and how they are covered in news reports? **RQ5**: Does team size affect the certainty of scientific findings and how they are covered in news reports?

RQ1–RQ3 focus on changes to the description of certainty in the science communication process. While studies have found that news reports tend to describe uncertain findings as more certain (Weiss and Singer, 1988; Fahnestock, 2009), some studies suggest that news articles may also add more uncertainty to science finding in some cases (Friedman et al., 1999). Our model and dataset allow us to study (1) if certainty is changed, (2) if so, what aspects are changed, and (3) what drives the change. RQ4–RQ5 examine external factors that may affect how journalists and scientists present certainty. We focus on (4) the prestige/quality of the journal, asking whether lower- or higher-quality journals differ in how certain their findings are, and (5) in the era of team science (Stokols et al., 2008; Hall et al., 2018), whether team size influences how certain the authors describe the results, given known differences in the nature of team science research outputs (Wu et al., 2019; Sud and Thelwall, 2016; Thelwall and Sud, 2016).

Data and method For RQ1 and RQ2, to control the effects of the content of the finding, we propose a method to match the same scientific finding in news and paper abstracts. For each extracted finding in a paper abstract, we identify the paraphrased findings in the corresponding news article reporting on that paper. We first remove all punctuation and stop words and then stem all the words in each sentence. Next, we calculate the overlap and Jaccard similarity between each pair of findings in the news and abstract. We manually evaluated the matched findings and set word overlap $>=3$ and Jaccard similarity >0.3 as the threshold. Based on the findings from 52,406 news-paper pairs, we identify 6,586 unique finding pairs from news and abstracts. We manually annotated 70 matched finding pairs, and 63 (90%) refer to the same science finding, indicating high precision of our matching process. We construct separate regressions predicting the sentence-level (RQ1) and each aspect-level (RQ2) certainty in findings with the source of the finding (i.e., news or abstract). We further control the fields, author and affiliation ranking, journal impacts, finding length, and Flesch reading ease score. For RQ3, we construct a regression predicting the overall sentence-level certainty in news findings with the aspect-level certainty in the corresponding finding from the paper abstract. Except for all the IVs above, we further control the news outlet and the sentence-level certainty of the finding in the abstract.

For RQ4 and RQ5, we construct a regression predicting the sentence-level certainty in 265,758 findings presented in 55,178 paper abstracts using journal impacts factors and the number of authors. Recognizing its limitations (Kurmis, 2003), we use journal impact factor as a proxy for the quality of science based on prior use (Saha et al., 2003). We include controls the field of the research, author, and affiliation ranking extracted from the Microsoft Academic Graph (Wang et al., 2019a), finding length and Flesch reading ease score to remove the potential confounds. To test the connection between certainty in news findings and external factors, we construct another regression to predict the level of uncertainty in 72,013 findings presented in 27,000 news articles.

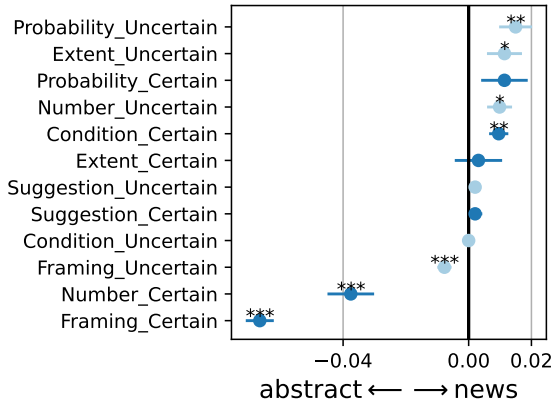


Figure 3.5: Controlling for multiple factors in RQ2 (e.g., topic, news outlet), the marginal effects show the relative probability of finding each aspect described in the abstract (left) versus news (right), revealing that some aspects like numeric certainty are much more likely to be described in one source.

Besides all the IVs regarding abstract and authors, we also control the outlet to remove potential confounds.

RQ1: Are findings in news more certain than those in paper abstracts? The regression analysis indicates that news descriptions have *lower* overall sentence-level certainty than abstract descriptions of the same finding ($p<0.01$). Although existing studies suggest that science news tends to remove hedges and describe science findings with increased certainty (e.g., Weiss and Singer, 1988; Fahnestock, 2009), our study over the paired findings finds the opposite: findings in news are less certain compared with findings in abstract, even when controlling the content and many contextual factors.

RQ2: Do journalists and scientists differ in their use of aspect-level certainty? Yes, as shown in fig:aspect-certainty-source, findings in abstracts are associated with more certainties about FRAMING and NUMBER. Findings in news are associated with uncertainties about PROBABILITY, EXTENT and NUMBER, indicating that the journalists tend to play down the certainty of some aspects, especially regarding numeric information. Existing studies suggest that laypeople with lower numeracy tend to focus more on narrative instead of numeric information (Dieckmann et al., 2009); one potential explanation for this difference is that journalists could be intentionally simplifying numerical information with hedges like "roughly " instead of the detailed number to better engage the lay audiences. Further, journalists are more likely to discard expressions of the scholar's uncertainty (FRAMING) when presenting the results, potentially aiming to make the work seem more objective. Our result suggests a potential mechanism for lower sentence-level certainty

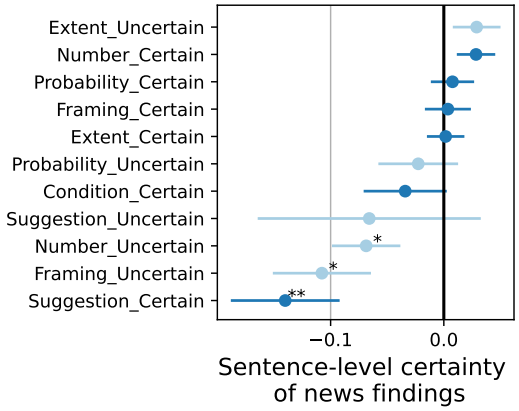


Figure 3.6: Averaged marginal effects of aspect-level certainty in abstract findings on the level of certainty for matched news findings.

in news compared with abstracts.

RQ3: Does aspect-level certainty in abstract findings affect the sentence-level certainty in news findings? As shown in figure 3.6, uncertainty about NUMBER and FRAMING are associated with decreased certainty in news findings, indicating that their uncertainty expressions are readily perceived by journalists. However, we also find that the certainty about SUGGESTION in abstracts are also associated with decreased certainty in news, suggesting that journalists may play down the certainty when presenting the findings involving suggestions or future actions even when it is certain. While existing studies suggest that journalists may exaggerate the potential benefits of science (Wilson et al., 2010), our result indicates that journalists can be very careful when reporting findings involving suggestions or future actions.

RQ4: Are findings in high-impact journals more certain than findings in low-impact journals? No. As shown in Figure 3.7, findings in the *lower*-impact journals are written with the highest level of certainty, while findings appearing in relatively higher-impact journals are described with comparatively less certainty. One potential explanation for this phenomenon is that high-quality papers published in journals with more strict reviewing processes[2] present certainty more precisely, which leads to a lower overall certainty compared with findings in low-impact journals. As a comparison, the certainty of findings written by journalists is not significantly associated with journal impact factors, suggesting that the prestige of a journal does not affect how journalists present scientific findings.

[2]Journals with higher impact factors generally have longer reviews than low impact journals (Publons, 2018, p. 36).

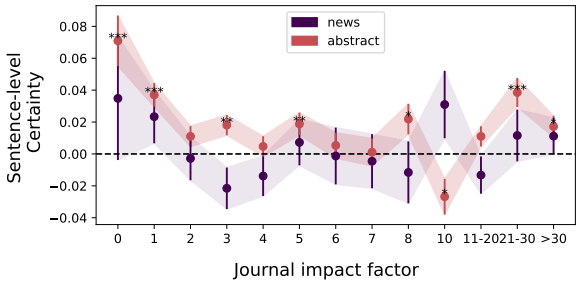


Figure 3.7: Averaged marginal effects of journal impact factors on level of certainty. Findings appearing on journals with lower impact factors are associated with higher levels of certainty.

RQ5: Are findings from small teams more certain than findings from large teams? We find a linear relationship between the number of authors and the overall level of certainty in scientific findings , even with controls for fields and authors. Multiple mechanisms may explain this behavior. Larger teams may themselves be more capable of producing more certain results due to more individuals participating and checking results or due to the scale of the experiments capable in team science (Bozeman and Corley, 2004; Jeong and Choi, 2015). Furthermore, our result also connects to Wu et al. (2019)'s finding that small teams generate new disruptive ideas while large teams tend to develop old, existing ideas, as new ideas are often associated with more uncertainties. Alternatively, larger teams likely have more authors who are further away from the experiments and therefore may not be able to describe certainty with the same nuance and precision as smaller teams (or single individuals) who are intimately familiar with the methodological details (MacKenzie, 1998); this distance may increase the perception of certainty when writing. However, this linear trend does not persist in science news; instead, the sentence-level certainty of findings in science news stays relatively steady across different numbers of authors. While team size has been found to be associated with the novelty and impact of science (Lee et al., 2015; Thelwall and Sud, 2016; Sud and Thelwall, 2016), our results indicate that the journalist is largely not influenced by the size of the research team in describing the certainty of their findings.

Across these results, our study suggests that the journalists report scientific findings with *less* certainty than scientists (RQ1). This result contradicts the existing findings that the journalists are overstating science (Weiss and Singer, 1988; Fahnestock, 2009). The fine-grained analysis over aspect-level certainty provides further details for such a change: journalists may play down the certainty about several core aspects of science findings like SUGGESTION even when they are certain in the abstract (RQ2,3). Moreover, we find that the certainty of scientific findings in research articles varies with journal impact factors and team size, while such a pattern does not persist in science news (RQ4,5), suggesting that journalists may not alter scientific uncertainty according to these factors.

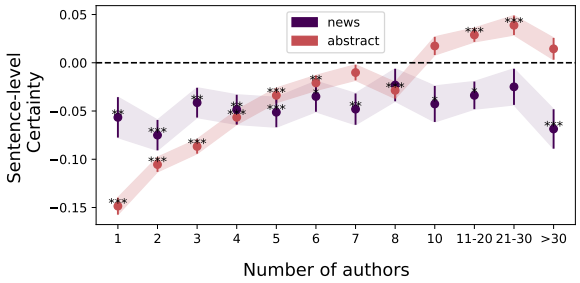


Figure 3.8: Averaged marginal effects of number of authors on level of certainty. Findings written by more authors are associated with a higher level of certainty, while such a pattern does not persist in science news.

3.7 Discussion

In this paper, we propose a new taxonomy, data, and models for certainty in science communications. Using the model, we analyzed a large scientific finding dataset and answered a series of important research questions on science communications. However, we also note the following limitations of our study. (1) We only use the abstract rather than the full text of research articles due to open access restrictions from copyright. Although authors normally present the core findings in the abstract, findings in abstracts could still be presented differently from findings in the main texts. (2) We use report verbs to extract findings in science news, which may miss findings that are presented without them. How to identify scientific findings in science news is still an open question and we call for future studies in this direction. (3) In our analysis, we use word-based heuristic methods (word overlap and Jaccard similarity) to match findings in news and abstracts, while the same scientific finding can be paraphrased with different sets of words. In future studies, we will develop better methods to identify paraphrases of scientific findings.

3.8 Conclusion

Our study represents a new step towards modeling certainty in text and demonstrates that sentence-level and aspect-level certainty are natural and feasible ways to model and annotate certainty. The proposed computational framework for certainty in scientific findings could support and inspire new studies on certainty in general language as well as new approaches to study science communication.

CHAPTER 4

Modeling Information Change in Science Communication

4.1 Introduction

Science communication disseminates scholarly information to audiences outside the research community, such as the public and policymakers (National Academies of Sciences, Engineering, and Medicine, 2017). This process usually involves translating highly technical language to non-technical, less-formal language that is engaging and easily understandable for lay people (Salita, 2015). The public relies on the media to learn about new scientific findings, and media portrayals of science affect people's trust in science while at the same time influencing their future actions (Gustafson and Rice, 2019; Fischhoff, 2012a; Kuru et al., 2021). However, not all scientific communication accurately conveys the original information, as shown in Figure 4.1. Identifying cases where scientific information has changed is a critical but challenging task due to the complex translating and paraphrasing done by effective communicators. Our work introduces a new task of measuring scientific information change, and through developing new data and models aims to address the gap in studying faithful scientific communication.

Though efforts exist to track and flag when popular media misrepresent science,[1] the sheer volume of new studies, reporting, and online engagement make purely manual efforts both intractable and unattractive. Existing studies in NLP to help automate the study of science communication have examined exaggeration (Wright and Augenstein, 2021b), certainty (Pei and Jurgens, 2021), and fact checking (Boissonnet et al., 2022; Wright et al., 2022b), among others. However, these studies skip over the key first step needed to compare scientific texts for information change: automatically identifying content from both sources that describe the **same** scientific finding. In other words, to answer relevant questions about and analyze changes in scientific information at scale, one must first be able to identify both the original information and the paraphrases.

[1] See e.g. https://www.healthnewsreview.org/ and https://sciencefeedback.co/

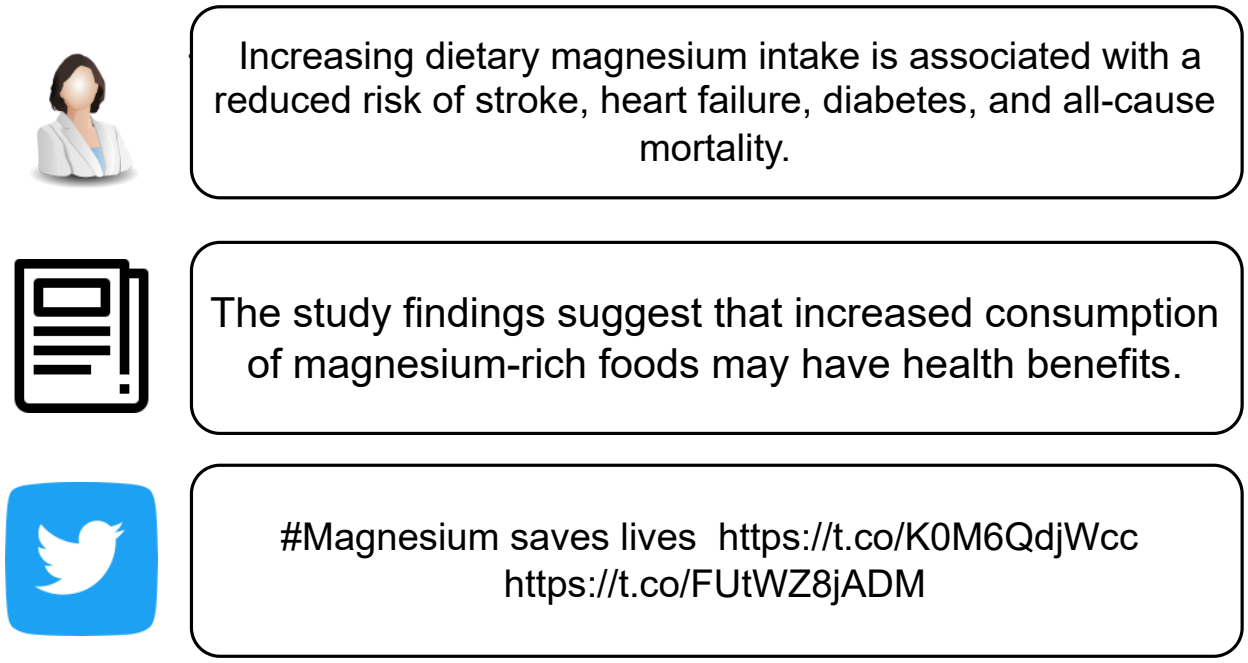


Figure 4.1: We are interested in measuring the information similarity of statements about scientific findings between different sources, including scientific papers, news, and tweets, shown here with real examples. The finding in this figure comes from Fang et al. (2016) and the news quote is from this Reuters story.

To enable automated analysis of science communication, this work offers the following **contributions** (marked by **C**). First, we present the SCIENTIFIC PARAPHRASE AND INFORMATION CHANGE DATASET dataset (SPICED), a manually annotated dataset of paired scientific findings from news articles, tweets, and scientific papers (**C1**, §4.3). SPICED has the following merits: (1) existing datasets focus purely on semantic similarity, while SPICED focuses on differences in the *information* communicated in scientific findings; (2) scientific text datasets tend to focus solely on titles or paper abstracts, while SPICED includes sentences extracted from the full-text of papers and news articles; (3) SPICED is largely multi-domain, covering the 4 broad scientific fields that get the most media attention (namely: medicine, biology, computer science, and psychology) and includes data from the whole science communication pipeline, from research articles to science news and social media discussions.

In addition to extensively benchmarking the performance of current models on SPICED (**C2**, §4.4), we demonstrate that the dataset enables multiple downstream applications. In particular, we perform an applied analysis on unlabelled tweets and news articles where we show (1) media tend to exaggerate findings in the limitations sections of papers; (2) press releases and SciTech tend to have less informational change than general news outlets; and (3) organizations' Twitter accounts tend to discuss science more faithfully than verified users on Twitter and users with more followers

(**C4**, §4.5).

4.2 Related Work

The analysis of scientific communication directly relates to fact checking, scientific language analysis, and semantic textual similarity. We briefly highlight our connections to these.

Fact Checking Automatic fact checking is concerned with verifying whether or not a given claim is true, and has been studied extensively in multiple domains (Thorne et al., 2018; Augenstein et al., 2019) including science (Wadden et al., 2020; Boissonnet et al., 2022; Wright et al., 2022b). Fact checking focuses on a specific type of information change, namely veracity. Additionally, the task generally assumes access to pre-existing knowledge resources, such as Wikipedia or PubMed, from which evidence can be retrieved that either supports or refutes a given claim. Our task is concerned with a more general type of information change beyond categorical falsehood and is a required task to complete prior to performing any kind of fact check.

Scientific Language Analysis Automating tasks beneficial for understanding changes in scientific information between the published literature and media is a growing area of research (Wright and Augenstein, 2021b; Pei and Jurgens, 2021; Boissonnet et al., 2022; Dai et al., 2020; August et al., 2020b; Tan and Lee, 2014; Vadapalli et al., 2018; August et al., 2020a; Ginev and Miller, 2020). The three tasks most related to our work are understanding writing strategies for science communication (August et al., 2020b), detecting changes in certainty (Pei and Jurgens, 2021), and detecting changes in causal claim strength i.e. exaggeration (Wright and Augenstein, 2021b). However, studying these requires access to paired scientific findings. To be able to do so at scale will require the ability to pair such findings automatically.

Semantic Similarity The topic of semantic similarity is well-studied in NLP. Several datasets exist with explicit similarity labels, many of which come from SemEval STS shared tasks (e.g., Cer et al., 2017) and paraphrasing datasets (Ganitkevitch et al., 2013). It is possible to build unlabelled datasets of semantic similarity automatically, which is the main method that has been used for scientific texts (Cohan et al., 2020; Lo et al., 2020). However, such datasets fail to capture more subtle aspects of similarity, particularly when the focus is solely on the scientific findings conveyed by a sentence. And as we will show, approaches based on these datasets are insufficient for the task we are concerned with in this work, motivating the need for a new resource.

4.3 SPICED

We introduce SPICED, a new large-scale dataset of *scientific findings* paired with how they are communicated in news and social media. Communicating scientific findings is known to have a broad impact on public attitudes (Weigold, 2001) and to influence behavior, e.g., the way vaccines are framed in the media has an effect on vaccine uptake (Kuru et al., 2021). Building upon prior work in NLP (Wright and Augenstein, 2021a; Pei and Jurgens, 2021; Sumner et al., 2014; Bratton et al., 2019), we define a scientific finding as **a statement that describes a particular research output of a scientific study, which could be a result, conclusion, product, etc.** This general definition holds across fields; for example, many findings from medicine and psychology report on effects on some dependent variable via manipulation of an independent variable, while in computer science many findings are related to new systems, algorithms, or methods. Following, we describe how the pairs of scientific findings were selected and annotated.

4.3.1 Data Collection

An initial dataset of unlabelled pairs of scientific communications was collected through Altmetric (https://www.altmetric.com/) a platform tracking mentions of scientific articles online. This initial pool contains 17,668 scientific papers, 41,388 paired news articles, and 733,755 tweets—note that a single paper may be communicated about multiple times. The scientific findings were extracted in different ways for each source. Similar to Prabhakaran et al. (2016), we fine-tune a RoBERTa (Liu et al., 2019b) model to classify sentences into methods, background, objective, results and conclusions using 200K paper abstracts from PubMed that had been self-labeled with these categories (Canese and Weis, 2013). This sentence classifier attained 0.92 F1 score on a held-out 10% sampleand then the classifier was applied to each sentence of the news stories and paper fulltexts. Given the domain difference between scientific abstracts and news, we additionally manually annotated a sample of 100 extracted conclusions; we find that the precision of the classifier is 0.88, suggesting that it is able to accurately identify scientific findings in news as well. We extract each sentence classified as "result" or "conclusion" and create pairs with each finding sentence from news articles written about it. This yields 45.7M potential pairs of ⟨news, paper⟩ findings. For tweets, we take full tweets as is, yielding 35.6M potential pairs of ⟨tweet, paper⟩ findings.

4.3.2 Data Sampling

Pairing every finding from a news story with every finding from its matched paper results in an untenable amount of data to annotate. Additionally, it has been shown that proper data selection

Paper finding	News Finding	Similarity Score	IMS
However, the consistency of the erythritol results in both the central adiposity and usual glycemia comparisons lends strength to the findings, and the cluster of metabolites has biological plausibility.	Young adults who exhibited central adiposity gain over the course of 35 weeks had plasma erythritol levels 15-times higher at baseline than those with stable adiposity over the same period.	0.88	1
Our results showed that most of the official adult-onset men began their antisocial activities during early childhood.	Beckley, who is in the department of psychology and neuroscience at Duke, said the adult-onset group had a history of anti-social behavior back to childhood, but reported committing relatively fewer crimes.	0.38	4.4

Table 4.1: Annotated information matching score (IMS) and the similarity score estimated by SBERT (Reimers and Gurevych, 2019) for selected finding pairs from SPICED. These examples demonstrate that simple similarity scores may not reflect whether the two sentences are covering the same scientific finding.

can reduce the need to annotate every possible sample (MacKay, 1992; Holub et al., 2008; Houlsby et al., 2011). Therefore, to obtain a sample of paired findings covering a range of similarities, we first filter our pool of unlabelled matched findings based on the semantics using Sentence-BERT (SBERT, Reimers and Gurevych (2019)), a Siamese BERT network trained for semantic text similarity, trained on over 1B sentence pairs We use this model to score pairs of findings from news articles and papers based on their embeddings' cosine similarity and conduct a pilot study to determine which data to annotate.

For the pilot, we sample 400 pairs evenly for every 0.05 increment bucket in the range $[0, 1]$ of similarity scores (20 per bucket). Each sample is annotated by two of the authors of this study with a binary label of "matching" vs "not matching", yielding a Krippendorff's alpha of 0.73.[2] From this sample, we observed that there were no matches below 0.3 and only 2 ambiguous matches below 0.4. At the same time, the vast majority of samples from the entire dataset have a similarity score of less than 0.4. Additionally, above 0.9 we saw that each pair was essentially equivalent. Given the distribution of matched findings across the similarity scale, in order to balance the number of annotations we can acquire, the yield of positive samples, and the sample difficulty, we sampled data as follows based on their cosine similarity:

1. Below 0.4 = automatically unmatched.

2. Above 0.9 with a Jaccard index above 0.5 = automatically matched.

[2]Note that many discussions about what constitutes matching vs. not matching were had in pilot work, leading to high agreement.

3. Sample an equal number of pairs from each 0.05 increment bin between 0.4 and 0.9 for human expert annotation.

We sample 600 ⟨news, paper⟩ finding pairs from the four fields which receive the most media attention (medicine, biology, computer science, and psychology) using this method. This yields 2,400 pairs to be annotated.

We follow a similar procedure to sample pairs from papers and Twitter for annotation. However, rather than use the SBERT similarity scores, we instead first obtain annotations for news pairs using the scheme to be described later in §4.3.3 in order to train an initial model on our task (CiteBERT, Wright and Augenstein 2021a). We then use the trained model to obtain scores in the range [0,1] for each pair and sample an equal number of pairs from bins in 0.05 increments, for a total of 1,200 pairs (300 from each field of interest).

4.3.3 Finding Matching Annotation

We perform our final annotation based on the sampling scheme above using the Prolific platform (https://www.prolific.co/) as it allows prescreening annotators by educational background. We require each annotator to have at least a bachelor's degree in a relevant field to work on the task. Annotators are asked to label "whether the two sentences are discussing the same scientific finding" for 50 finding pairs with a 5-point Likert schema where each value indicates that "The information in the findings is..." (1): Completely different (2): Mostly different (3): Somewhat similar (4): Mostly the same, or (5): Completely the same. We call this the INFORMATION MATCHING SCORE (IMS) of a pair of findings. Annotation was performed using POTATO (Pei et al., 2022). Notably, annotators were instructed to mark how similar the information in the *findings* was, as opposed to how similar the sentences are. Further, they were instructed to ignore extraneous information like "The scientists show..." and "our experiments demonstrate...".

Post processing To improve the reliability of the annotations, we use MACE (Hovy et al., 2013) to estimate the competence score of each annotator and removed the labels from the annotators with the lowest competence scores. MACE is a Bayesian method that learns distributions over true class labels and annotator competence based on crowd-sourced labels. We further manually examine pairs with the most diverse labels (standard deviation of ratings >1.2) and manually replace the outliers with our expert annotations. The overall Krippendoff's α is 0.52, 0.57, 0.53, and 0.52 for CS, Medicine, Biology, and Psychology respectively, indicating that the final labels are reliable. While many annotators considered the task challenging, our quality control strategies allow us

STSB	SNLI	SPICED	News	Tweets
0.401	0.631	**0.726**	*0.712*	*0.749*

Table 4.2: The average normalized edit distance between matching pairs for various datasets shows that SPICED includes more pairs that are lexically dissimilar. For SPICED and STSB, pairs are considered matching if their similarity score is greater than 3. For SNLI, pairs are considered matching if the label is "entailment".

to collect reliable annotations.[3] For all the annotated pairs, we average the ratings as the final similarity score. In addition to the 3,600 manually annotated pairs, we include an extra 2,400 automatically annotated pairs as determined in §4.3.2 (unmatched pairs get an IMS of 1, matched pairs get an IMS of 5), for a total of 6,000 pairs. Given that there can be multiple pairs from a single news-paper pair, to avoid overlaps between training and test sets, we split the dataset 80%/10%/10% based on the paper DOI and balance across subjects.

Selected Examples To highlight the difficulty of SPICED, we show a pair of samples from our final dataset in Table 4.1. The IMS is compared to the cosine similarity between embeddings produced by SBERT. For the first case, SBERT presumably picks up on similarities in the discussed topics, such as erythritol and its relationship to adiposity, but the paper finding is concerned with the consistency of results and its biological implications while the news finding explicitly mentions a relationship between erythritol and adiposity. The second case expresses the opposite effect; the news finding contains a lot of extraneous information for context, but one of the core findings it expresses is the same as the paper finding, giving it a high rating in SPICED.

Comparison with existing datasets To further characterize the difficulty of SPICED compared to existing datasets, we show the average normalized edit distance between matching pairs in SPICED, STSB (Cer et al., 2017), and SNLI (Bowman et al., 2015). STSB is a semantic text similarity dataset consisting of pairs of sentences scored with their semantic similarity, sourced from multiple SemEval shared tasks. SNLI is a natural language inference corpus, and consists of pairs of sentences labeled for if they entail each other, contradict each other, or are neutral. We calculated the mean normalized edit distance across all pairs of *matching* sentences in each dataset's training data; For SPICED and STSB, pairs are considered matching if their IMS or similarity score is greater than 3, respectively. For SNLI, pairs are considered matching if the label is "entailment".

We find that there is a much greater lexical difference between the matching pairs in SPICED (0.726) than existing general domain paired text datasets (0.401 for STSB and 0.631 for SNLI). This gap between STSB and SPICED also emphasizes the difference between traditional semantic

[3]For example, one participant commented "It was pretty hard to consider both the statements and their context then comparing them for similarities, but i enjoyed it"

textual similarity tasks and the information change task we describe here. Within SPICED, Twitter pairs had a higher distance (0.749) than news pairs (0.712), suggesting stronger domain differences.

Relationship of SPICED to Fact Checking The task introduced by SPICED captures information change more broadly than veracity as in automatic fact checking, as the task is concerned with the degree to which two sentences describe the same scientific information—indeed, two similar sentences may describe the same information equally poorly. Our task is similar to the sentence selection stage in the fact checking pipeline, and we later demonstrate that models trained on SPICED data are useful for this task for science in **??**. However, our task and annotation are agnostic to whether a pair of sentences entail one another. This is especially useful if one wants to compare how a particular finding is presented across different media. Fact-checking datasets are also explicitly constructed to contain claims which are about a single piece of information—SPICED is not restricted in this way, focusing on a more general type of information change beyond categorical falsehood. Finally, we note two more unique features of SPICED: 1) SPICED contains naturally occurring sentences, while fact checking datasets like FEVER and SciFact often contain manually written claims. 2) The combination of domains in SPICED is unique; sentences are paired between (news, science) and (tweets, science), and these pairings don't exist currently.

4.4 Scientific Information Change Models

(b) Pearson correlation (r)

(a) Mean Squared Error

Figure 4.2: (a) Mean Squared Error (MSE, ↓ better) and (b) Pearson correlation (r, ↑ better) on the test set of SPICED. Grey = zero-shot transfer models, red = MLM models fine-tuned on SPICED,blue = SBERT models fine-tuned on SPICED. Results are averaged across 5 random seeds. Best results are given in bold.

We now use SPICED to evaluate models for estimating the IMS of finding pairs in two settings: zero-shot transfer and supervised fine-tuning.

4.4.1 Experimental setup

We use the following four models to estimate zero-shot transfer performance.

- **Paraphrase**: RoBERTa (Liu et al., 2019b) pre-trained for paraphrase detection on an adversarial paraphrasing task (Nighojkar and Licato, 2021). We convert the output probability of a pair being a paraphrase to the range [1,5] for comparison with our labels.
- **Natural Language Inference (NLI)**: RoBERTa pre-trained on a wide range of NLI datasets (Nie et al., 2020). The final score is the model's measured probability of entailment mapped to the range [1,5].
- **MiniLM**: SBERT with MiniLM as the base network (Wang et al., 2020); we obtain sentence embeddings for pairs of findings and measure the cosine similarity between these two embeddings, clip the lowest score to 0, and convert this score to the range [1,5]. Note that this model was trained on over 1B sentence pairs, including from scientific text, using a contrastive learning approach where the embeddings of sentences known to be similar are trained to be closer than the embeddings of negatively sampled sentences. SBERT models represent a very strong baseline on this task and have been used in the context of other matching tasks for fact-checking including detecting previously fact-checked claims (Shaar et al., 2020).
- **MPNet**: The same setting and training data as MiniLM but with MPNet as the base network (Song et al., 2020).

We fine-tune the following six models on SPICED to estimate IMS as a comparison with zero-shot transfer.

- **MiniLM-FT**: The same MiniLM model from the zero-shot transfer setup but further fine-tuned on SPICED. The training objective is to minimize the distance between the IMS and the cosine similarity of the output embeddings of the pair of findings.
- **MPNet-FT**: The same setup as MiniLM-FT but using MPNet as the base network.
- **RoBERTa**: The RoBERTa (Liu et al., 2019b) base model; We perform a regression task where the model is trained to minimize the mean-squared error between the prediction and IMS.
- **SciBERT**: A transformer model trained using masked language modeling on a large corpus of scientific text (Beltagy et al., 2019b). The fine-tuning setup is the same as for the RoBERTa model.

- **CiteBERT**: A SciBERT model further fine-tuned on the task of citation detection, and was shown to have improved performance on downstream tasks using scientific text (Wright and Augenstein, 2021a). The training setup is the same as for the RoBERTa model.

For the fine-tuned models, we train on the entire training set of SPICED, including both news findings and tweets and show test performance overall and split between news and tweets. For the test set we only use manually annotated pairs. Performance is measured in terms of mean-squared error (MSE) and Pearson correlation (r).All results are reported as the average and standard deviation for each model across 5 random seeds.

4.4.2 Results

Paraphrase detection and natural language inference models perform very poorly for zero-shot transfer on this task (Figure 4.2, grey bars), with NLI having slightly better transfer, supporting our hypothesis that transferring from existing tasks to this domain is challenging. Fine-tuned models with Masked Language Model (MLM) pretraining can learn the task decently well (Figure 4.2, red bars), but surprisingly RoBERTa performs just as well as SciBERT and CiteBERT which were specifically pretrained on scientific texts. We posit that this could be due to the fact that RoBERTa was pretrained on a wider range of texts that are reflective of the domains in SPICED, including news texts, while SciBERT and CiteBERT were trained solely on scientific papers.

SBERT models trained on large amounts of pretraining sentences perform well in the zero-shot transfer setup, with the MiniLM based model outperforming MPNet. The best setup was using SBERT fine-tuned on SPICED (Figure 4.2, blue bars), which yields up to 3.9 points gained overall in Pearson correlation and a reduction of 0.3 in terms of MSE (MPNet to MPNet-FT). We also note that there is a large gap between performance on this data and general semantic similarity datasets such as STSB, which see correlation scores in the 90s. As such, there is potentially much room to grow in terms of raw performance on this dataset.

Models performed worse for pairs with tweets versus those from news. This performance difference is in line with our expectations, as there is a large domain shift between tweets and scientific texts and our base models were not exposed to tweets during pre-training. All models, including the zero-shot transfer SBERT models, perform much worse on that split of the data. Additionally, we only see minor gains in performance in terms of MSE for MiniLM when fine-tuned on tweets. We see larger gains for MPNet. Interestingly, the best performance (Pearson r) for Tweets is RoBERTa, though the overall MSE is still best for MPNet-FT. We note the difficulty of these samples, including a need to understand that, for example, higher numbers are equivalent to "an improvement", and needing to hone in on relevant information in the sentences as opposed to extraneous details.

4.5 Application: Modeling Information Change in Science Communication

Whether the media faithfully communicate scientific information has long been a core question to the science community (National Academies of Sciences, Engineering, and Medicine, 2017). Our dataset and models allow us to conduct a large-scale analysis to study information change in science communication. Here, we focus on three research questions:

- **RQ1:** Do findings reported by different types of outlets express different degrees of information change from their respective papers?
- **RQ2:** Do different types of social media users systematically vary in information change when discussing scientific findings?
- **RQ3:** Which parts of a paper are more likely to be miscommunicated by the media?

RQ1-2 focus on the holistic information change captured in IMS, while RQ3 focuses on what types of information might be changing.

4.5.1 RQ1: Comparing Media Outlets

Different types of media target different audiences and tend to report the same issue differently (Richardson, 1990; Mencher and Shilton, 1997). While good science journalism requires outlets to prioritize quality, in real practices, journalists may adopt different writing strategies for different types of audiences (Roland, 2009). Thus, we investigate if findings reported by different types of outlets express different levels of information change, focusing on three types of outlets: General News (e.g., NYTimes), Press Releases (e.g., Science Daily), and Science & Technology (e.g., Popular Mechanics). We use our best-performing MPNet-FT model to estimate the IMS of over 1B pairs and keep those with IMS > 3, which finally leads to 1.1M paired findings from 26,784 news stories and 12,147 papers. We then build a linear mixed effect regression model (Gałecki and Burzykowski, 2013) to predict IMS for matching pairs from news stories and research articles. We include a fixed effect for the type of news outlet, using General News as the reference category. To account for reporting differences across fields and variations specific to highly-publicized papers, we also include a fixed effect for the scientific subject and a random effect for each paper with 30+ pairs (all other papers are pooled in a single random effect).

Results. Compared with General News, Science & Technology news outlets and Press Releases report findings that more closely match those from the original paper (Figure 4.3 shows the regression coefficients). This difference likely is due to some form of audience design where the

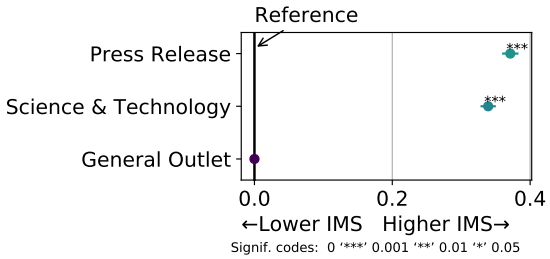


Figure 4.3: Scientific findings covered by Press Release and SciTech generally have less informational changes compared with findings presented in General Outlets

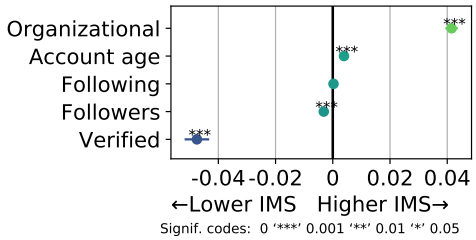


Figure 4.4: Organizational Twitter accounts keep more original information from the paper finding while verified users and those with more followers change more information when tweeting about a scientific finding.

journalist is writing for a more science-savvy readership in the latter two, whereas General News journalists must more heavily paraphrase the results for lay people.

4.5.2 RQ2: Comparing Social Media Accounts

Social media play an important role in disseminating scientific findings (Zakhlebin and Horvát, 2020), so what factors affect the presentation of scientific information on social media becomes an important question. Here, we focus on the types of Twitter users who tweet about scientific findings. Based on 182K matched tweets and paper findings, we again build a linear mixed effect regression model to predict IMS. We include fixed effects of (1) if the account is run by an organization, as inferred using M3 (Wang et al., 2019b), (2) if the account is verified (3) the number of followers and following, both log-transformed, and (4) the account age in years. We use the same field fixed effects and paper random effects as in RQ1.

Results The type of user strongly influences how faithful the tweets are to the original findings (Figure 4.4). Accounts from organizations tend to be more faithful to the original paper findings, which could be due to intentional actions of image management to build trust (Saffer et al., 2013). Surprisingly, verified accounts were far more likely to change information away from its original

meaning; similarly, accounts with more followers had the same trend. Given their prominent roles in Twitter communication (Bakshy et al., 2011; Hentschel et al., 2014), multiple mechanisms may explain this gap such as adding more commentary or trying to translate original scientific findings to lay language to make the findings easier to understand. Our results point to a need to study systematic behavior based on who is communicating science.

4.5.3 RQ3: What Information Changes

Most studies on scientific misinformation focus on paper titles and abstracts (e.g., Sumner et al., 2014), which cannot fully reflect the information presented in the full papers. Analyzing the information change of findings paired from all sections of papers could help to better understand the mechanisms behind scientific misinformation and develop strategies to reduce them. We use the same 1.1M finding pair dataset as RQ1 and analyze what information might have changed using two models trained for changes in scientific communication: identifying exaggerations (Wright and Augenstein, 2021b) and certainty (Pei and Jurgens, 2021).

Results Journalists tend to downplay the certainty and strength of findings from abstracts (Figure 4.5), mirroring the results of Pei and Jurgens (2021). However, this pattern does not persist for findings in other parts of papers, especially the limitations. Existing studies suggest that journalists might fail to report the limitations of scientific findings (Fischhoff, 2012a), and our results here suggest that findings presented in limitations are more likely to be exaggerated and overstated. However, it is also possible that scientists may adopt different discourse strategies for different parts of a paper (Clark, 2013). Nonetheless, our result obviates the necessity of analyzing the full text of a paper when studying science communication.

4.6 Conclusion

Faithful communication of scientific results is critical for disseminating new information and establishing public trust in science. Given the challenge of—and occasional failures in—communicating science, new resources and models are needed to evaluate how science is reported. Here, we introduce SPICED, a new science communication paraphrases dataset labeled with information similarity. Extensive experiments demonstrate that models can predict the degree to which two reports of a scientific finding have the same information but that this is a challenging task even for current SOTA pre-trained language models. In downstream applications, using the trained model to perform a large-scale analysis of information change in science communication, we show systematic behaviors in how different people and news outlets faithfully convey scientific results. Data, code,

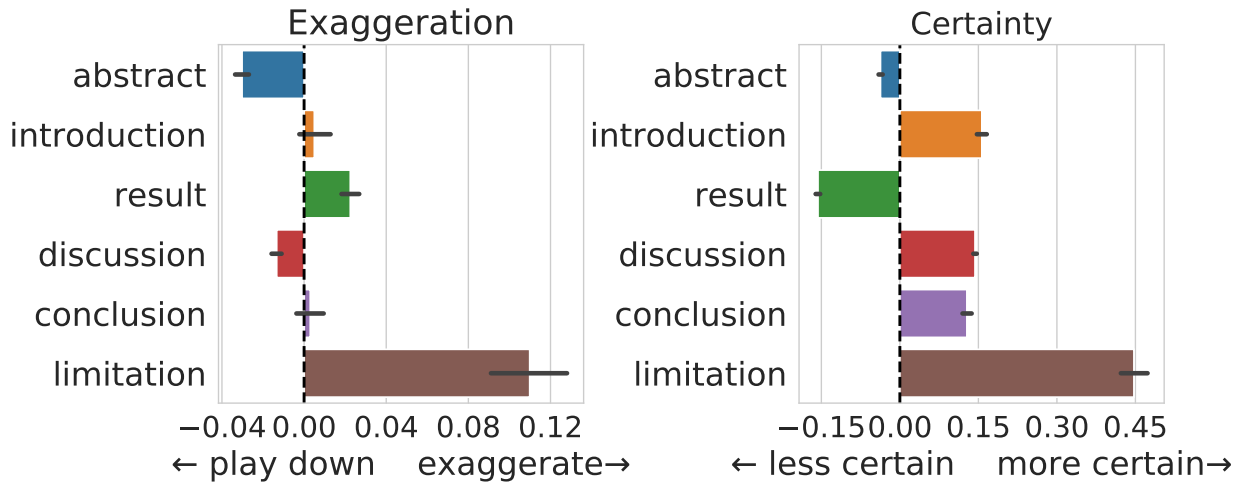


Figure 4.5: Journalists tend to downplay the certainty and strength of findings in abstracts, but overstate findings discussed in limitations sections.

and pretrained models are available at http://www.copenlu.com/publication/2022_emnlp_wright/.

Limitations

We note three limitations of our study. Our data and analysis in social media is limited to only one platform, Twitter, and includes only tweets directly linked to the original paper, as indicated through Altmetric. While Twitter is among the largest social media platforms and is the most common in the Altmetric data, our data potentially omits other kinds of scientific communication about papers that do not directly link to a paper or tweets that link to a paper that cannot be easily identified to a DOI (e.g., linking to a PDF hosted on a personal website). Other types of tweets may be omitted from our dataset such as those written in a thread, or in a tweetorial, about a paper (Gero et al., 2021), which may include additional tweets that describe a paper's findings. While our models would likely still be able to effectively analyze such tweets, these additional forms of scientific communication could add new variety. We leave identifying and collecting such tweets to future work.

Second, our study focuses on only four large scientific fields. While these fields do cover a broad selection of papers, we were unable to annotate additional fields due to the annotation budget and limitations from the Prolific platform. On Prolific, not all potential domains had sufficient numbers of qualified annotators (we required at least a Bachelor's degree in the domain), and the number of unique surveys to run scaled linearly with the number of domains, creating a significant human overhead. However, we will open-source our annotation interface and pipeline and we encourage further efforts to build a larger dataset across more scientific domains.

Finally, while our models achieve moderately high performance at inferring the information matching (Figure 4.2), performance is not perfect, which potentially limits our ability in downstream models and tasks. While we show the data is still useful in training for related tasks (sec:evidence-retrieval) and a trained model can be used to identify systematic behavior by types of users and outlets (sec:social-media-analysis), more accurate models would likely be needed to identify any trends for finer-grained settings, such as looking at the behavior of a specific outlet. For this reason,w e have kept our analyses at a higher level (e.g., outlet categories).

Ethics and Impacts

Miscommunication of scientific information can have negative impacts on many aspects of our society. Our study contributes to a large research program on the science of science communications (National Academies of Sciences, Engineering, and Medicine, 2017). Our dataset and model could be used to keep track of information change in science communication, enable large-scale analyses to understand the current science communication ecosystem, and finally help to facilitate better and more effective science communications.

Crowdsourcing ethics Annotating paired findings requires deep attention and may lead to annotator burnout. We carefully designed our annotation pipeline to provide a good annotation experience for the annotators. We designed a user-friendly Web-based annotation interface that allows annotators to do annotations using keyboard shortcuts. All the annotators are encouraged to leave comments and answer several questions about their annotation experience. More than 95% of the annotators are satisfied with their annotation experience and many people suggest that our study helps them to better understand the science communication process[4] and our annotation interface makes their task easier.[5]

[4]For example, one participant said "Nice learning experience, Helps to understand the news can be far more different then the research paper cited"

[5]For example, one participant said "i liked the option of using my keyboard, it made the experience more comfortable and efficient."

CHAPTER 5

Modeling Public Perception of Science in News Media

5.1 Introduction

Science communication plays a key role in disseminating scientific information and promoting public engagement with science (Fischhoff, 2013). Effectively communicating science to the public holds significance for both scientists and the broader audience outside the science community. On the one hand, individuals living in modern society rely on scientific knowledge to make important life decisions (Bell and Lederman, 2003), and in general, the public shows great interest in scientific information (Durant et al., 1989). On the other hand, public engagement and discussions about scientific information could also help to promote societal understanding of science and inform new scientific developments (Stilgoe et al., 2014)

Despite its crucial role in our society and the considerable efforts devoted to it, effectively communicating science to the public remains challenging and confronts numerous obstacles (Cheng et al., 2008; Scheufele, 2013). One major challenge in science communication is understanding the public (Schäfer et al., 2018). Assumptions about the public represent an important latent dimension that distinguishes between the two classic models of science communication. In the deficit model, the public is simply viewed as a group of individuals that are passive and ignorant, and science communication is to simply "sell science" to the public (Sturgis and Allum, 2004). Such a model has long been criticized because it ignores the variety of the public (Sturgis and Allum, 2004). The engagement model calls for science communicators to better contextualize the communication of science to promote science engagement with diverse groups of people (Stilgoe et al., 2014). However, in practice, understanding the public is not as easy as it seems to be, even for professional journalists (Atkin et al., 1983). Additionally, existing research also suggests that many current science communication practices may only attract people who are already well-informed about certain science issues but fail to engage with a broader and more diverse set of audiences (Goidel and Nisbet, 2006; Nisbet and Scheufele, 2009). Therefore, understanding the

broad audience of science communication could provide further insights into facilitating effective science communication.

Existing studies have explored the interactions between science and diverse audiences from two primary perspectives. Firstly, researchers have investigated how various social factors influence people's attitudes and attention toward scientific information (Weinburgh, 1995; Achterberg et al., 2017; McCright et al., 2013). These studies often employ surveys, where participants rate specific scientific issues such as gene technology (Olofsson et al., 2006) and climate change (Shi et al., 2015). Another relevant area of research examines how different framing and language usage impact people's responses to scientific information (Feldman and Hart, 2021). However, within this line of research, while some studies have explored how pre-existing beliefs or political orientations shape individuals' consumption of scientific information (Yeo et al., 2015; Hart and Nisbet, 2012), there remains a gap in understanding the nuanced interactions between content and individual backgrounds.

In short, we identify two major gaps in existing studies: (1) a tendency to narrowly focus on very specific topics (e.g. climate change), overlooking the nuanced variations across and within domains; (2) the nuanced interactions between people's backgrounds and individual scientific information presented in media are underexplored. (3) lack of datasets and computational tools to understand public perception of scientific information in media at scale. In this study, we aim to help promote public engagement with science by addressing the three questions: (1) What to measure regarding the public perception of scientific information? (2) How to create a large-scale dataset to better understand the public perception of scientific information in media? (3) Can we automatically estimate the potential public perceptions of scientific information in different contexts?

We introduce a computational pipeline designed to assess the public perception of scientific information in media. Firstly, we propose a comprehensive framework comprising 12 dimensions, including NEWSWORTHINESS, UNDERSTANDABILITY, EXPERTISE, INTERESTINGNESS, FUN, IMPORTANCE, BENEFIT, SHARING, READING WILLINGNESS, EXAGGERATION, SURPRISINGNESS, and CONTROVERSY. Following this framework, we curated a large-scale science news perception dataset containing 10,489 annotations from 2,101 participants across 1,506 science news stories. Our dataset draws from representative samples of the US and UK populations, ensuring diversity and representativeness in measuring public perceptions. Each participant evaluated a variety of science news stories based on our proposed framework. Additionally, we developed NLP models to automatically predict public perception of science news. Our experiments demonstrate that a fine-tuned ROBERTA-LARGE model on this dataset reliably estimates the perceived values of scientific information in both news stories and Reddit posts.

Leveraging this model and the dataset, we conduct large-scale analysis to understand the vari-

ations of public perception in different contexts. We found that while the public perception of scientific information depends on the domain of the research, perceived values can also vary at individual article levels, highlighting the need to measure public attitudes toward scientific information beyond topics and domains. Furthermore, we found that individuals' science news consumption frequency and trust in science are the strongest factors affecting their perceptions of science news stories. In the end, by analyzing a large-scale social media dataset of scientific information, we found that the estimated perceptions correlate with public engagement with scientific information. For example, posts with higher estimated SURPRINGNESS, IMPORTANCE, and FUN tend to receive more upvotes and comments on Reddit. Overall our analysis highlights the significance of modeling public perception of scientific information in the media. It not only helps to better understand the public interest in science but also holds the potential to predict engagement patterns with scientific information.

Overall, our study makes the following contributions:

1. We create a novel resource to model direct public perceptions of scientific information presented in media.
2. We build NLP models that are able to analyze the public perception of science news stories
3. We demonstrate that the curated dataset and the model could help to address important research questions on science communication.

5.2 Framework for Modeling the Public Perception of Scientific Information in Media

Modeling public perception of scientific information is a key step towards effective science communication (Longnecker, 2023). However, public perception is a subjective and multi-dimensional construct. Here, drawing from existing literature on news values, public understanding of science, and science communication, we develop a comprehensive framework to analyze the public perception of scientific information in media.

NEWSWORTHINESS In science journalism practices, newsworthiness is regarded as a core dimension that guides the selection and coverage of information (Harcup and O'neill, 2017). Similarly, scholars have studied the news values of scientific information in news stories (Badenschier and Wormer, 2011) or research articles (Nishal and Diakopoulos, 2022). While many existing studies focus on modeling journalists' understanding of newsworthiness, some also emphasize the importance of studying perceived newsworthiness in various settings (Lee and Chyi, 2014; Zhang

et al., 2016). Here, we approach the newsworthiness of scientific information from the perspective of the general public. Specifically, our goal is to measure whether the public views a news article as "should be published in news" or not. NEWSWORTHINESS represents the holistic evaluation of the perceived value of the presented information

UNDERSTANDABILITY In science communication, presenting content that is accessible and comprehensible to the audience is crucial, as the value of scientific information in news is contingent upon its understandability (Fischhoff, 2012a). Without clear comprehension, the lay public is unlikely to benefit from the scientific news presented and further participate in the public discussion about it (Varner, 2014). However, despite existing efforts to make science more accessible to the public, existing studies suggest that understandability remains a challenge in current science news reporting (Morosoli et al., 2024). Therefore, it is important for us to measure how the public perceives the understandability of science presented in the media.

EXPERTISE Scientific information usually contains jargon or specialized knowledge that the general public may not easily understand (Hirst, 2003; Sharon and Baram-Tsabari, 2014). Here, we aim to quantify whether a given science news story requires specialized knowledge for the public to understand. UNDERSTANDABILITY models an individual's capability to understand the content, which could also depend on the individual's background in the relevant topic. Different from understandability, EXPERTISE focuses on the relatively more objective "difficulty" of the content.

IMPORTANCE While scientific findings may be highly valued within specific scientific communities, their significance to the broader society can vary. Thus, assessing whether the science covered is of sufficient importance to the general audience is crucial in understanding the overall perception of science news (Burns et al., 1995). Importance has been considered as a key dimension in evaluating science news values (Badenschier and Wormer, 2011) and here we focus on the public perception of scientific importance. By doing so, we aim to gain insights into how the public perceives the significance of research outputs presented in news media.

FUN Entertainment is recognized as an essential news factor in journalism (Harcup and O'neill, 2017). Despite science often being seen as a serious topic, making scientific information engaging is essential for science communicators to foster public engagement with science (Stilgoe et al., 2014; Fischhoff, 2013). Therefore, assessing the entertainment value or the fun aspect of reading science news stories is important in evaluating the overall perceptions.

SURPRINGNESS Surprisingness, characterized by content that contrasts with people's existing knowledge or perceptions, is widely recognized as a core dimension of news value (Galtung and Ruge, 1965). Here, we aim to measure how the public perceives the surprisingness of scientific information presented in media. An existing study suggests that findings that surprise the audience may be deemed to have higher news value (Htoo et al., 2023), as they not only inform but also engage by introducing novel perspectives or unexpected information.

CONTROVERSY Scientific information in the media can be controversial due to the natural uncertainty of scientific research and the significance of the topic it tackles (Sarewitz, 2004; Friedman et al., 2012). Furthermore, existing studies also suggest that controversies discussed in the media can also potentially affect scientific publications (Stewart, 2008). Therefore, it is important for us to measure how the public perceives the scientific information presented in news media to be controversial.

EXAGGERATION Science reporting has long grappled with the issue of exaggeration or hype (Caulfield and Condit, 2012). While exaggeration may initially attract more audience attention, existing studies suggest that perceived exaggeration could erode public trust in science (Intemann, 2022; Neureiter et al., 2021). Hence, measuring how the public perceives exaggerations in science news stories is crucial for a comprehensive evaluation of scientific information in the media.

INTERESTINGNESS INTERESTINGNESS represents a subjective but essential dimension in assessing science news, tailored to the preferences of specific audiences. In journalism research, perceived interestingness is considered a core factor affecting people's perceptions of overall newsworthiness (Lee and Chyi, 2014). One of the primary goals of science communication is to engage the public with information that captures their interest, therefore it is essential for us to measure the perceived interestingness of scientific information presented in news articles.

BENEFIT Scientific knowledge plays a key role in informing individuals' decision-making and facilitating societal transformations (Yin et al., 2022). Consequently, the direct applicability or usability of science news stories becomes a critical dimension of their value (Badenschier and Wormer, 2011). This dimension underscores the importance of presenting scientific findings in a way that is not only informative but also practically relevant to the audience's daily lives and decisions.

SHARING WILLINGNESS Social media platforms have gradually become one of the primary sources of news information for the general public (Shearer and Grieco, 2019). Therefore, studying

the sharability of news stories and users' sharing behaviors has attracted attention from journalism scholars (Bednarek, 2016; Harcup and O'neill, 2017). In this context, we consider the likelihood of readers sharing scientific information as an important dimension to measure the overall perception of science news, reflecting the impact and reach of science communication in the digital age. We use SHARING to represent this category in the rest of the paper.

READING WILLINGNESS Reading represents one of the most basic types of engagement with scientific information. READING WILLINGNESS measures whether the public would potentially be willing to read about the content in different venues. Good science news should attract the audience and prompt them to read further about the topic. Articles that the audience does not even want to read are less likely to facilitate further discussions and engagements with the corresponding topic. Therefore, we consider reading willingness as an important dimension in our framework. We use READING to represent this category in the rest of the paper.

Collectively, the 12 dimensions provide a comprehensive evaluation of public perceptions of scientific information and can be used to measure science presented in diverse types of media including news stories, social media posts, and research articles.

5.3 Building a New Dataset to Model Public Perception of Scientific Information

Data resources are key to scientific research. To support long-term research on public perception of scientific information and promote public engagement with science, we create a new large-scale dataset containing nuanced ratings for individual science news stories as well as detailed background information of the participants. In this section, I describe the creation process of this dataset.

5.3.1 Raw Dataset

Our raw dataset is obtained from altmetric.com which contains 2.4 million media mentions of research papers. We follow the processing pipeline from (Peng et al., 2020) and further extract a subsample of 128,942 news/article pairs where we are able to collect the full text of the news articles. Our data spans 273 different news outlets and 57,807 different scientific articles in over 100 scientific domains.

Category	Statement	Description
NEWSWORTHINESS	The science news story should be published in the news	In journalism, newsworthiness is considered as the core dimension of news stories and is usually operationalized as "worthy of being published as news" (Caple, 2018). Here we measure the perceived newsworthiness from the perspectives of the public.
UNDERSTANDABILITY	I can understand the science news story	Understandability measures whether a given science news story can be understood by the public. Existing studies consider understandability as an important dimension for news stories (Dalecki et al., 2009; Agarwal, 1992).
EXPERTISE	Understanding the science news story requires specialized knowledge	Scientific information may involve specialized knowledge and the EXPERTISE dimension measures whether a given news story requires specialized knowledge to understand.
IMPORTANCE	The science news story tackles an important issue	Importance has been considered as one of the core value factors of science news stories (Badenschier and Wormer, 2011). Here we measure how the public perceives the importance of the scientific information presented in a news article.
FUN	The science news story is fun to read	Entertainment has long been considered as one of the core news values in journalism (Harcup and O'neill, 2017) and the public also actively perceives the entertainment value of science (Bourk et al., 2018). Here we measure whether a given scientific information is perceived as fun and entertaining to the public.
SURPRISINGNESS	The scientific finding seems surprising to me	Science sometimes provides surprising results that contradict people's existing beliefs and surprisingness has been considered as an important factor for science news values (Badenschier and Wormer, 2011). Here we measure whether the public perceives a given science news story as surprising.
CONTROVERSY	The scientific finding could be controversial	Science tackles many controversial topics (Narasimhan, 2001) and here we measure whether a given science is perceived as controversial by the public.
EXAGGERATION	This science news story is overstated or exaggerated	Exaggeration is a common issue in science reporting (Caulfield and Condit, 2012) and can potentially undermine the public trust in science (Intemann, 2022). Here, we measure the subjective perceptions of exaggeration in scientific information.
INTERESTINGNESS	The science news story sounds interesting to me The science news story could be interesting to the general public	Perceived interestingness is considered as a key factor affecting people's perception of the overall newsworthiness (Lee and Chyi, 2014). Here we measure the perceived interestingness from two dimensions: (1) the perspective from an individual and (2) the perceived interestingness for
BENEFIT	Who will benefit from learning about this finding? - The general public - A segment of the public - Policy makers - Companies in the related industries	Benefit is one of the core dimensions of science news values (Badenschier and Wormer, 2011). Here, we measure the perceived benefits of a given science news story from multiple latent dimensions.
	I learned something useful from the science news story Knowing about this science could benefit a lot of people Knowing about this science could benefit a specific segment of the population	
SHARING	I would share this science news story with someone I know directly I would share this science news story with a wider forum like a mailing list, Twitter, Reddit I would be unlikely to share this science news story with anyone	The sharability of a news article has been considered as an important dimension of the news values (Harcup and O'neill, 2017). Here we measure the public's sharing willingness in different imagined settings (e.g. forums and friends).
READING	If I'm browsing news articles from __ , I would be likely to open and read the science news story above -General news outlets (e.g. BBC, New York Times, Fox News) -Science and Technology media (e.g. Scientific American, National Geographic) -Other popular printing media (e.g. Vogue, GQ, Elle) -Other popular media (e.g. TV and radio) -It should not be published in public media outside the science community	Reading willingness measures the audience's willingness to read the presented information in different venues (e.g. general news outlet, science and technology media and other popular media). Modeling this dimension could further help to understand the public's interest in reading about science news stories.

Table 5.1: The twelve-dimension framework to measure public perceptions of scientific information in media.

5.3.2 Crowdsourcing with Representative Samples

To measure the public perception of science news stories, we designed the following pipeline to create a large-scale crowdsourcing annotation task.

Data Cleaning For our annotation task, we aim to present the full news articles while excluding information that might bias the annotators' perception of scientific information. To this end, we designed a data-cleaning pipeline that removes the following types of information: (1) Outlet: people's perception of news articles might be biased by their pre-existing attitudes towards the outlet (2) Other meta information like author, date, city, and URL, as such information might create an impression that the presented news article has already been published, which may further bias the ratings.

Data Sampling for Annotation The overall goal of our data sampling is to provide a generally balanced sample for annotation across different settings. We design a sampling procedure that balances the type of outlet, the science domain, and the popularity of the paper. Below we describe our sampling process for a batch of around 1000 news articles:

1. Sampling based on coverage types. We first categorize each outlet into three categories: General, Press Release, and SciTech blogs. Furthermore, we balancedly sample the same amount of papers (e.g. 80 in this case) from three settings: (1) mentioned in all three types of outlets (2) mentioned in both PR and SciTech outlets (3) mentioned in both PR and General outlets. For each selected paper, we randomly sample 1 article from the corresponding category. Such a process will lead to 560 sampled news articles.

2. Sampling based on outlet types. We further expand our sample by randomly sampling 50 news articles from each type of outlet separately. This step leads to 810 news stories.

3. Upsampling based on domains. To balance the stories in different science domains, we further upsample another 50 news stories in Social Science, 50 in Humanities, and 100 in Engineering, which leads to 1,010 stories.

4. Upsampling based on popularity. Previous steps may lead to a sample of papers with relatively low popularity, which might have lower news values than papers that are covered in many news stories. In this step, we further upsample another 30 news articles that cover popular research articles (coverage count larger than 30).

After this process, we are able to obtain a sample of news stories that are balanced for the type of coverage, type of outlet, domain, and popularity of the paper.

Category	Statement	K's α	C's α
NEWSWORTHINESS	Should be published in news	0.130	
IMPORTANCE	Tackles an important issue	0.311	
EXPERTISE	Needs specialized knowledge	0.248	
UNDERSTANDABILITY	I understand the news story	0.185	
FUN	News story is fun to read	0.122	
CONTROVERSY	Finding could be controversial	0.147	
SURPRISINGNESS	Finding is surprising to me	0.096	
EXAGGERATION	This science news story is overstated or exaggerated	0.045	
INTERESTINGNESS	Could interest the general public	0.137	0.77
INTERESTINGNESS	The science news story sounds interesting to me	0.092	
BENEFIT	Could benefit the policy makers	0.233	
BENEFIT	Could benefit many people	0.220	
BENEFIT	Could benefit related industry companies	0.156	0.84
BENEFIT	Could benefit the general public	0.166	
BENEFIT	Could benefit a segment of the public	0.113	
BENEFIT	I learned something useful	0.082	
SHARING	I would share this science news story with someone I know directly	0.050	
SHARING	I would be unlikely to share this science news story with anyone	0.024	0.81
SHARING	I would share this science news story with a wider forum like a mailing list, Twitter, Reddit	0.022	
READING	Would be willing to read it from General news outlets (e.g., BBC, New York Times, Fox News)	0.083	
READING	Would be willing to read it from Other popular media (e.g., TV and radio)	0.074	
READING	Would be willing to read it from Other popular print media (e.g., Vogue, GQ, Elle, People, Time and Reader's Digest)	0.082	0.83
READING	Would be willing to read it from my normal feed/content/timeline on a social media site (e.g., Facebook, Instagram and Twitter)	0.047	
READING	Would be willing to read it from Science and Technology media (e.g., Scientific American, National Geographic)	0.070	
READING	It should not be published in public media outside the science community	0.012	

Table 5.2: Krippendorff's α and Cronbach's α for rating the public perceptions of science news stories.

Particpants We recruit participants from Prolific.com as it allows us to collect annotations from representative samples for the US and UK population. More specifically, we recruit 1,057 participants from the US and 1,044 participants from the UK. Both are drawn from a representative sample balanced for age, sex, and race. We require each participant to be fluent in English to participate in our study as the science news articles are written in English. For each news article, we set to collect 4 labels from the US and the UK sample separately.

Annotation Procedure We first ask each annotator to read the introduction of our study and then consent to participate in the annotation. Once they agree to participate, we display 5 randomly sampled science news stories in a sequence. To measure the public perception of science news values, we follow the framework introduced in Sec 5.2 and ask each annotator to rate the given science news stories based on the 25 statements in Table 5.1. Each annotator is asked to rate the 25 statements for each science news story following a 1-5 Likert scale, where 1 means strongly disagree and 5 means strongly agree. After the annotation phase, each annotator is asked to answer a list of questions regarding their demographic information, educational background, experience/interest in science, and political orientation.

5.3.3 Quality Check and Post Processing

Our final dataset contains 10,489 annotations for 1,506 news articles. To evaluate the quality of the created dataset, we calculate Krippendorrf's α and Cronbach's α to evaluate the reliability of

the inter-annotator agreement and the internal consistency of the scales.

Inter-annotator Agreement Table 5.2 shows the inter-annotator agreement as calculated by Krippendorff's α (Hayes and Krippendorff, 2007) for each dimension. The average IAA across all the statements is 0.11, suggesting that raters' perception of science news stories is subjective. Despite its subjectivity, we observe that the IAA for rating certain statements is generally higher, for example, the IAA for "The finding needs specialized knowledge", "Could benefit many people" and "Tackles an important issue" are generally above 0.2, suggesting that the annotators have more agreement when rating relatively more objective dimensions. While the overall IAA seems low in our data, it is worth noting that a similar level of IAA was observed in a similar annotation task conducted by Nishal and Diakopoulos (2022). Nishal and Diakopoulos (2022) collected newsworthiness ratings from both expert raters and crowd workers. The IAA between expert raters is 0.16 and ranges from 0.09 to 0.18 for crowd workers, suggesting that the low IAA in our task is potentially caused by natural subjectivity instead of low-quality annotations.

Internal Consistency To evaluate the internal consistency of the instruments used for measuring public perceptions, we calculate Cronbach's α (Tavakol and Dennick, 2011) for each group of statements including INTERESTINGNESS, BENEFIT, SHARING, and READING and the C's α is 0.77, 0.84, 0.81, 0.83 respectively, suggesting that the scales used to measure public perceptions are generally reliable.

Post Processing For dimensions measured with multiple latent statements (e.g. READING and SHARING), we average all the relevant ratings from an annotator to create a unified score for the dimension. After this step, each news article receives 12 scores from a single annotator. We then average the ratings of all the annotators to obtain the overall perception value. As a result, the final dataset contains perception scores for 1,506 news stories, and each story has 12 scores representing the public perceptions of it on the 12 dimensions.

Usage for Modeling Training One of the final goals of our study is to build NLP models that are able to automatically estimate public perceptions of scientific information. Therefore, it is worth discussing the suitability of using this data for building machine learning models. First of all, our post-processing steps ensure that we are modeling the mean relative perceptions instead of the numerical ratings. More specifically, when a news article receives a high average newsworthiness rating, it is more likely to be more newsworthy than another article with a low average score. To quantify the reliability of our final average scores as a relativity measurement of news perceptions, we follow a similar procedure as Dittrich et al. (2007) to convert the Likert ratings into ranking

scores. We then calculate the correlation between the average rating and the ranking score on all the dimensions. The average correlation across all the dimensions is 0.8, suggesting that the average rating could reliably measure the relative perception orders across articles. Secondly, while we observe relatively low inter-annotator agreement, existing studies have suggested that low-IAA data can still be used to train reliable machine learning models (Reidsma and Carletta, 2008; Artstein and Poesio, 2008). For example, in a similar task modeling the newsworthiness of research articles (Nishal and Diakopoulos, 2022), while the IAAs between crowd workers are around 0.1, a model trained on the annotated data could still precisely predict the expert ratings for newsworthiness. As we are following a similar post-processing pipeline as (Nishal and Diakopoulos, 2022), we argue that our annotations, while with relatively low IAA (from 0.01 to 0.3), could still be used to train NLP models for automatically predicting public perceptions of scientific information.

5.4 Computational Modeling of Public Perceptions

To model the public perceptions of scientific information at scale, we develop a data-driven NLP pipeline that is able to estimate the public perception of each science news article.

Model To allow automatic modeling of perceived news values for each news story, we fine-tune a RoBERTa-Large model (Liu et al., 2019a) on the annotated dataset. We split the annotated data into train, validation, and test sets following the ratio of 7:1:2. We concatenate the title and the body of each news article as the input document. We follow a multi-task regression setting and on the output layer, the model predicts a score ranging from 1-5 for each statement so that for each science news article, we could directly obtain the predictions for all the dimensions. We fine-tune each model for 10 epochs and then choose the best-performing model on the validation set as the final model and further evaluate the model performance on the hold-out test set.

Result Table 5.3 shows the model performances on the test set. We found that the model is generally able to predict the perception of science news stories on most dimensions. The correlations for 7 out of the 12 dimensions are above 0.5. However, the model performs poorly on dimensions that are relatively more subjective. For example, the prediction performances for exaggerating and sharing are only 0.31 on news stories. Overall our experiment suggests that automatically measuring the public perceptions of scientific information in media is doable but also challenging.

Model Generalizability on Social Media Text To evaluate the model performance in other domains, we created another test set for Reddit posts and tweets mentioning research articles or news stories. A similar annotation pipeline was adopted to annotate 50 Reddit posts and 50 tweets about

Dimension	News	Reddit	Twitter	Paper
NEWSWORTHINESS	0.54	0.69	0.42	0.52
READING	0.44	0.52	0.49	
UNDERSTANDABILITY	0.71	0.75	0.79	0.65
EXPERTISE	0.74	0.40	0.39	0.66*
FUN	0.60	0.76	0.72	
IMPORTANCE	0.77	0.62	0.20	
INTERESTINGNESS	0.52	0.34	0.14	
SHARING	0.31	0.69	0.48	
BENEFIT	0.67	0.35	0.24	
SURPRISINGNESS	0.41	0.56	0.57	
CONTROVERSY	0.49	0.16	0.05	0.44
EXAGGERATION	0.31	0.38	0.16	
Overall	0.54	0.52	0.39	

Table 5.3: Model performances as measured by Pearson's r over different test sets. The model generally performs well on News and Reddit posts but performs poorly on tweets. "Paper" refers to the Arxiv abstract newsworthiness dataset created by Nishal and Diakopoulos (2022). We calculated the correlation between our model's predictions and crowd ratings for the corresponding category. Overall our model also shows good performance on the Arxiv abstract newsworthiness dataset. * means that an absolute correlation score is used because the two measured dimensions are conceptually opposite.

scientific information. Each post is annotated by 8 annotators and Table 5.4 shows the internal consistency and inter-annotator agreement for the two datasets. We further run the trained model over this dataset and the results are shown in Table 5.3. Overall we observe that the model performs well on Reddit posts but performs poorer on tweets. This is potentially because the tweets are usually shorter and may contain more noise.

Model Generalizability on Paper Abstracts To test the generalizability of our model on research articles, we also evaluate model performance on the Arxiv newsworthiness dataset created by Nishal and Diakopoulos (2022), which contains readability, newsworthiness, and controversy labels for computer science paper abstracts. We calculate the correlation between our model predictions and the ground truth crowd labels for overlapped dimensions. More specially, for NEWSWORTHINESS and CONTROVERSY, we directly calculate the correlations. For UNDERSTANDABILITY and EXPERTISE, we calculate the correlation with the readability labels in the Arxiv newsworthiness dataset. As shown in Table 5.3, we found moderate to high correlations between our model predictions and the ground truth labels. For example, the Pearson's r for NEWSWORTHINESS is 0.52 and is 0.65 for UNDERSTANDABILITY. As a comparison, the correlation between the ground truth expert ratings is 0.33. Since the Arxiv abstracts were all sampled from computer science papers, such a result suggests that our model is able to capture nuanced differences for content in the same science domain despite the low IAA of the annotations.

	K's α Reddit	**K's α** Twitter	**C's α** Reddit	**C's α** Twitter
Newsworthiness	0.41	0.32		
Understandability	0.36	0.45		
Expertise	0.33	0.53		
Fun	0.31	0.06		
Importance	0.60	0.46		
Surprisingness	0.26	0.13		
Controversy	0.31	0.28		
Exaggeration	0.08	0.23		
Interestingness	0.40	0.33	0.79	0.75
Benefit	0.43	0.41	0.86	0.84
Sharing	0.15	0.23	0.73	0.81
Reading	0.30	0.24	0.81	0.77

Table 5.4: Krippendorff's α and Cronbach's α for rating the public perceptions of Reddit and Twitter posts. We observe similar or higher levels of reliability and internal consistency compared with news annotations.

5.5 Applications

How the public perceives scientific information has long been a core question for the science community (National Academies of Sciences, Engineering, and Medicine, 2017). Our dataset and model allow us to analyze the nuanced interactions between individuals and scientific information in the media. In this section, we explore the potential usage of our dataset and model for science communication research. More specially, we focus on three research questions: **RQ1:** Does the public perception of scientific information in news media vary across and within fields? **RQ2:** What background factors have the largest effect on individuals' ratings for science news stories? **RQ3:** Can we use the estimated perceptions to predict public engagement with scientific information on social media?

Overall these three questions touch on different aspects of science communication. RQ1 focuses on the content side and examines how the public perceives scientific information across and within fields. RQ2 focuses on the audience of science communication and could help to better understand the most important factors affecting public reception of scientific information in the media. In the end, RQ3 leverages the developed NLP model for estimating public perceptions of scientific information and could further help to provide new insights on predicting public engagement with science on social media. Collectively, the three studies could generate new insights about public understanding and engagement with science.

5.5.1 RQ1: Public Perceptions of Scientific Information Across and Within Fields

How the public receives and uses scientific information from different fields has long been an important question for the science community (Bauer et al., 1994; Boulter, 1999). However, most of the existing studies either focus on the public attitudes toward science as a broad concept (Bauer et al., 1994) or focus on very specific issues like climate change (Cheng and Gonzalez-Ramirez, 2021). Therefore, it is unclear how the public perception of scientific information varies across and within scientific fields. Here, we focus on studying the role of scientific fields in affecting people's perception of scientific information in media. More specially, whether the public perception of science news stories varies across and within scientific domains. Leveraging the created science news perception dataset, we analyzed the public's ratings for scientific information from five broad domains: Engineering, Science, Social Science, Medicine, and Humanities. It is worth noting that we use the phrase "scientific information" to refer to any knowledge produced from scholarly research and therefore consider humanities as a special domain similar to (Yin et al., 2022).

Result Are public perceptions of science field-dependent? As shown in Figure 5.1, the public perceptions of scientific information vary across different fields and dimensions. For example, while humanities and social science are considered more understandable and fun, the public perceives them to be less important and newsworthy than medical research. On the contrary, while medical news is perceived as more beneficial and important, the public also perceives it to be more controversial, potentially because of the nature of the medical research (Kopelman, 1978). Despite the field-level differences, public perception also varies largely within each field. For example, while medical news stories are perceived as more important than other fields in general, as shown in Figure 5.1, outliers still exist and are perceived as not important by the public. Overall, our result suggests that while the public perception of scientific information is field-dependent, people's attitudes toward scientific information still vary at the individual article level. Moreover, public attitudes towards scientific information are multi-dimensional and can vary largely across dimensions. Therefore, it is important to move beyond field- or topic-level measurement of public attitudes and directly model the public perceptions of scientific information in a more fine-grained manner.

5.5.2 RQ2: Background Factors Affecting Perceptions of Scientific Information

Public attitudes toward science are subjective and can be affected by individuals' backgrounds like gender (Weinburgh, 1995) and education (Bak, 2001). However, existing studies majorly focus on

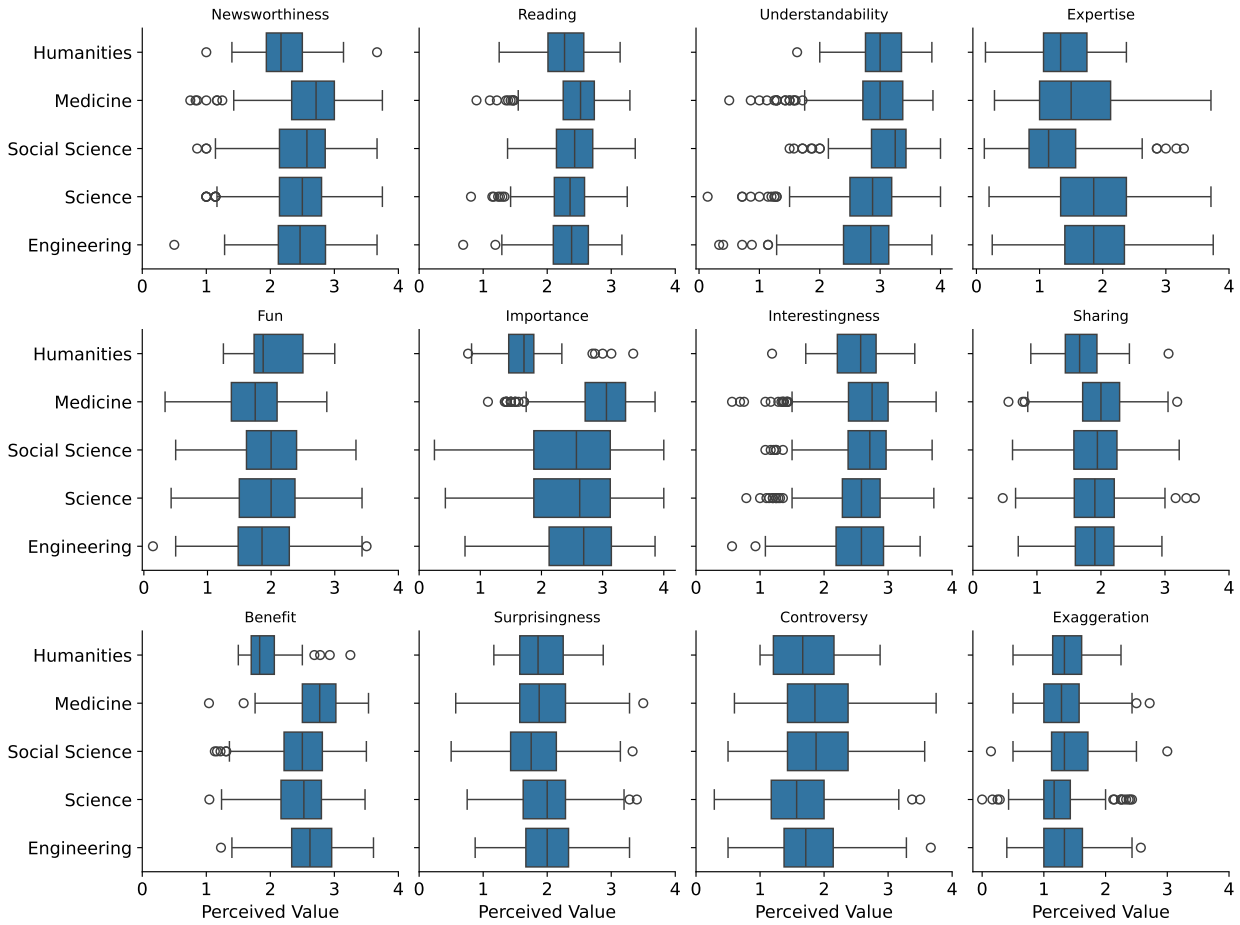


Figure 5.1: Public perception of scientific information varies within and across scientific fields.

peoples' broad attitudes toward science, and it is unclear how different background factors affect people's perceptions of scientific information presented in media. In this study, leveraging the nuanced background information as well as individual ratings for science news stories, we explore the connection between background factors and public perceptions of scientific information in news articles.

Experiment Setup To understand how background factors affect people's perceptions of individual science news stories, we construct separate mixed-effect regression models leveraging the background factors to predict each dimension of public perceptions. The background factors include three major dimensions: (1) demographic information, including gender, age, and educational level (2) experience with science, including science news consumption frequency and trust in science (3) political orientations, including the attitudes towards women's reproductive health rights, legalizing drugs, government investments in public services, gender identity rights, taxing the rich and government interfering the market. To account for the effect of article-level variations,

we control the article id as the random effect. Additionally, we also control the scientific field and the outlet type as fixed effects. Figure 5.2 presents the aggregated results from the regressions.

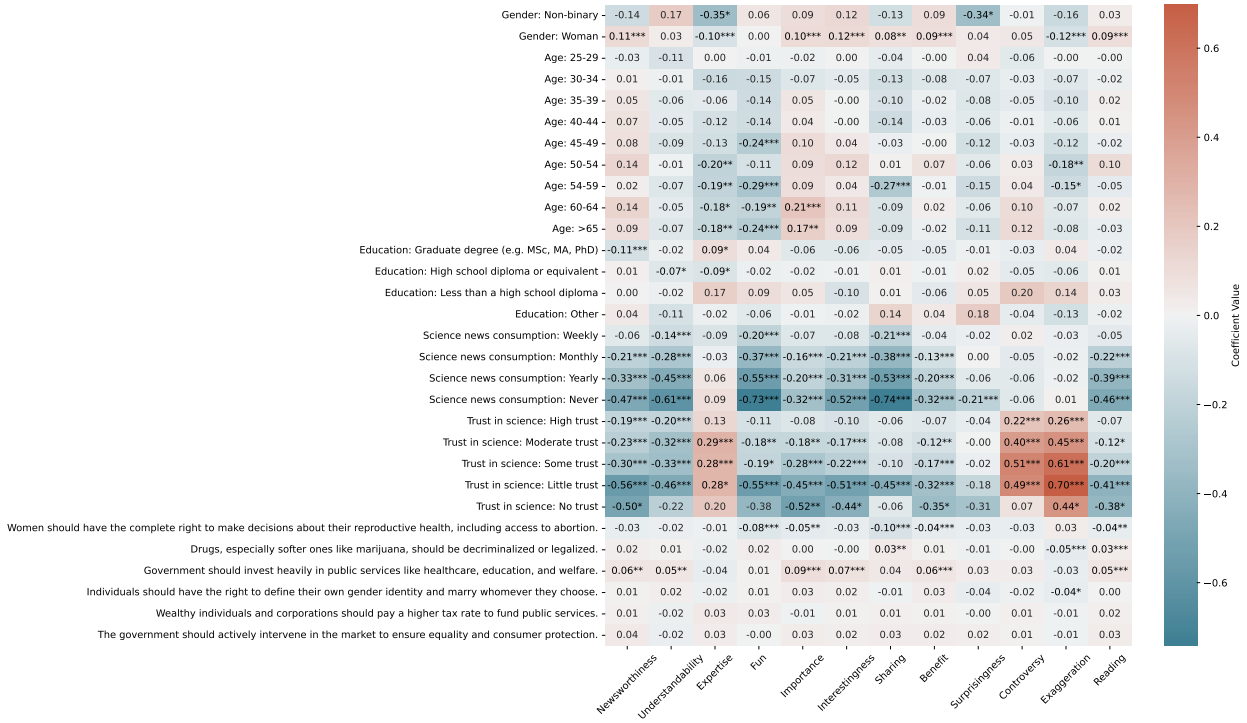


Figure 5.2: Numbers in each cell show the regression coefficient. Experience with science and individuals' backgrounds could both affect people's perception of scientific information. People's science news consumption frequency and trust in science have the largest effect sizes.

Demographics Compared with men, women tend to perceive the same scientific information as more newsworthy, more important, and more interesting. While some existing studies suggest that men hold more positive attitudes towards science (Weinburgh, 1995; Jones et al., 2000) our results suggest that women may actually show greater interest in scientific information presented in media with the control of other content and background factors. For age, we found no significant perception differences for most of the dimensions across age groups when controlling for other content and background factors. Such a result suggests that age may not be a factor that affects people's perceptions of scientific information.

Furthermore, we found that for most of the dimensions, the educational level does not significantly affect people's perceptions of science news stories. Moreover, while existing studies generally suggest a positive correlation between people's education level and their attitudes towards science (Osborne et al., 2003), we observe that people with a graduate degree may actually consider the same scientific information with lower newsworthiness compared with individuals

with a college degree. One potential explanation is that people with a higher degree may have a more stringent perception of news values due to their graduate-level training.

Political orientations Existing studies suggest that individuals' political orientations could affect their attitudes toward science (Blank and Shaw, 2015). Do political orientations also affect people's perception of scientific information in news media? In this study, we explore the connection between individuals' attitudes toward a series of politician orientation statements and their ratings of scientific information. The last six rows of figure 5.2 show the results from our regressions. Surprisingly, most of the political orientation factors do not significantly affect individual ratings for science news stories besides support for women's reproductive health rights and government investment in public services. People with higher levels of support for government investment in public services are more likely to give a higher rating of newsworthiness, importance, and interestingness for the same science news story. Additionally, a higher degree of support for women's reproductive health rights is associated with a decreased level of perceived fun and sharing willingness. Such a result suggests a nuanced connection between peoples' attitudes towards politics and scientific information in the media. However, the effect sizes are relatively small compared with science news consumption and trust in science, suggesting that political orientations may have very small or no effects on people's perception of science news stories in general.

Experience with Science As shown in Figure 5.2, both science news consumption frequency and trust in science significantly affect individuals' rating for scientific information. For example, people who consume science news on a daily basis tend to rate the same scientific information 0.47 higher than people who never consume science news stories on a 5-point scale. The effect size becomes even larger for "Fun" and "Sharing", suggesting that people's science news consumption frequency significantly affects their tendency to enjoy and share science. Similarly, trust in science also has a large effect on people's perceptions of scientific information regarding newsworthiness and interestingness. Additionally, trust in science significantly affects individuals' ratings for controversy and exaggeration, while science news consumption does not. Such a result suggests that prior trust in science could potentially affect individuals' trust in scientific information. Overall our results suggest that individuals' science news consumption frequency and trust in science play a significant role in shaping people's perceptions of scientific information compared with other background factors.

5.5.3 RQ3: Estimated Perceptions and Public Engagement with Science

One of the fundamental goals of science communication is to promote public engagement with scientific information (Stilgoe et al., 2014). Social media platforms offer unique opportunities

for the public to engage with scientific information. However, predicting public coverage and engagement with science has been notoriously hard (MacLaughlin et al., 2018) because many types of information are competing for the public's attention in the current media environment (Scheufele, 2013). Here, we explore the connection between estimated perceptions and public engagement with scientific information on Reddit, a popular social media platform that contains many active science-sharing communities.

Experienment Setup To study the connections between estimated perceptions and public engagement with social media posts about science, we construct a dataset using Reddit posts. We first collect all the posts from 2017 to 2021 shared in r/science, a subreddit where users actively share science news stories. We then extract all the shared links in these posts and further extract all the posts on Reddit mentioning the links. We further removed all the duplicated posts and URLs that are only shared once. The final dataset contains 95,465 Reddit posts mentioning 29,099 science-related content. We conduct two analyses to understand the connection between the perceived values and engagements: (1) a descriptive analysis showing the overall trends and (2) a regression analysis controlling for multiple confounders. More specifically, we construct mixed effect models using the estimated perceptions to predict two engagement metrics: the number of comments and the post scores, both log-transformed. Post scores are calculated by subtracting the number of downvotes from the total upvotes. Additionally, comments about scientific information represent an important engagement with scientific information as they mark open discussions about science. We control the URL as a random effect and also control the subreddit, the domain of the URL, and whether a URL was shared in a subreddit for the first time as a fixed effect. Furthermore, given that public perception correlated with each other, we followed the step-wise regression procedure to gradually remove the variables with the highest variance inflation factor score. The final regression contains five perception dimensions (i.e. IMPORTANCE, SURPRISINGNESS, FUN, CONTROVERSY and EXPERTISE) and other control variables.

Results Figure 5.3 and Figure 5.4 show the overall trend of estimated perceptions and public engagement patterns on Reddit. In general, public engagement with science-related posts is positively correlated with the estimated perceptions on most of the dimensions. For example, posts with higher estimated newsworthiness are significantly more likely to receive more upvotes and comments. On the contrary, posts that require specialized knowledge to understand and receive lower scores and fewer comments. Such a result suggests that the estimated perception values could be used to predict public engagement with science-related posts on social media.

Figure 5.5 and Figure 5.6 show the results of the regression analysis. In this setting, we explore the effect of estimated perceptions on the final engagement when posts are sharing the same

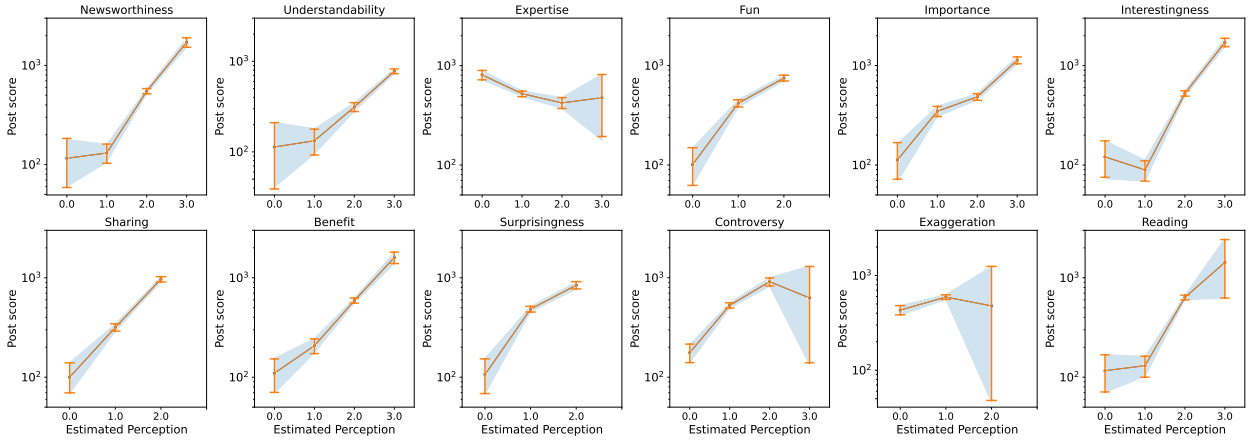


Figure 5.3: Posts with higher estimated perception values receive higher scores on Reddit.

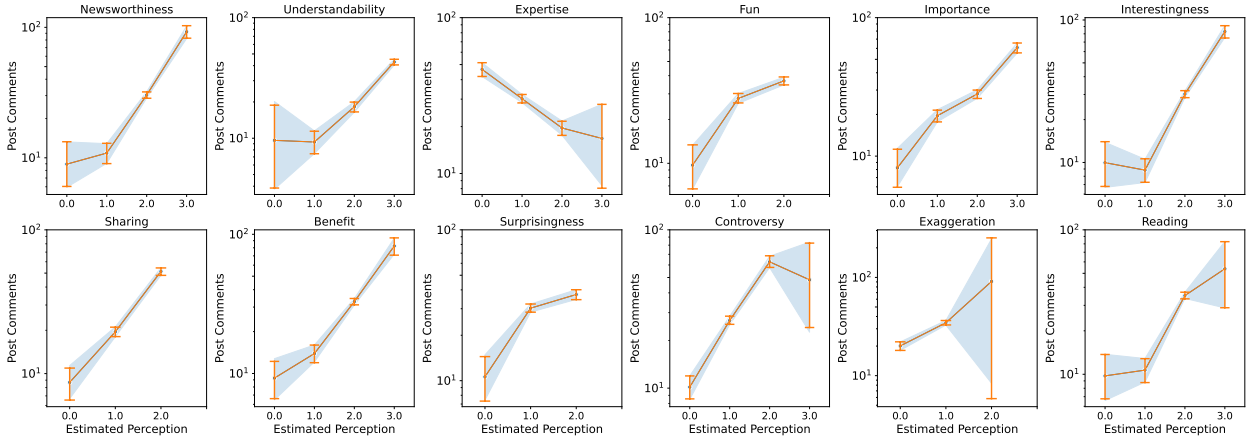


Figure 5.4: Posts with higher estimated perception values receive more comments on Reddit.

science news stories. We found that posts with higher importance, surprisingness and fun are more likely to receive higher scores and more comments. On the contrary, posts that require specialized knowledge to understand receive lower scores and fewer comments, suggesting that science jargon or specialized topics can reduce public engagement with science. For controversy, we observe different effects for comments and scores: posts with higher controversy values receive significantly more comments, but not scores. Such a result suggests that controversial content attracts more engagement through comments. While the number of comments is usually highly correlated with the post scores, in this setting, it is possible that controversial posts also receive more downvotes, which removes the effect of increased engagement. Our result also connects to the literature on framing (Pan and Kosicki, 1993; Scheufele, 1999). Existing research has found that framing the same issue in different ways could lead to different communication outcomes and audience perceptions (D'Angelo and Lombard, 2008). While we didn't study the linguistic features that contribute

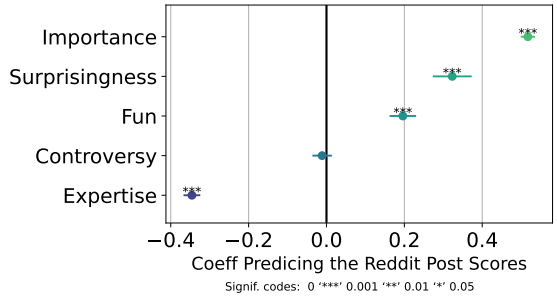


Figure 5.5: When sharing the same science news story, posts with higher estimated importance, surprisingness and fun receive higher scores on Reddit. Posts that require more specialized knowledge to understand receive lower scores.

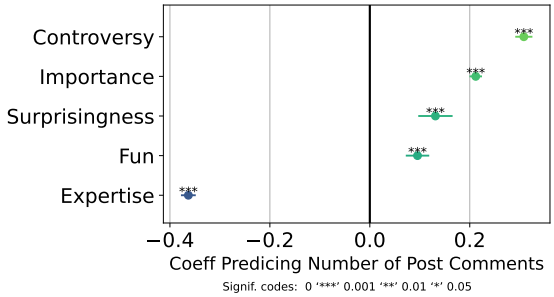


Figure 5.6: When sharing the same science news story, posts with higher estimated importance, surprisingness, fun, and controversy receive more comments on Reddit. Posts that require more specialized knowledge to understand receive less comments.

to different potential perceptions, our model could allow future researchers or journalists to directly estimate the potential public perceptions of the same information but with different frames.

5.6 Discussions

In this study, we explore computational modeling of public perceptions of individual science news stories and demonstrate the potential applications of the built dataset and NLP models. However, the current study also has the following limitations: (1) It primarily focuses on science news stories. While science news is crucial for science communication, the recent rise of short video platforms has transformed how people consume information. Communication on platforms like TikTok differs significantly from traditional science journalism practices. In the future, we plan to explore science communication on more emerging media platforms like TikTok. (2) When modeling public perceptions, we only focus on holistic perceptions. However, as the results have shown, people may perceive the same scientific information differently based on their personal background. In the future, I plan to build background-information-aware prediction models that can help science communicators to better tailor to the needs of diverse audiences. (3) While in this study we demonstrate that the estimated perceptions can be used to better predict the engagement with scientific information, it is still relatively indirect regarding effective science communication. In the long term, I plan to build new systems that can directly help journalists and scientists to better create science news stories for targeted audiences. (4) While we show that the developed models reliably estimate the perceptions of scientific information on most of the dimensions, dimensions like EXAGGERATION are hard to predict. However, despite the relatively low performance, downstream

analysis shows that the estimated perceptions could still be used to predict public engagement with social media posts. Such a result suggests that the developed models could still capture meaningful variations of scientific information. However, we acknowledge that because of the potential prediction inaccuracy, our findings may potentially face the risk of Type I errors.

5.7 Ethics

Despite its potential to help journalists to better understand the public perception of scientific information, it is essential to discuss the potential risks and ethical concerns for the developed tool. First of all, such a tool could be potentially used to promote exaggerated content as it allows large-scale analysis of exaggeration and controversy. However, on the other hand, it also allows researchers and government agencies to identify potential miscommunication of science at a large scale. Therefore, it is important to make sure that such a tool is being responsibly used in real-world applications. To this end, when releasing the models, we plan to create a special user agreement and will not distribute it automatically at scale. Furthermore, we plan to work with journalists to understand their perceptions and usages of such tools. Such a step could inform the design of responsible and effective technologies for science journalism. Understanding the audience and selecting newsworthy information have long been the core task for journalists and journalism. Despite all the convenience and automation technology could bring, we believe that technologies should never be used to replace human journalists. On the contrary, given all the potential risks, we believe that it becomes even more important to work with journalists to ensure the responsible deployment of such tools.

5.8 Conclusions

Modeling public perceptions of scientific information is important for effective science communication. In this study, we create a computational pipeline to comprehensively model public perceptions of science news. We developed a twelve-dimension framework for modeling public perceptions of individual science news stories and constructed a large-scale dataset covering ratings for science news stories as well as detailed background information of the annotators. We show that the public perception of scientific information could be affected by both the field of the content as well as individual's experience with science. We further develop a new NLP model that can automatically estimate public perceptions of individual news stories. Further analysis on Reddit posts shows that the estimated perceptions could help to predict public engagement with scientific information. Overall, our study could help science communicators to better understand the audience of

science communication and could further help to promote the public engagement with science in the long-term.

CHAPTER 6

Conclusion

In the final chapter, I summarize the findings and the contributions of the three studies. Then I discuss the limitations of my dissertation and future directions on using computational methods to facilitate better science communication.

6.1 Summary of Findings

The first study focuses on modeling expressed certainty and uncertainty in scholarly communications. I build a new computational framework to model sentence-level and aspect-level certainty in scientific claims written by scientists and journalists, and further conduct large-scale analysis to uncover how journalists and scientists communicate certainty and uncertainty for science writing. Despite existing debates on whether journalists overstate scientific certainty, our large-scale analysis over 431K scientific findings suggests that the journalist may actually add more hedges and downplay the certainty of scientific claims. Furthermore, I found that team sizes and the journal impact factor affect the way scientists present certainty in their findings. Findings presented by larger teams and in low-impact journals tend to have a higher degree of certainty, while such a pattern does not persist for the journalists' descriptions of scientific findings. Such a result suggests a potential information loss in the science communication process.

The second study builds upon the first by expanding the focus beyond a single language dimension of science communication, aiming to model the holistic information change in the process. In this study, I introduce a new measure named the "information matching score (IMS)," which quantifies the degree of informational alignment for any pair of scientific findings. I further developed a computational pipeline to automatically model IMS and conducted large-scale analyses to demonstrate the usage of IMS for science communication research. Our analysis revealed that in communicating scientific findings, press releases tend to adhere more closely to the original content, while general news outlets' descriptions of scientific findings often undergo more informational changes. Additionally, I extended the analysis of certainty changes from abstracts to

full-text papers. I observed a similar pattern in paper abstracts, where journalists' descriptions exhibited lower certainty and less exaggeration compared to the original abstracts. However, a different pattern emerged when comparing these descriptions to findings in other sections of research papers. For instance, journalists' descriptions showed higher certainty when compared to findings presented in limitation sections. Such a result highlights the importance of considering the entire research paper rather than just abstracts when analyzing scholarly communications.

The third study shifts focus towards communication audiences, aiming to comprehend how the public perceives scientific information received from the news media. I introduce a new multi-dimensional framework to model the public perception of scientific information and construct a large-scale science news perception dataset to support research on understanding the public reception of scientific information. Additionally, I develop NLP models capable of automatically estimating the potential perception of scientific information across news articles, social media texts, and paper abstracts. Through extensive analysis, I demonstrate that the created dataset and the model offer valuable insights into how the public perceives and engages with scientific information presented in media. This study bridges the gap between scientific communication and public understanding, shedding light on the dynamics of information perception in the public domain.

6.2 Contributions

Overall my dissertation contributes to the literature of both science communication and natural language processing.

On the science communication side, my dissertation provides new insights into information change and public engagement with science. For example, existing research has long been debating whether journalists overstate scientific uncertainty. Through large-scale analysis of 431K scientific findings, I provide empirical evidence that journalists may actually play down the certainty of scientific findings. Moreover, the computational methods developed in my dissertation present new opportunities for science communicators to better contextualize science communication according to the diverse needs of their audience. Previous research has highlighted the importance of modeling perceptions of scientific information across broad audiences (Schäfer et al., 2018; Atkin et al., 1983). In my third study, I introduce new NLP methods capable of automatically estimating public perceptions of scientific information and further demonstrate that the estimated perceptions can be used to predict public engagement with science on social media. Overall, my dissertation showcases the feasibility and value of leveraging NLP methods to enhance science communication, offering valuable insights into information change and facilitating more meaningful public engagement with science In summary, via developing new computational methods for analyzing science communication data, my research enables large-scale analysis to not only validate existing

findings (Study 1) but also answer new research questions about science communication (study 3). The developed methods could also be used by other researchers to analyze science communication at scale and answer important research questions.

For natural language processing, my dissertation introduces new frameworks, datasets, and models to understand nuanced information expressed in language. For instance, previous studies on detecting uncertainty in language often rely on hedges as a proxy, yet the effectiveness of hedges as proxies for uncertainty had not been quantitatively evaluated. Through the construction of a dedicated dataset, I was able to quantitatively examine the effectiveness of hedges on modeling textual uncertainty. I found that hedges may not effectively capture the full variation of both sentence-level and aspect-level uncertainty. I further propose new computational methods to measure both sentence-level and aspect-level certainty in text. These methods serve as efficient tools for automatically analyzing textual certainty, contributing to a deeper understanding of how language expresses nuanced information. As another major contribution, my dissertation introduces new theories, resources, and models to effectively model subjective social constructs that are integral to human communications. For instance, in the third study, I focus on modeling public perceptions of scientific information, representing a novel subjective NLP modeling task. Despite existing efforts in modeling subjective concepts in language, large-scale datasets with both nuanced annotations and detailed information about annotators' backgrounds are rare. The dataset generated from my dissertation studies fills this gap and could serve as a valuable resource for future research in developing social-aware NLP models that resonate with people's diverse interests.

Science communication not only concerns the research community but connects to broader audiences in the entire society. Through large-scale analysis of science communication data, my dissertation studies also have broader implications for stakeholders outside the science community like journalists and policymakers. For example, in the first study, we provide empirical evidence that journalists are not overstating science, contrasting with many existing findings and societal perceptions. Such a result could help society to better understand the current science communication practices and the role of science journalists in science communications.

6.3 Limitations

Despite the contributions of my dissertation studies, they have the following limitations.

Modality One of the major limitations of my dissertation is it primarily focuses on textual communication including news stories and social media texts. However, scientific information can be communicated in various modalities including text, videos, and audio. For example, video platforms like TikTok and YouTube allow both professional and amateur science communicators to

engage with lay audiences and existing research suggests that user-generated content may actually be more popular than content generated by professionals (Welbourne and Grant, 2016). Therefore, it is important to explore how amateur science communicators disseminate scientific information through videos.

Communication Channels The scope of my dissertation is confined to science communication through news (Study 1-3) and social media platforms (Study 2-3). While news stories and social media are two prominent channels for science communication, lay people also actively engage scientific knowledge in other forms like textbooks, novels, and movies (Barriga et al., 2010). For example, Barnett et al. (2006) found that science fiction films could shape student's perception of science. Therefore, studying how popular culture portrays science and how such portrayals affect the public understanding of science is also very important for science communication research.

Practical Implications The ultimate goal of science communication is to promote public engagement with scientific information. While my dissertation introduces new NLP models for analyzing scientific text and new empirical findings for understanding current practices in science communication, there remains a gap between these models and a comprehensive system that directly assists scientists and journalists in effectively communicating science to the public. For example, one of the major findings in the first study is journalists do not overstate the certainty of scientific findings. While such a finding provides important evidence to address a longstanding debate about whether journalists overstate science, it does not provide direct guidance on how science should be communicated.

6.4 Future Work

Effectively communicating science to the public is important yet challenging. In the long term, I plan to build better models and systems to help promote science communication in our society. More specially, I am interested in the following three directions.

Evaluating LLM's Capability on Science Communication Tasks Science communication encompasses many language-processing tasks from language generation to misinformation detection. Large Language Models have shown promising performances on various language processing tasks. However, it is unclear how they perform on NLP tasks related to science communication. The scientific text represents a special domain of text that is knowledge-intensive, and effectively communicating science requires not only language understanding capabilities but also the ability to understand domain-specific knowledge. Therefore, science communication offers a unique and

important setting to evaluate the capabilities of general-purpose language models. In the future, I plan to build a comprehensive benchmark to evaluate LLMs' capabilities for science communication tasks

Pesonalized Science News Generation The science community produces a large volume of research outputs every day, but only a very small portion of them get picked up by the media and further communicated to the public. However, different individuals have different needs for scientific information, and it is nearly impossible for traditional media to tailor to individuals' needs for science. Recent advances in Large Language Models offer huge opportunities for building personalized science news generation systems that can be used to fulfill the diverse needs of each individual. In the future, I plan to develop models that are able to automatically select and present information for individual users. Such a system could help to speed up the science communication process and promote societal-level understanding of science.

Modeling Science Communication in Short Videos Short videos have become a popular platform where people actively share and engage with scientific information. However, analyzing the content of short videos at scale poses a new challenge. In the future, I plan to develop new computational methods that can be used to analyze scientific information shared in videos. Furthermore, I plan to conduct a large-scale analysis on TikTok and YouTube to study how people engage with scientific information in videos and whether it differs from textual content.

BIBLIOGRAPHY

Achterberg, P., De Koster, W., and Van der Waal, J. (2017). A science confidence gap: Education, trust in scientific methods, and trust in scientific institutions in the united states, 2014. *Public Understanding of Science*, 26(6):704–720.

Agarwal, T. (1992). *An analysis of the readability of science and technology reporting in four national newspapers*. PhD thesis, Loughborough University of Technology.

Ahir, H., Bloom, N., and Furceri, D. (2018). The world uncertainty index. *Available at SSRN 3275033*.

Angler, M. (2017). *Science journalism: an introduction*. Routledge.

Armeni, C. (2004). From whom the general public chooses to learn science: a case study.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Atkin, C. K., Burgoon, J. K., and Burgoon, M. (1983). How journalists perceive the reading audience. *Newspaper research journal*, 4(2):51–63.

Augenstein, I., Baldwin, T., Cha, M., Chakraborty, T., Ciampaglia, G. L., Corney, D., DiResta, R., Ferrara, E., Hale, S., Halevy, A. Y., Hovy, E. H., Ji, H., Menczer, F., Míguez, R., Nakov, P., Scheufele, D. A., Sharma, S., and Zagni, G. (2023). Factuality challenges in the era of large language models. *ArXiv*, abs/2310.05189.

Augenstein, I., Lioma, C., Wang, D., Chaves Lima, L., Hansen, C., Hansen, C., and Simonsen, J. G. (2019). MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

August, T., Card, D., Hsieh, G., Smith, N. A., and Reinecke, K. (2020a). Explain like I am a scientist: The linguistic barriers of entry to r/science. In Bernhaupt, R., Mueller, F. F., Verweij, D., Andres, J., McGrenere, J., Cockburn, A., Avellino, I., Goguey, A., Bjøn, P., Zhao, S., Samson, B. P., and Kocielnik, R., editors, *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–12. ACM.

August, T., Kim, L., Reinecke, K., and Smith, N. A. (2020b). Writing strategies for science communication: Data and computational analysis. In *Proceedings of the 2020 Conference on*

Empirical Methods in Natural Language Processing (EMNLP), pages 5327–5344, Online. Association for Computational Linguistics.

Badenschier, F. and Wormer, H. (2011). Issue selection in science journalism: Towards a special theory of news values for science news? In *The sciences' media connection–public communication and its repercussions*, pages 59–85. Springer.

Bak, H.-J. (2001). Education and public attitudes toward science: Implications for the "deficit model" of education and support for science and technology. *Social Science Quarterly*, 82(4):779–795.

Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011). Everyone's an influencer: quantifying influence on twitter. In King, I., Nejdl, W., and Li, H., editors, *Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011*, pages 65–74. ACM.

Barnett, M., Wagner, H., Gatling, A., Anderson, J., Houle, M., and Kafka, A. (2006). The impact of science fiction film on student understanding of science. *Journal of Science Education and Technology*, 15:179–191.

Barriga, C. A., Shapiro, M. A., and Fernandez, M. L. (2010). Science information in fictional movies: Effects of context and gender. *Science Communication*, 32(1):3–24.

Batteux, E., Avri, B., Johnson, S. G., and Tuckett, D. (2021). The negative consequences of failing to communicate uncertainties during a pandemic: The case of covid-19 vaccines. *medRxiv*.

Bauer, M., Durant, J., and Evans, G. (1994). European public perceptions of science. *international Journal of Public opinion research*, 6(2):163–186.

Bauer, R. A. (1958). The communicator and the audience. *Journal of Conflict Resolution*, 2(1):67–77.

Bednarek, M. (2016). Investigating evaluation and news values in news items that are shared through social media. *Corpora*, 11(2):227–257.

Bell, A. (1994). Media (mis) communication on the science of climate change. *Public understanding of science*, 3(3):259.

Bell, R. L. and Lederman, N. G. (2003). Understandings of the nature of science and decision making on science and technology based issues. *Science education*, 87(3):352–377.

Beltagy, I., Lo, K., and Cohan, A. (2019a). SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Beltagy, I., Lo, K., and Cohan, A. (2019b). SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Berg, J. M., Bhalla, N., Bourne, P. E., Chalfie, M., Drubin, D. G., Fraser, J. S., Greider, C. W., Hendricks, M., Jones, C., Kiley, R., et al. (2016). Preprints for the life sciences. *Science*, 352(6288):899–901.

Besley, J. C. (2013). The state of public opinion research on attitudes and understanding of science and technology. *Bulletin of Science, Technology & Society*, 33(1-2):12–20.

Besley, J. C., Kramer, V. L., Yao, Q., and Toumey, C. (2008). Interpersonal discussion following citizen engagement about nanotechnology: What, if anything, do they say? *Science Communication*, 30(2):209–235.

Blank, J. M. and Shaw, D. (2015). Does partisanship shape attitudes toward science and public policy? the case for ideology and religion. *The ANNALS of the American Academy of Political and Social Science*, 658(1):18–35.

Board, N. S. (2012). Science and technology: Public attitudes and understanding.

Boiarsky, C. (2016). *Risk communication and miscommunication: Case studies in science, technology, engineering, government, and community organizations*. University Press of Colorado.

Boissonnet, A., Saeidi, M., Plachouras, V., and Vlachos, A. (2022). Explainable assessment of healthcare articles with QA. In Demner-Fushman, D., Cohen, K. B., Ananiadou, S., and Tsujii, J., editors, *Proceedings of the 21st Workshop on Biomedical Language Processing, BioNLP@ACL 2022, Dublin, Ireland, May 26, 2022*, pages 1–9. Association for Computational Linguistics.

Boulter, D. (1999). Public perception of science and associated general issues for the scientist. *Phytochemistry*, 50(1):1–7.

Bourk, M., León, B., and Davis, L. S. (2018). Entertainment in science: Useful in small doses. In *Communicating science and technology through online video*, pages 90–106. Routledge.

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

Bozeman, B. and Corley, E. (2004). Scientists' collaboration strategies: implications for scientific and technical human capital. *Research policy*, 33(4):599–616.

Bratton, L., Adams, R. C., Challenger, A., Boivin, J., Bott, L., Chambers, C. D., and Sumner, P. (2019). The Association Between Exaggeration in Health-Related Science News and Academic Press Releases: A Replication Study. *Wellcome open research*, 4.

Brillouin, L. (2014). *Scientific uncertainty, and information*. Academic Press.

Bruine de Bruin, W. and Bostrom, A. (2013). Assessing what to address in science communication. *Proceedings of the National Academy of Sciences*, 110(supplement_3):14062–14068.

Bubela, T., Nisbet, M. C., Borchelt, R., Brunger, F., Critchley, C., Einsiedel, E., Geller, G., Gupta, A., Hampel, J., Hyde-Lay, R., et al. (2009). Science communication reconsidered. *Nature biotechnology*, 27(6):514–518.

Budescu, D. V., Broomell, S., and Por, H.-H. (2009). Improving communication of uncertainty in the reports of the intergovernmental panel on climate change. *Psychological science*, 20(3):299–308.

Budescu, D. V., Por, H.-H., and Broomell, S. B. (2012). Effective communication of uncertainty in the ipcc reports. *Climatic change*, 113(2):181–200.

Burns, R. B., Moskowitz, M. A., Osband, M. A., and Kazis, L. E. (1995). Newspaper reporting of the medical literature. *Journal of general internal medicine*, 10:19–24.

Burns, T. W., O'Connor, D. J., and Stocklmayer, S. M. (2003). Science communication: a contemporary definition. *Public understanding of science*, 12(2):183–202.

Canese, K. and Weis, S. (2013). Pubmed: the bibliographic database. *The NCBI handbook*, 2(1).

Caple, H. (2018). News values and newsworthiness. In *Oxford research encyclopedia of communication*.

Castell, S., Charlton, A., Clemence, M., Pettigrew, N., Pope, S., Quigley, A., Shah, J. N., and Silman, T. (2014). Public attitudes to science 2014. *London, Ipsos MORI Social Research Institute*, 194:28.

Caulfield, T. and Condit, C. (2012). Science and the sources of hype. *Public Health Genomics*, 15(3-4):209–217.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., and Specia, L. (2017). SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Cheng, D., Claessens, M., Gascoigne, T., Metcalfe, J., Schiele, B., and Shi, S. (2008). Communicating science in social contexts. *New Models*.

Cheng, H. and Gonzalez-Ramirez, J. (2021). Trust and the media: Perceptions of climate change news sources among us college students. *Postdigital Science and Education*, 3:910–933.

Clark, S. K. (2013). *Writing strategies for science*. Teacher Created Materials.

Cohan, A., Feldman, S., Beltagy, I., Downey, D., and Weld, D. (2020). SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Collins, K., Shiffman, D., and Rock, J. (2016). How are scientists using social media in the workplace? *PloS one*, 11(10):e0162680.

Corley, P. C. and Wedeking, J. (2014). The (dis) advantage of certainty: The importance of certainty in language. *Law & Society Review*, 48(1):35–62.

Dai, E., Sun, Y., and Wang, S. (2020). Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. In Choudhury, M. D., Chunara, R., Culotta, A., and Welles, B. F., editors, *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, pages 853–862. AAAI Press.

Dalecki, L., Lasorsa, D. L., and Lewis, S. C. (2009). The news readability problem. *Journalism Practice*, 3(1):1–12.

D'Angelo, P. and Lombard, M. (2008). The power of the press: The effects of press frames in political campaign news on media perceptions. *Atlantic Journal of Communication*, 16(1):1–32.

Davenport, T. H. and Beck, J. C. (2001). The attention economy. *Ubiquity*, 2001(May):1–es.

De Semir, V., Ribas, C., and Revuelta, G. (1998). Press releases of science journal articles and subsequent newspaper stories on the same topic. *Jama*, 280(3):294–295.

Dewey, J. (1929). The quest for certainty.

Dickson, D. (2005). The case for a 'deficit model'of science communication. *SciDev. net*, 27:1–6.

Dieckmann, N. F., Slovic, P., and Peters, E. M. (2009). The use of narrative evidence and explicit likelihood by decisionmakers varying in numeracy. *Risk Analysis: An International Journal*, 29(10):1473–1488.

Dittrich, R., Francis, B., Hatzinger, R., and Katzenbeisser, W. (2007). A paired comparison approach for the analysis of sets of likert-scale responses. *Statistical Modelling*, 7(1):3–28.

Dixon, G. N. and Clarke, C. E. (2013). Heightening uncertainty around certain science: Media coverage, false balance, and the autism-vaccine controversy. *Science Communication*, 35(3):358–382.

Dogan, R. I., Leaman, R., and Lu, Z. (2014). Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Druzdzel, M. J. (1989). Verbal uncertainty expressions: Literature review. *Pittsburgh, PA: Carnegie Mellon University, Department of Engineering and Public Policy*, pages 1–13.

Dunwoody, S. (2021). Science journalism: Prospects in the digital age. In *Routledge handbook of public communication of science and technology*, pages 14–32. Routledge.

Durant, J. R., Evans, G. A., and Thomas, G. P. (1989). The public understanding of science. *Nature*, 340(6228):11–14.

D'andrea, L. and Declich, A. (2005). The sociological nature of science communication. *JCom*, 4(2):1–9.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *The quarterly journal of economics*, pages 643–669.

Evans, G. and Durant, J. (1995). The relationship between knowledge and attitudes in the public understanding of science in britain. *Public understanding of science*, 4(1):57.

Evans, J. A. (2008). Electronic publication and the narrowing of science and scholarship. *science*, 321(5887):395–399.

Fahnestock, J. (1986). Accommodating science: The rhetorical life of scientific facts. *Written communication*, 3(3):275–296.

Fahnestock, J. (2009). The rhetoric of the natural sciences. *The SAGE handbook of rhetorical studies*, pages 175–195.

Falk, J. H., Storksdieck, M., and Dierking, L. D. (2007). Investigating public science interest and understanding: Evidence for the importance of free-choice learning. *Public understanding of Science*, 16(4):455–469.

Fang, X., Wang, K., Han, D., He, X., Wei, J., Zhao, L., Imam, M. U., Ping, Z., Li, Y., Xu, Y., et al. (2016). Dietary magnesium intake and the risk of cardiovascular disease, type 2 diabetes, and all-cause mortality: a dose–response meta-analysis of prospective cohort studies. *BMC medicine*, 14(1):1–13.

Farkas, R., Vincze, V., Móra, G., Csirik, J., and Szarvas, G. (2010). The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*, pages 1–12, Uppsala, Sweden. Association for Computational Linguistics.

Feldman, L. and Hart, P. S. (2021). Upping the ante? the effects of "emergency" and "crisis" framing in climate change news. *Climatic Change*, 169(1-2):10.

Fischhoff, B. (2012a). Communicating uncertainty fulfilling the duty to inform. *Issues in Science and Technology*, 28(4):63–70.

Fischhoff, B. (2012b). Communicating uncertainty: Fulfilling the duty to inform. *Issues in Science and Technology*, 28.

Fischhoff, B. (2013). The sciences of science communication. *Proceedings of the National Academy of Sciences*, 110(supplement_3):14033–14039.

Fischhoff, B., Kadvany, J., and Kadvany, J. D. (2011). *Risk: A very short introduction*. Oxford University Press, USA.

Fleerackers, A., Shores, K., Chtena, N., and Alperin, J. P. (2024). Unreviewed science in the news: The evolution of preprint media coverage from 2014-2021. *Quantitative Science Studies*, pages 1–40.

Fortner, R. W., Lee, J.-Y., Corney, J. R., Romanello, S., Bonnell, J., Luthy, B., Figuerido, C., and Ntsiko, N. (2000). Public understanding of climate change: Certainty and willingness to act. *Environmental Education Research*, 6(2):127–141.

Fox, C. R. and Ülkümen, G. (2011). Distinguishing two dimensions of uncertainty. *Perspectives on thinking, judging, and decision making*, 14.

Franklin, J. (2008). The end of science journalism. In *Journalism, Science and Society*, pages 155–168. Routledge.

French, S. (1995). Uncertainty and imprecision: modelling and analysis. *Journal of the Operational Research Society*, 46(1):70–79.

Friedman, S. M., Dunwoody, S., and Rogers, C. L. (1999). *Communicating uncertainty: Media coverage of new and controversial science*. Routledge.

Friedman, S. M., Dunwoody, S., and Rogers, C. L. (2012). *Communicating uncertainty: Media coverage of new and controversial science*. Routledge.

Funk, C., GOTTFRIED, J., and MITCHELL, A. (2017). Science news and information today. *Pew Research Center*.

Gałecki, A. and Burzykowski, T. (2013). Linear mixed-effects model. In *Linear mixed-effects models using R*, pages 245–273. Springer.

Galtung, J. and Ruge, M. H. (1965). The structure of foreign news: The presentation of the congo, cuba and cyprus crises in four norwegian newspapers. *Journal of peace research*, 2(1):64–90.

Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2013). PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, Atlanta, Georgia. Association for Computational Linguistics.

Garett, R. and Young, S. D. (2021). Online misinformation and vaccine hesitancy. *Translational Behavioral Medicine*, 11:2194 – 2199.

Garvey, W. D. and Griffith, B. C. (1971). Scientific communication: Its role in the conduct of research and creation of knowledge. *American psychologist*, 26(4):349.

Gero, K. I., Liu, V., Huang, S., Lee, J., and Chilton, L. B. (2021). What makes tweetorials tick: How experts communicate complex topics on twitter. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–26.

Ginev, D. and Miller, B. R. (2020). Scientific statement classification over arXiv.org. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1219–1226, Marseille, France. European Language Resources Association.

Goidel, K. and Nisbet, M. (2006). Exploring the roots of public participation in the controversy over embryonic stem cell research and cloning. *Political Behavior*, 28:175–192.

Gonçalves, S., Cortez, P., and Moro, S. (2018). A deep learning approach for sentence classification of scientific abstracts. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, pages 479–488. Springer.

Gustafson, A. and Rice, R. E. (2019). The effects of uncertainty frames in three science communication topics. *Science Communication*, 41(6):679–706.

Gyasi, W. K. (2013). The role of readability in science education in ghana: A readability index analysis of ghana association of science teachers textbooks for senior high school. *IOSR Journal of Research & Method in Education*, 2(1):9–19.

Hall, K. L., Vogel, A. L., Huang, G. C., Serrano, K. J., Rice, E. L., Tsakraklides, S. P., and Fiore, S. M. (2018). The science of team science: A review of the empirical evidence and research gaps on collaboration in science. *American Psychologist*, 73(4):532.

Han, P. K., Klein, W. M., and Arora, N. K. (2011). Varieties of uncertainty in health care: a conceptual taxonomy. *Medical Decision Making*, 31(6):828–838.

Harcup, T. and O'neill, D. (2017). What is news? news values revisited (again). *Journalism studies*, 18(12):1470–1488.

Hart, P. S. and Nisbet, E. C. (2012). Boomerang effects in science communication: How motivated reasoning and identity cues amplify opinion polarization about climate mitigation policies. *Communication research*, 39(6):701–723.

Hart, R. P. and Childers, J. P. (2004). Verbal certainty in american politics: An overview and extension. *Presidential Studies Quarterly*, 34(3):516–535.

Hayes, A. F. and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Henriksson, A. and Velupillai, S. (2010). Levels of certainty in knowledge-intensive corpora: an initial annotation study. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 41–45, Uppsala, Sweden. University of Antwerp.

Hentschel, M., Alonso, O., Counts, S., and Kandylas, V. (2014). Finding users we trust: Scaling up verified twitter users using their communication patterns. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Hirst, R. (2003). Scientific jargon, good and bad. *Journal of technical writing and communication*, 33(3):201–229.

Holmes, J. (1982). Expressing doubt and certainty in english. *RELC journal*, 13(2):9–28.

Holub, A., Perona, P., and Burl, M. C. (2008). Entropy-based active learning for object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2008, Anchorage, AK, USA, 23-28 June, 2008*, pages 1–8. IEEE Computer Society.

Houlsby, N., Huszar, F., Ghahramani, Z., and Lengyel, M. (2011). Bayesian active learning for classification and preference learning. *CoRR*, abs/1112.5745.

Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., and Hovy, E. (2013). Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Howard, R. A. (1988). Uncertainty about probability: A decision analysis perspective. *Risk Analysis*, 8(1):91–98.

Hoxha, I., Agahi, R., Bimbashi, A., Aliu, M., Raka, L., Bajraktari, I., Beqiri, P., and Adams, L. V. (2022). Higher covid-19 vaccination rates are associated with lower covid-19 mortality: A global analysis. *Vaccines*, 11.

Htoo, T. H. H., Jin-Cheon, N., and Thelwall, M. (2023). Why are medical research articles tweeted? the news value perspective. *Scientometrics*, 128(1):207–226.

Hyland, K. (1996). Writing without conviction? hedging in science research articles. *Applied linguistics*, 17(4):433–454.

Hyland, K. (2005). *Metadiscourse: Exploring Interaction in Writing*. London: Continuum.

Hyland, K. and Salager-Meyer, F. (2008). Scientific writing. *Annual review of information science and technology*, 42(1):297.

Intemann, K. (2022). Understanding the problem of "hype": Exaggeration, values, and trust in science. *Canadian Journal of Philosophy*, 52(3):279–294.

Jeong, S. and Choi, J. Y. (2015). Collaborative research for academic knowledge creation: How team characteristics, motivation, and processes influence research impact. *Science and Public Policy*, 42(4):460–473.

Jones, M. G., Howe, A., and Rua, M. J. (2000). Gender differences in students' experiences, interests, and attitudes toward science and scientists. *Science education*, 84(2):180–192.

Jones, R. (2008). Fearing the fear of nanotechnology. *Nature*, 9.

Jurgens, D., Kumar, S., Hoover, R., McFarland, D., and Jurafsky, D. (2018). Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Kaiser, J. (2017). The preprint dilemma.

Kearnes, M. B., Macnaghten, P., and Wilsdon, J. (2006). Governing at the nanoscale: people, policies and emerging technologies.

Kopelman, L. (1978). Ethical controversies in medical research: the case of xyy screening. *Perspectives in Biology and Medicine*, 21(2):196–204.

Krieghbaum, H. (1940). The background and training of science writers. *Journalism Quarterly*, 17(1):15–18.

Kringelum, J. V., Kjærulff, S. K., Brunak, S., Lund, O., Oprea, T. I., and Taboureau, O. (2016). Chemprot-3.0: a global chemical biology diseases mapping. *Database: The Journal of Biological Databases and Curation*, 2016.

Kurmis, A. P. (2003). Understanding the limitations of the journal impact factor. *JBJS*, 85(12):2449–2454.

Kuru, O., Stecula, D., Lu, H., Ophir, Y., Chan, M.-p. S., Winneg, K., Hall Jamieson, K., and Albarracín, D. (2021). The effects of scientific messages and narratives about vaccination. *PLoS One*, 16(3):e0248328.

Lamers, W. S., Boyack, K., Larivière, V., Sugimoto, C. R., van Eck, N. J., Waltman, L., and Murray, D. (2021). Meta-research: Investigating disagreement in the scientific literature. *Elife*, 10:e72737.

Laugksch, R. C. (2000). Scientific literacy: A conceptual overview. *Science education*, 84(1):71–94.

Lee, A. M. and Chyi, H. I. (2014). When newsworthy is not noteworthy: Examining the value of news from the audience's perspective. *Journalism studies*, 15(6):807–820.

Lee, Y.-N., Walsh, J. P., and Wang, J. (2015). Creativity in scientific teams: Unpacking novelty and impact. *Research policy*, 44(3):684–697.

Leshner, A. I. (2003). Public engagement with science.

Levy, R., Bilu, Y., Hershcovich, D., Aharoni, E., and Slonim, N. (2014). Context dependent claim detection. In *International Conference on Computational Linguistics*.

Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C.-H., Leaman, R., Davis, A. P., Mattingly, C. J., Wiegers, T. C., and Lu, Z. (2016). Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Lin, H., Ma, J., Cheng, M., Yang, Z., Chen, L., and Chen, G. (2021). Rumor detection on twitter with claim-guided hierarchical graph attention networks. In *Conference on Empirical Methods in Natural Language Processing*.

Lin, W.-H., Wilson, T., Wiebe, J., and Hauptmann, A. (2006). Which side are you on? identifying perspectives at the document and sentence levels. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 109–116, New York City. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019a). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019b). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Lo, K., Wang, L. L., Neumann, M., Kinney, R., and Weld, D. (2020). S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Longnecker, N. (2023). Good science communication considers the audience. In *Teaching science students to communicate: a practical guide*, pages 21–30. Springer.

López-Goñi, I. and Sánchez-Angulo, M. (2018). Social networks as a tool for science communication and public engagement: focus on twitter. *FEMS Microbiology letters*, 365(2):fnx246.

MacKay, D. J. C. (1992). Information-based objective functions for active data selection. *Neural Comput.*, 4(4):590–604.

MacKenzie, D. (1998). The certainty trough. In *Exploring expertise*, pages 325–329. Springer.

MacLaughlin, A., Wihbey, J., and Smith, D. (2018). Predicting news coverage of scientific articles. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Magerman, T., Van Looy, B., and Song, X. (2010). Exploring the feasibility and accuracy of latent semantic analysis based text mining techniques to detect similarity between patent documents and scientific publications. *Scientometrics*, 82(2):289–306.

McCright, A. M., Dentzman, K., Charters, M., and Dietz, T. (2013). The influence of political ideology on trust in science. *Environmental Research Letters*, 8(4):044029.

Meena, A. and Prabhakar, T. (2007). Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis. In *European conference on information retrieval*, pages 573–580. Springer.

Mencher, M. and Shilton, W. P. (1997). *News reporting and writing*. Brown & Benchmark Publishers Madison, WI.

Morosoli, J. J., Colodro-Conde, L., Barlow, F. K., and Medland, S. E. (2024). Scientific clickbait: Examining media coverage and readability in genome-wide association research. *Plos one*, 19(1):e0296323.

Mosleh, A. and Bier, V. M. (1996). Uncertainty about probability: a reconciliation with the subjectivist viewpoint. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 26(3):303–310.

Murray, D. S., Lamers, W., Boyack, K. W., Larivière, V., Sugimoto, C. R., van Eck, N. J., and Waltman, L. (2019). Measuring disagreement in science. In *ISSI*, pages 2370–2375.

Narasimhan, M. G. (2001). Controversy in science. *Journal of biosciences*, 26(3):299–304.

National Academies of Sciences, E., Medicine, et al. (2016). Science literacy: Concepts, contexts, and consequences.

National Academies of Sciences, Engineering, and Medicine (2017). Communicating science effectively: A research agenda.

Neureiter, A., Stubenvoll, M., Kaskeleviciute, R., and Matthes, J. (2021). Trust in science, perceived media exaggeration about covid-19, and social distancing behavior. *Frontiers in Public Health*, 9:670485.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., and Kiela, D. (2020). Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Nighojkar, A. and Licato, J. (2021). Improving paraphrase detection with the adversarial paraphrasing task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7106–7116, Online. Association for Computational Linguistics.

Nisbet, M. C. and Goidel, R. K. (2007). Understanding citizen perceptions of science controversy: bridging the ethnographic—survey research divide. *Public Understanding of science*, 16(4):421–440.

Nisbet, M. C. and Scheufele, D. A. (2009). What's next for science communication? promising directions and lingering distractions. *American journal of botany*, 96(10):1767–1778.

Nishal, S. and Diakopoulos, N. (2022). From crowd ratings to predictive models of newsworthiness to support science journalism. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–28.

Olofsson, A., Öhman, S., and Rashid, S. (2006). Attitudes to gene technology: the significance of trust in institutions. *European Societies*, 8(4):601–624.

Osborne, J., Simon, S., and Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. *International journal of science education*, 25(9):1049–1079.

Pan, Z. and Kosicki, G. M. (1993). Framing analysis: An approach to news discourse. *Political communication*, 10(1):55–75.

Pei, J., Ananthasubramaniam, A., Wang, X., Zhou, N., Dedeloudis, A., Sargent, J., and Jurgens, D. (2022). Potato: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Pei, J. and Jurgens, D. (2021). Measuring sentence-level and aspect-level (un)certainty in science communications. In Moens, M., Huang, X., Specia, L., and Yih, S. W., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 9959–10011. Association for Computational Linguistics.

Peng, H., Teplitskiy, M., and Jurgens, D. (2020). Author mentions in science news reveal widespread ethnic bias. *arXiv preprint arXiv:2009.01896*.

Peters, E., Hart, P. S., Tusler, M., and Fraenkel, L. (2014a). Numbers matter to informed patient choices: a randomized design across age and numeracy levels. *Medical Decision Making*, 34(4):430–442.

Peters, E., Meilleur, L., and Tompkins, M. K. (2014b). Numeracy and the affordable care act: Opportunities and challenges. In *Health literacy and numeracy: Workshop summary*. National Academies Press (US).

Politi, M. C., Han, P. K., and Col, N. F. (2007). Communicating the uncertainty of harms and benefits of medical interventions. *Medical Decision Making*, 27(5):681–695.

Powell, M. and Lee Kleinman, D. (2008). Building citizen capacities for participation in nanotechnology decision-making: the democratic virtues of the consensus conference model. *Public Understanding of Science*, 17(3):329–348.

Prabhakaran, V., Hamilton, W. L., McFarland, D., and Jurafsky, D. (2016). Predicting the rise and fall of scientific topics from trends in their rhetorical framing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1170–1180, Berlin, Germany. Association for Computational Linguistics.

Prabhakaran, V., Rambow, O., and Diab, M. (2012). Predicting overt display of power in written dialogs. In *Proceedings of NAACL*, pages 518–522.

Publons (2018). Global state of peer review. *Clarivate Analytics*.

Recasens, M., Danescu-Niculescu-Mizil, C., and Jurafsky, D. (2013). Linguistic models for analyzing and detecting biased language. In *Proceedings of ACL*, pages 1650–1659.

Reidsma, D. and Carletta, J. (2008). Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Richardson, L. (1990). *Writing strategies: Reaching diverse audiences*, volume 21. Sage Publications.

Rinott, R., Dankin, L., Perez, C. A., Khapra, M. M., Aharoni, E., and Slonim, N. (2015). Show me your evidence - an automatic method for context dependent evidence detection. In *Conference on Empirical Methods in Natural Language Processing*.

Roland, M.-C. (2009). Quality and integrity in scientific writing: prerequisites for quality in science communication. *Journal of Science Communication*, 8(2):A04.

Rosen, R. (1996). On the limitations of scientific knowledge. *Boundaries and barriers: On the limits to scientific knowledge, Reading, MA: Perseus Books*, pages 199–214.

Rowlands, I., Nicholas, D., and Huntington, P. (2004). Scholarly communication in the digital environment: what do authors want? *Learned Publishing*, 17(4):261–273.

Rubin, V. L. (2007). Stating with certainty or stating with doubt: Intercoder reliability results for manual annotation of epistemically modalized statements. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 141–144, Rochester, New York. Association for Computational Linguistics.

Rubin, V. L., Liddy, E. D., and Kando, N. (2006). Certainty identification in texts: Categorization model and manual tagging results. In *Computing attitude and affect in text: Theory and applications*, pages 61–76. Springer.

Saffer, A. J., Sommerfeldt, E. J., and Taylor, M. (2013). The effects of organizational twitter interactivity on organization–public relationships. *Public relations review*, 39(3):213–215.

Saha, S., Saint, S., and Christakis, D. A. (2003). Impact factor: a valid measure of journal quality? *Journal of the Medical Library Association*, 91(1):42.

Salita, J. T. (2015). Writing for lay audiences: A challenge for scientists. *Medical Writing*, 24:183–189.

Sarewitz, D. (2004). How science makes environmental controversies worse. *Environmental science & policy*, 7(5):385–403.

Saurí, R. and Pustejovsky, J. (2009). Factbank: a corpus annotated with event factuality. *Language resources and evaluation*, 43(3):227–268.

Schäfer, M. S., Füchslin, T., Metag, J., Kristiansen, S., and Rauchfleisch, A. (2018). The different audiences of science communication: A segmentation analysis of the swiss population's perceptions of science and their information and media use patterns. *Public understanding of science*, 27(7):836–856.

Scheufele, D. A. (1999). Framing as a theory of media effects. *Journal of communication*, 49(1):103–122.

Scheufele, D. A. (2013). Communicating science in social settings. *Proceedings of the National Academy of Sciences*, 110(supplement_3):14040–14047.

Scheufele, D. A. and Krause, N. M. (2019). Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, 116(16):7662–7669.

Schouten, K. and Frasincar, F. (2015). Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):813–830.

Shaar, S., Babulkov, N., Da San Martino, G., and Nakov, P. (2020). That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, Online. Association for Computational Linguistics.

Sharon, A. J. and Baram-Tsabari, A. (2014). Measuring mumbo jumbo: A preliminary quantification of the use of jargon in science communication. *Public Understanding of Science*, 23(5):528–546.

Shearer, E. and Grieco, E. (2019). Americans are wary of the role social media sites play in delivering the news. *Pew Research Center*, 2.

Shi, J., Visschers, V. H., and Siegrist, M. (2015). Public perception of climate change: The importance of knowledge and cultural worldviews. *Risk Analysis*, 35(12):2183–2201.

Shuchman, M. and Wilkes, M. S. (1997). Medical scientists and health news reporting: a case of miscommunication. *Annals of Internal Medicine*, 126(12):976–982.

Siipi, H. and Ahteensuu, M. (2011). The deficit model and the forgotten moral values.

Sinayev, A., Peters, E., Tusler, M., and Fraenkel, L. (2015). Presenting numeric information with percentages and descriptive risk labels: a randomized trial. *Medical decision making*, 35(8):937–947.

Smithson, M. (2012). *Ignorance and uncertainty: Emerging paradigms*. Springer Science & Business Media.

Soleimani, A., Monz, C., and Worring, M. (2019). Bert for evidence retrieval and claim verification. *Advances in Information Retrieval*, 12036:359 – 366.

Sompel, H. v. d., Payette, S., Erickson, J., Lagoze, C., and Warner, S. (2004). Rethinking scholarly communication: Building the system that scholars deserve. *D-Lib Magazine; 2004 [10] 9*.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T. (2020). Mpnet: Masked and permuted pre-training for language understanding. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Soni, S., Mitra, T., Gilbert, E., and Eisenstein, J. (2014). Modeling factuality judgments in social media text. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 415–420, Baltimore, Maryland. Association for Computational Linguistics.

Stewart, C. O. (2008). How a media controversy can influence a scientific publication. *Discourse Approaches to Politics, Society and Culture (DAPSAC)*.

Stilgoe, J., Lock, S. J., and Wilsdon, J. (2014). Why should we promote public engagement with science? *Public understanding of science*, 23(1):4–15.

Stokols, D., Hall, K. L., Taylor, B. K., and Moser, R. P. (2008). The science of team science: overview of the field and introduction to the supplement. *American journal of preventive medicine*, 35(2):S77–S89.

Sturgis, P. and Allum, N. (2004). Science in society: Re-evaluating the deficit model of public attitudes. *Public understanding of science*, 13(1):55–74.

Sud, P. and Thelwall, M. (2016). Not all international collaboration is beneficial: The m endeley readership and citation impact of biochemical research collaboration. *Journal of the Association for Information Science and Technology*, 67(8):1849–1857.

Suldovsky, B. (2016). In science communication, why does the idea of the public deficit always return? exploring key influences. *Public understanding of science*, 25(4):415–426.

Sullivan, J. W. N. (2018). *Limitations of science*. Read Books Ltd.

Sumner, P., Vivian-Griffiths, S., Boivin, J., Williams, A., Venetis, C. A., Davies, A., Ogden, J., Whelan, L., Hughes, B., Dalton, B., et al. (2014). The Association Between Exaggeration in Health Related Science News and Academic Press Releases: Retrospective Observational Study. *BMJ*, 349.

Szarvas, G., Vincze, V., Farkas, R., Móra, G., and Gurevych, I. (2012). Cross-genre and cross-domain detection of semantic uncertainty. *Computational Linguistics*, 38(2):335–367.

Tan, C. and Lee, L. (2014). A corpus of sentence-level revisions in academic writing: A step towards understanding statement strength in communication. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 403–408, Baltimore, Maryland. Association for Computational Linguistics.

Tan, N. Ö., Nguyen, T. B., Bensemann, J., Peng, A. Y., Bao, Q., Chen, Y., Gahegan, M., and Witbrock, M. (2023). Multi2claim: Generating scientific claims from multi-choice questions for scientific fact-checking. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Tavakol, M. and Dennick, R. (2011). Making sense of cronbach's alpha. *International journal of medical education*, 2:53.

Thelwall, M. and Sud, P. (2016). Mendeley readership counts: An investigation of temporal and disciplinary differences. *Journal of the Association for Information Science and Technology*, 67(12):3036–3050.

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

Thunnissen, D. P. (2003). Uncertainty classification for the design and development of complex systems. In *3rd annual predictive methods conference*, pages 1–16. Newport Beach CA.

Tichenor, P. J., Olien, C. N., Harrison, A., and Donohue, G. (1970). Mass communication systems and communication accuracy in science news reporting. *Journalism Quarterly*, 47(4):673–683.

Trench, B. (2008). Towards an analytical framework of science communication models. *Communicating science in social contexts: New models, new practices*, pages 119–135.

Vadapalli, R., Syed, B., Prabhu, N., Srinivasan, B. V., and Varma, V. (2018). When science journalism meets artificial intelligence : An interactive demonstration. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 163–168, Brussels, Belgium. Association for Computational Linguistics.

Valdeón, R. A. (2018). On the use of the term 'translation'in journalism studies. *Journalism*, 19(2):252–269.

Van Der Bles, A. M., van der Linden, S., Freeman, A. L., and Spiegelhalter, D. J. (2020). The effects of communicating uncertainty on public trust in facts and numbers. *Proceedings of the National Academy of Sciences*, 117(14):7672–7683.

Varner, J. (2014). Scientific outreach: toward effective public engagement with biological science. *BioScience*, 64(4):333–340.

Vold, E. T. (2006). Epistemic modality markers in research articles: a cross-linguistic and cross-disciplinary study. *International Journal of Applied Linguistics*, 16(1):61–87.

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

Wade, S. and Schramm, W. (1969). The mass media as sources of public affairs, science, and health knowledge. *Public Opinion Quarterly*, 33(2):197–209.

Wang, K., Shen, Z., Huang, C., Wu, C.-H., Eide, D., Dong, Y., Qian, J., Kanakia, A., Chen, A., and Rogahn, R. (2019a). A review of microsoft academic services for science of science studies. *Frontiers in Big Data*, 2:45.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Wang, Z., Hale, S. A., Adelani, D. I., Grabowicz, P. A., Hartmann, T., Flöck, F., and Jurgens, D. (2019b). Demographic inference and representative population estimates from multilingual social media data. In Liu, L., White, R. W., Mantrach, A., Silvestri, F., McAuley, J. J., Baeza-Yates, R., and Zia, L., editors, *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2056–2067. ACM.

Weigold, M. F. (2001). Communicating science: A review of the literature. *Science communication*, 23(2):164–193.

Weinburgh, M. (1995). Gender differences in student attitudes toward science: A meta-analysis of the literature from 1970 to 1991. *Journal of Research in science Teaching*, 32(4):387–398.

Weingart, P. and Guenther, L. (2016). Science communication and the issue of trust. *Journal of Science communication*, 15(05):C01.

Weiss, C. and Singer, E. (1988). *Reporting of social science in the national media*. Russell Sage Foundation.

Welbourne, D. J. and Grant, W. J. (2016). Science communication on youtube: Factors that affect channel and video popularity. *Public understanding of science*, 25(6):706–718.

Wilson, A., Robertson, J., McElduff, P., Jones, A., and Henry, D. (2010). Does it matter who writes medical news stories? *PLOS medicine*, 7(9):e1000323.

Windschitl, P. D. and Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied*, 2(4):343.

Woloshin, S. and Schwartz, L. M. (2002). Press releases: translating research into news. *Jama*, 287(21):2856–2858.

Wolpert, L. (1994). *The unnatural nature of science*. Harvard University Press.

Wood, W. and Eagly, A. H. (2009). Advantages of certainty and uncertainty. *The handbook of research synthesis and meta-analysis*, 2:455–472.

Wright, D. and Augenstein, I. (2021a). CiteWorth: Cite-worthiness detection for improved scientific document understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1796–1807, Online. Association for Computational Linguistics.

Wright, D. and Augenstein, I. (2021b). Semi-supervised exaggeration detection of health science press releases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10824–10836.

Wright, D., Wadden, D., Lo, K., Kuehl, B., Cohan, A., Augenstein, I., and Wang, L. L. (2022a). Generating scientific claims for zero-shot scientific fact checking. *ArXiv*, abs/2203.12990.

Wright, D., Wadden, D., Lo, K., Kuehl, B., Cohan, A., Augenstein, I., and Wang, L. L. (2022b). Generating scientific claims for zero-shot scientific fact checking. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2448–2460. Association for Computational Linguistics.

Wu, L., Wang, D., and Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, 566(7744):378–382.

Wuhrl, A. and Klinger, R. (2021). Claim detection in biomedical twitter posts. In *Workshop on Biomedical Natural Language Processing*.

Yeo, S. K., Xenos, M. A., Brossard, D., and Scheufele, D. A. (2015). Selecting our own science: How communication contexts and individual traits shape information seeking. *The ANNALS of the American Academy of Political and Social Science*, 658(1):172–191.

Yin, Y., Dong, Y., Wang, K., Wang, D., and Jones, B. F. (2022). Public use and public funding of science. *Nature human behaviour*, 6(10):1344–1350.

Zakhlebin, I. and Horvát, E.-A. (2020). Diffusion of scientific articles across online platforms. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 762–773.

Zhang, J., Xia, F., Wang, W., Bai, X., Yu, S., Bekele, T. M., and Peng, Z. (2016). Cocarank: A collaboration caliber-based method for finding academic rising stars. In *Proceedings of the 25th International Conference Companion on World Wide Web (WWW)*, pages 395–400. International World Wide Web Conferences Steering Committee.